

**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 THỜ MỜ TRỰC CẢM VÀ TÔPÔ SUY RỘNG**

**LUẬN ÁN TIẾN SĨ NGÀNH HỆ THỐNG THÔNG TIN**

**Hà Nội - Năm 2023**

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 TÔPO SUY RỘNG

LUẬN ÁN TIẾN SĨ NGÀNH HỆ THỐNG THÔNG TIN  
Mã số: 9 48 01 04

Xác nhận của Học viện  
Khoa học và Công nghệ

Người hướng dẫn 1  
(Ký, ghi rõ họ tên)

Người hướng dẫn 2  
(Ký, ghi rõ họ tên)

Hà Nội - Năm 2023

## LỜI CAM ĐOAN

Tôi xin được cam đoan đây là công trình nghiên cứu của riêng tôi dưới sự hướng dẫn của PGS. TS Nguyễn Long Giang và GS. TS Vũ Đức Thi tại Viện Công nghệ thông tin - Viện Hàn lâm Khoa học và Công nghệ Việt Nam. Các kết quả nghiên cứu lý thuyết và thực nghiệm trong luận án này được trình bày chính xác, trung thực và không sao chép từ bất kỳ nguồn tài liệu nào và dưới bất kỳ hình thức nào. Việc tham khảo các nguồn tài liệu được trích dẫn và ghi nguồn đầy đủ.

*Hà Nội, ngày 3 tháng 10 năm 2023*

**Trần Thanh Đại**

## LỜI CẢM ƠN

Luận án này được hoàn thiện với sự nỗ lực và cố gắng không ngừng của tác giả cùng với sự ân cần chỉ bảo, giúp đỡ của các thầy hướng dẫn, sự góp ý xác đáng của các chuyên gia, nhà khoa học, sự động viên về tinh thần của gia đình, bạn bè và đồng nghiệp trong suốt quá trình học tập và nghiên cứu của tác giả.

Trước tiên, tác giả xin bày tỏ lòng biết ơn đến PGS. TS Nguyễn Long Giang, GS. TS Vũ Đức Thi đã tận tình chỉ bảo, hướng dẫn và động viên tác giả hoàn thành luận án đúng mục tiêu và đúng tiến độ. Tác giả cũng xin được bày tỏ lời cảm ơn sâu sắc đến PGS. TS Lê Hoàng Sơn đã có những góp ý quý giá trong suốt quá trình thực hiện luận án này.

Tác giả xin gửi lời cảm ơn tới các thầy, cô giáo và cán bộ phòng quản lý nghiên cứu sinh tại Học viện Khoa học và Công nghệ, Viện Hàn lâm Khoa học và Công nghệ Việt Nam. Đặc biệt tác giả xin trân trọng cảm ơn phòng quản lý sau đại học của Viện Công nghệ thông tin đã nhiệt tình giúp đỡ, tạo ra môi trường nghiên cứu thuận lợi cho tác giả hoàn thành luận án đúng tiến độ và đúng quy định của Học viện.

Tác giả xin chân thành cảm ơn nhóm nghiên cứu AI 4.0 tại Viện Công nghệ thông tin (ITI) - Đại học Quốc gia Hà Nội đã giúp đỡ tác giả về mặt chuyên môn và tinh thần nghiên cứu trong suốt quá trình trao đổi và nghiên cứu học thuật tại ITI.

Tác giả xin chân thành cảm ơn tới Ban Giám Hiệu Trường Đại học Kinh tế Kỹ thuật Công nghiệp đã động viên tinh thần và tạo nhiều điều kiện thuận lợi trong suốt quá trình học tập và nghiên cứu.

Đặc biệt tác giả xin bày tỏ lòng biết ơn sâu sắc tới Gia đình và người thân đã hi sinh vô điều kiện, tạo điều kiện tốt nhất về tinh thần và thời gian cho tác giả trong suốt quá trình học tập và làm nghiên cứu.

**NCS Trần Thanh Đại**

## MỤC LỤC

<b>LỜI CAM ĐOAN</b> . . . . .	i
<b>LỜI CẢM ƠN</b> . . . . .	ii
<b>MỤC LỤC</b> . . . . .	v
<b>DANH MỤC CÁC THUẬT NGỮ, CÁC CHỮ VIẾT TẮT</b> . . . . .	vi
<b>DANH MỤC CÁC KÝ HIỆU</b> . . . . .	vii
<b>DANH MỤC HÌNH VẼ</b> . . . . .	ix
<b>DANH MỤC BẢNG BIỂU</b> . . . . .	xi
<b>MỞ ĐẦU</b> . . . . .	1
<b>CHƯƠNG 1. TỔNG QUAN BÀI TOÁN RÚT GỌN THUỘC TÍNH THEO TIẾP CẬN RS VÀ TÔPÔ</b>	<b>8</b>
1.1 Mở đầu . . . . .	8
1.2 Các khái niệm cơ bản . . . . .	10
1.2.1 Hệ thông tin và mô hình RS truyền thống . . . . .	10
1.2.2 RS mờ trực cảm . . . . .	12
1.2.3 Không gian tôpô . . . . .	16
1.2.4 Tập rút gọn . . . . .	17
1.3 Một số công thức tính toán độ thành viên . . . . .	18
1.3.1 Chuẩn hóa dữ liệu . . . . .	18
1.3.2 Độ đo độ tương tự . . . . .	19
1.4 Phương pháp đánh giá reduct . . . . .	21
1.4.1 Các tiêu chí đánh giá . . . . .	21
1.4.2 Mô hình và dữ liệu đánh giá . . . . .	22
1.4.3 Chỉ số đánh giá . . . . .	22
1.5 Một số phương pháp giảm thuộc tính . . . . .	24

1.5.1	Phương pháp giảm thuộc tính theo tiếp cận ma trận phân biệt . . . . .	24
1.5.2	Phương pháp giảm thuộc tính theo tiếp cận độ đo . . . . .	26
1.5.3	Phương pháp giảm thuộc tính theo tiếp cận tôpô . . . . .	31
1.6	Kết luận Chương 1 . . . . .	34
<b>CHƯƠNG 2. PHƯƠNG PHÁP RÚT GỌN THUỘC TÍNH TRONG BẢNG QUYẾT ĐỊNH THEO TIẾP CẬN RS MỜ TRỰC CẢM</b>		<b>35</b>
2.1	Mở đầu . . . . .	35
2.2	Xây dựng độ đo khoảng cách mờ trực cảm . . . . .	36
2.2.1	Khoảng cách giữa hai tập mờ trực cảm . . . . .	36
2.2.2	Khoảng cách giữa hai phân hoạch mờ trực cảm . . . . .	38
2.3	Giảm thuộc tính trong bảng quyết định sử dụng độ đo khoảng cách mờ trực cảm . . . . .	40
2.3.1	Đề xuất thuật toán tìm reduct theo phương pháp lai ghép filter - wrapper, sử dụng độ đo khoảng cách mờ trực cảm . . . . .	40
2.3.2	Thực nghiệm và đánh giá thuật toán . . . . .	45
2.4	Kết luận Chương 2 . . . . .	61
<b>CHƯƠNG 3. PHƯƠNG PHÁP RÚT GỌN THUỘC TÍNH TRONG BẢNG QUYẾT ĐỊNH THEO TIẾP CẬN TÔPÔ MỜ TRỰC CẢM</b>		<b>62</b>
3.1	Mở đầu . . . . .	62
3.2	Đề xuất cấu trúc tôpô mờ trực cảm . . . . .	63
3.3	Đề xuất độ đo tương đồng của hai tôpô mờ trực cảm . . . . .	67
3.4	Giảm thuộc tính trong bảng quyết định theo tiếp cận tôpô mờ trực cảm	68
3.4.1	Đề xuất thuật toán tìm reduct trong bảng quyết định theo phương pháp filter, sử dụng cấu trúc tôpô mờ trực cảm . . . . .	68
3.4.2	Đề xuất thuật toán tìm reduct trong bảng quyết định theo phương pháp lai ghép filter - wrapper, sử dụng cấu trúc tôpô mờ trực cảm . . . . .	71

3.4.3	Thực nghiệm và đánh giá các thuật toán . . . . .	75
3.5	Kết luận Chương 3 . . . . .	89
<b>CHƯƠNG 4. PHƯƠNG PHÁP RÚT GỌN THUỘC TÍNH TRONG BẢNG</b>		
<b>QUYẾT ĐỊNH THEO TIẾP CẬN TÔPÔ HAUSDORFF</b>		<b>90</b>
4.1	Mở đầu . . . . .	90
4.2	Đề xuất cấu trúc tôpô từ không gian xấp xỉ mờ ngưỡng $\beta$ . . . . .	91
4.3	Đề xuất cấu trúc tôpô Hausdorff . . . . .	97
4.4	Giảm thuộc tính trong bảng quyết định theo tiếp cận tôpô Hausdorff .	98
4.4.1	Đề xuất thuật toán tìm reduct trong bảng quyết định theo phương pháp lai ghép filter - wrapper, sử dụng cấu trúc tôpô Hausdorff . . . . .	98
4.4.2	Thực nghiệm và đánh giá thuật toán . . . . .	101
4.5	Kết luận Chương 4 . . . . .	117
<b>KẾT LUẬN</b>		<b>118</b>
<b>DANH MỤC CÁC CÔNG TRÌNH NGHIÊN CỨU</b>		<b>120</b>
<b>TÀI LIỆU THAM KHẢO</b>		<b>121</b>

## DANH MỤC CÁC THUẬT NGỮ, CÁC CHỮ VIẾT TẮT

IS	Hệ thống tin (Information System)
DT	Bảng quyết định (Decision Table)
FN	False Negative (Phủ định sai)
FP	False Positive (Khẳng định sai)
TN	True Negative (Phủ định đúng)
TP	True Positive (Khẳng định đúng)
Base	Cơ sở
Subbase	Cơ sở con
IF-base	Cơ sở mờ trực cảm (Intuitionistic Fuzzy Base)
IF-subbase	Cơ sở con mờ trực cảm (Intuitionistic Fuzzy Subbase)
FRS	Tập thô mờ (Fuzzy Rough Set)
IFRS	Tập thô mờ trực cảm (Intuitionistic Fuzzy Rough Set)
IFS	Tập mờ trực cảm (Intuitionistic Fuzzy Set)
IFT	Tôpô mờ trực cảm (Intuitionistic Fuzzy Topology)
NRS	Tập thô lân cận (Neighborhood Rough set)
PRS	Tập thô xác suất (Probability Rough set)
VPRS	Tập thô điều chỉnh chính xác (Variable Precision Rough Set)
IFIE	Entropy thông tin mờ trực cảm (Intuitionistic Fuzzy Information Entropy)
IFD	khoảng cách mờ trực cảm (Intuitionistic Fuzzy Distance)
Hybrid	lai ghép
k-NN	k- láng giềng gần nhất (k - Nearest Neighbor)
SVM	Máy vector hỗ trợ (Support Vector Machine)



**DANH MỤC CÁC KÝ HIỆU**

C	Tập thuộc tính điều kiện
D	Tập thuộc tính quyết định
U	Tập đối tượng
$O$	Big-O
$\mathbb{R}$	Tập số thực
T	Thời gian thực hiện của mô hình phân lớp
M	Ma trận quan hệ
$M^T$	Ma trận chuyển vị
Model	Mô hình phân lớp
R	Quan hệ tương đương
$R^{\geq}$	Quan hệ ưu tiên
CLS	Miền đóng
INT	Miền trong
ACC	Độ chính xác (Accuracy)
$W_A$	wrapper thuộc tính (wrapper attribute)
$W_{\delta}$	wrapper theo giá trị delta
RAW	Dữ liệu ban đầu
C	Số lượng các thuộc tính điều kiện trong tập C
U	Số lượng các đối tượng trong tập U

## DANH MỤC CÁC HÌNH VẼ

2.1	Tác động của $\delta$ tới số lượng phần tử và khả năng phân lớp trên mô hình phân lớp SVM . . . . .	46
2.2	Tác động của $\delta$ tới số lượng phần tử và khả năng phân lớp trên mô hình phân lớp KNN . . . . .	47
2.3	Sự tương quan về số lượng phần tử và khả năng phân lớp chính xác của reduct thu được từ các thuật toán trên mô hình phân lớp SVM . . .	55
2.4	Sự tương quan về số lượng phần tử và khả năng phân lớp chính xác của reduct thu được từ các thuật toán trên mô hình phân lớp KNN . . .	56
3.1	Tập rút gọn thu được từ thuật toán F_IFT . . . . .	77
3.2	Sự tương quan về số lượng phần tử và khả năng phân lớp chính xác của reduct thu được từ các thuật toán so với F_IFT trên mô hình phân lớp KNN. . . . .	78
3.3	Sự tương quan về số lượng phần tử và khả năng phân lớp chính xác của reduct thu được từ các thuật toán so với F_IFT trên mô hình phân lớp SVM. . . . .	79
3.4	Biểu đồ đánh giá sự tương quan về số lượng phần tử reduct (trái) và chi phí tính toán (phải) với số lượng thuộc tính ban đầu của thuật toán F_IFT so với các thuật toán khác . . . . .	82
3.5	Sự tương quan về số lượng phần tử và khả năng phân lớp chính xác của reduct thu được từ các thuật toán so với FW_IFT trên mô hình phân lớp KNN. . . . .	84
3.6	Sự tương quan về số lượng phần tử và khả năng phân lớp chính xác của reduct thu được từ các thuật toán so với FW_IFT trên mô hình phân lớp SVM. . . . .	85

3.7	Biểu đồ đánh giá sự tương quan về số lượng phần tử reduct (trái) và chi phí tính toán (phải) với số lượng thuộc tính ban đầu của thuật toán FW_IFT so với các thuật toán khác trên mô hình phân lớp KNN . . . .	88
3.8	Biểu đồ đánh giá sự tương quan về số lượng phần tử reduct (trái) và chi phí tính toán (phải) với số lượng thuộc tính ban đầu của thuật toán FW_IFT so với các thuật toán khác trên mô hình phân lớp SVM . . . .	88
4.1	Biểu đồ phân tích sự tương quan giữa số lượng phần tử và khả năng phân lớp chính xác của reduct tại mỗi giá trị $\beta$ trên mô hình phân lớp SVM. . . . .	102
4.2	Biểu đồ phân tích sự tương quan giữa số lượng phần tử và khả năng phân lớp chính xác của reduct tại mỗi giá trị $\beta$ trên mô hình phân lớp KNN. . . . .	103
4.3	Biểu đồ phân tích sự tương quan giữa chi phí tính toán của thuật toán và $ U $ (left), giữa chi phí tính toán của thuật toán và $ C $ (right). . . . .	105
4.4	Biểu đồ phân tích sự tương quan giữa số lượng phần tử và khả năng phân lớp chính xác của reduct của mỗi thuật toán trên mô hình phân lớp SVM. . . . .	108
4.5	Biểu đồ phân tích sự tương quan giữa số lượng phần tử và khả năng phân lớp chính xác của reduct của mỗi thuật toán trên mô hình phân lớp KNN. . . . .	109

## DANH MỤC CÁC BẢNG BIỂU

1.1	Các toán tử T-chuẩn và T-đối chuẩn . . . . .	13
1.2	Các toán tử kéo theo chuẩn và đối chuẩn . . . . .	13
1.3	Mô tả cấu trúc bảng quyết định số . . . . .	18
1.4	Ma trận làm lẫn nhị phân . . . . .	23
1.5	Tổng hợp phương pháp giảm thuộc tính theo độ phụ thuộc . . . . .	28
1.6	Tổng hợp phương pháp giảm thuộc tính theo độ không chắc chắn . . .	29
1.7	Tổng hợp phương pháp giảm thuộc tính theo khoảng cách . . . . .	30
1.8	Tổng hợp phương pháp xây dựng tô pô theo tiếp cận RS . . . . .	31
2.1	Độ phức tạp của thuật toán IFD . . . . .	43
2.2	Bảng mô tả các tập dữ liệu thực nghiệm . . . . .	48
2.3	Mô tả sự tương quan về số lượng phần tử và khả năng phân lớp chính xác của reduct tại hai giai đoạn wrapper trên mô hình phân lớp SVM .	50
2.4	Mô tả sự tương quan về số lượng phần tử và khả năng phân lớp chính xác của reduct tại hai giai đoạn wrapper trên mô hình phân lớp KNN .	50
2.5	Mô tả số lượng phần tử thu được của reduct thu được từ các thuật toán	52
2.6	So sánh khả năng phân lớp của các reduct trên mô hình phân lớp SVM	52
2.7	So sánh khả năng phân lớp của các reduct trên mô hình phân lớp KNN	53
2.8	Mô tả chi phí tính toán của các thuật toán . . . . .	54
2.9	Mô tả reduct thu được từ các thuật toán . . . . .	57
3.1	Mô tả dữ liệu thực nghiệm . . . . .	76
3.2	So sánh số lượng phần tử của các reduct thu được từ các thuật toán theo tiếp cận filter . . . . .	80
3.3	So sánh khả năng phân lớp chính xác của reduct thu được từ các thuật toán theo tiếp cận filter trên mô hình phân lớp KNN . . . . .	80

3.4	So sánh khả năng phân lớp chính xác của reduct thu được từ các thuật toán theo tiếp cận filter trên mô hình phân lớp SVM . . . . .	81
3.5	Tập rút gọn thu được từ thuật toán FW_IFT trên mô hình phân lớp SVM	83
3.6	Tập rút gọn thu được từ thuật toán FW_IFT trên mô hình phân lớp KNN	86
3.7	So sánh số lượng phần tử của các reduct thu được từ các thuật toán theo tiếp cận filter - wrapper trên mô hình phân lớp SVM và KNN . .	87
3.8	So sánh khả năng phân lớp của các reduct thu được từ các thuật toán theo tiếp cận filter - wrapper trên mô hình phân lớp SVM và KNN . .	87
4.1	Mô tả các tập dữ liệu thực nghiệm . . . . .	104
4.2	So sánh số lượng phần tử của reduct thu được từ các thuật toán . . . .	106
4.3	So sánh khả năng phân lớp chính xác của reduct thu được từ các thuật toán trên mô hình phân lớp SVM . . . . .	107
4.4	So sánh khả năng phân lớp chính xác của reduct thu được từ các thuật toán trên mô hình phân lớp KNN . . . . .	110
4.5	So sánh chi phí tính toán của các thuật toán . . . . .	111
4.6	Mô tả reduct thu được từ các thuật toán . . . . .	112

## MỞ ĐẦU

### Tính cấp thiết của đề tài luận án

Giảm thuộc tính [1]–[3] hay chọn lọc thuộc tính là bước tiền xử lý dữ liệu quan trọng, được ứng dụng rộng 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]. Giảm 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. Giảm 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 RS cổ điển (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]. Giảm thuộc tính là một trong những ứng dụng quan trọng của mô hình RS cổ điển, 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 RS cổ điển, nhiều phương pháp đo lường độ quan trọng của thuộc tính được đề xuất để tìm tập thuộc tính rút gọn. Bên cạnh đó, không gian tôpô cũng là một khái niệm quan trọng trong mô hình RS cổ điển [18], [19]. Khái niệm tôpô theo tiếp cận RS cũng được Pawlack giới thiệu vào năm 1988 và nhận được nhiều quan tâm của các nhà nghiên cứu [4], [20].

Hơn ba thập kỷ vừa qua, hướng giảm thuộc tính theo tiếp cận RS [14] đã và đang thu hút được sự quan tâm của nhiều nhà nghiên cứu. Các kết quả nghiên cứu cho thấy phương pháp giảm thuộc tính theo tiếp cận RS 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 giảm 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 giảm thuộc tính trực tiếp trên các bảng quyết định gốc mà không phải qua quá trình rời rạc hóa dữ liệu.

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

### 1. RS mờ (Fuzzy Rough Set - FRS)

Mô hình RS mờ [21], [22] sử dụng khái niệm tương tự thay cho khái niệm không phân biệt được trong mô hình RS cổ điển truyền thống. Do đó, chúng ta không cần phải rời rạc hóa dữ liệu mà vẫn đánh giá chính xác sự tương quan của các đối tượng trong một tập. Cho đến nay, các hướng nghiên cứu giảm thuộc tính theo tiếp cận RS 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. RS 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 khả năng phân lớp cho các reduct 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 khả năng phân lớp ban đầu thấp. Gần đây, các công bố điển hình về giảm thuộc tính theo tiếp cận IFRS gồm có: phương pháp giảm 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 giảm thuộc

tính trực tiếp trên bảng quyết định số gồm có: luận án tiến sĩ của tác giả Cao Chính Nghĩa [3] nghiên cứu giảm 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 reduct 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 reduct 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 giảm 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 gọn thu được theo tiếp cận FRS còn chưa hiệu quả về số lượng phần tử và khả năng phân lớp trên các bộ dữ liệu nhiễu do không gian xấp xỉ mờ là chưa đủ để mô tả sự tương quan của các đối tượng trong một tập. Phương pháp giảm thuộc tính theo tiếp cận IFRS [15], [36] trên thế giới hiện nay còn chưa hiệu quả về số lượng phần tử của reduct và chi phí tính toán 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 giảm thuộc tính theo tiếp cận IFRS hiệu quả về thời gian, số lượng phần tử, cải thiện khả năng phân lớp đối với các tập dữ liệu nhiễu.

Bên cạnh các phương pháp giảm thuộc tính theo tiếp cận RS và RS mở rộng như đã được trình bày bên trên. Phương pháp giảm 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 RS cổ điển [37], [38].

Theo tiếp cận tôpô, khái niệm reduct theo cấu trúc tôpô lần đầu tiên được giới thiệu bởi Lashin và các cộng sự [37]. Để giảm 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 RS gồm có, các phương pháp xây dựng tôpô từ không gian xấp xỉ của RS [38], [40]–[42], các phương pháp xây dựng tôpô từ các phép toán xấp xỉ của RS [43]. Bên cạnh đó, sự tương quan của mô hình lý thuyết tôpô và RS 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 RS cổ điển với các phép toán định miền của mô hình lý thuyết tôpô [48]. Trên cơ sở đó, nhiều cấu trúc tôpô được đề xuất dựa trên việc xây dựng lại các phép toán xấp xỉ của RS [20], [45], [49]. Hơn nữa, dựa trên sự tương quan này, một số phương pháp cấu trúc lại mô hình RS 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 RS và RS từ tôpô nhằm nhấn mạnh sự tương quan lý thuyết chặt chẽ của hai mô hình này. Gần đây, Xie và các cộng sự [52] đã đề xuất phương pháp giảm 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 giảm 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 giảm thuộc tính.

### **Mục tiêu nghiên cứu**

Xuất phát từ những vấn đề còn tồn tại của các phương pháp giảm thuộc tính hiện nay, luận án đặt ra mục tiêu nghiên cứu một số phương pháp giảm thuộc tính trong bảng quyết định theo hướng tiếp cận RS mờ trực cảm và tôpô suy rộng, cụ thể như sau:

1) *Nghiên cứu phương pháp giảm thuộc tính theo tiếp cận RS mờ trực cảm*: Với phương pháp giảm thuộc tính theo tiếp cận RS mờ trực cảm, *vấn đề nghiên cứu trước tiên* là cần tìm hiểu cách thức mô tả sự tương quan 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 giảm thuộc tính hiệu quả trên các bộ dữ liệu nhiễu và có số chiều lớn trong thực tiễn.

2) *Nghiên cứu phương pháp giảm thuộc tính theo tiếp cận tô pô đại số*: Với phương pháp giảm thuộc tính theo tiếp cận tô pô đại số, *vấn đề nghiên cứu trước tiên* là cần tìm hiểu các phương pháp xây dựng cấu trúc tô pô, tìm hiểu các tính chất cơ sở của tô pô sao cho có thể đánh giá tô pô 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 tô pô nhằm xây dựng các phương pháp đánh giá, nhận diện độ quan trọng của thuộc tính, định nghĩa reduct thông qua cấu trúc tô pô, ứng dụng xây dựng thuật toán giảm 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 giảm 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 giảm thuộc tính trong bảng quyết định theo tiếp cận RS và tô pô đại số gồm có:

- Khảo sát các khái niệm cơ bản về RS, các độ đo được sử dụng để đánh giá độ quan trọng của thuộc tính và các phương pháp xây dựng thuật toán giảm thuộc tính theo tiếp cận Heuristic.
- Khảo sát các khái niệm cơ bản về tô pô theo tiếp cận RS, tô pô thu từ không gian xấp xỉ, tô pô thu từ quan hệ của các phép toán xấp xỉ, tính khả li trong không gian tô pô và tô pô rút gọn.

### **Phạm vi nghiên cứu**

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

- Nghiên cứu các mô hình RS mở rộng trên nền tập mờ và tập mờ trực cảm, ứng

dụng xây dựng thuật toán giảm thuộc tính trong bảng quyết định số.

- Nghiên cứu cấu trúc tôpô theo tiếp cận RS 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 giảm 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, RS, tập mờ, tập mờ trực cảm và entropy Shannon.

- *Góc độ nghiên cứu thực nghiệm:* các thuật toán được cài đặt và thực nghiệm trên các bộ dữ liệu từ UCI<sup>1</sup>. 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 reduct. So sánh chất lượng reduct từ thuật toán đề xuất với các thuật toán khác nhằm củng cố giá 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 giảm thuộc tính, phân loại các phương pháp giảm thuộc tính. Trình bày các khái niệm cơ bản về hệ thống tin, bảng quyết định và reduct. Trình bày các khái niệm cơ bản về mô hình RS truyền thống, không gian tôpô và tập mờ trực cảm. Trên cơ sở đó, luận án trình bày các phương pháp giảm thuộc tính theo tiếp cận độ đo và tiếp cận tôpô. Trình bày các chỉ số và phương pháp đánh giá chất lượng mô hình phân lớp dữ liệu. 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 giảm thuộc tính theo tiếp cận RS mờ trực cảm bao gồm các đề xuất chính như sau:

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

1) Đề xuất độ đo khoảng cách mờ trực cảm trên cơ sở mở rộng độ đo khoảng cách tri thức.

2) Đề xuất thuật toán reduct cho bảng quyết định theo phương pháp lai ghép filter - wrapper, sử dụng độ đo khoảng cách mờ trực cảm.

*Chương 3.* Luận án trình bày phương pháp giảm thuộc tính theo tiếp cận tôpô mờ trực cảm bao gồm các đề xuất chính như sau:

1) Đề xuất cấu trúc tôpô mờ trực cảm

2) Đề xuất độ đo đánh giá sự tương đồng của hai tôpô mờ trực cảm

3) Đề xuất thuật toán tìm reduct cho bảng quyết định theo phương pháp filter, sử dụng cấu trúc tôpô mờ trực cảm.

4) Đề xuất thuật toán tìm reduct cho bảng quyết định theo phương pháp lai ghép filter - wrapper, sử dụng cấu trúc tôpô mờ trực cảm.

*Chương 4.* Luận án trình bày phương pháp giảm thuộc tính theo tiếp cận tôpô Hausdorff bao gồm các đề xuất chính như sau:

1) Đề xuất cấu trúc tôpô theo tiếp cận RS

2) Đề xuất cấu trúc tôpô Hausdorff

3) Đề xuất thuật toán tìm reduct cho bảng quyết định theo phương pháp lai ghép filter - wrapper, sử dụng cấu trúc 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 RS 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 [53].

Có ba mô hình tiếp cận chính để xây dựng các thuật toán rút gọn thuộc tính gồm có mô hình filter, wrapper và embed thuộc tính. Trong đó tiếp cận Filter là tiếp cận được sử dụng rộng rãi nhất. Cách tiếp cận Wrapper thường được gắn với một mô hình phân lớp cụ thể để đánh giá tập thuộc tính rút gọn ứng viên tốt nhất về khả năng phân lớp. Tuy nhiên với  $C$  thuộc tính ta cần đánh giá tới  $2^C$  tập con thuộc tính, do đó cách tiếp cận này có chi phí tính toán rất lớn.

Gần đây, một số mô hình lai ghép filter và wrapper, trong đó kỹ thuật wrapper được cải tiến để đánh giá trên các reduct ứng viên có số lượng nhỏ. Với cách tiếp cận embed, việc đánh giá thuộc tính hay tập thuộc tính thường được gắn trực tiếp vào một mô hình cụ thể, ví dụ mô hình cây quyết định thì luật thu được từ mô hình này thường ngắn gọn hơn so với luật quyết định của bảng dữ liệu huấn luyện. Tuy nhiên tập thuộc tính rút gọn thu được theo cách tiếp cận Wrapper hay Embed chỉ phù hợp với một mô hình học máy cụ thể, nếu sử dụng reduct đó sang mô hình học máy khác có thể sẽ

không hiệu quả. Do đó, cho đến nay tiếp cận Filter vẫn được các nhà khoa học lựa chọn nhiều hơn cả do mục tiêu đánh giá tập thuộc tính rút gọn được khái quát theo tiêu chí bảo toàn thông tin của reduct so với tập dữ liệu gốc.

Để xây dựng các thuật toán rút gọn thuộc tính, mô hình chung cho các thuật toán gồm có hai thành phần chính như sau:

- Tiêu chuẩn chọn lọc thuộc tính: bao gồm các phương pháp đánh giá độ quan trọng của thuộc tính như dựa trên độ đo được định nghĩa hay cấu trúc tôpô được định nghĩa.

- Phương pháp tìm kiếm: chủ yếu dựa vào tiếp cận Heuristic như tìm kiếm thuộc tính quan trọng dựa trên tập thuộc tính ban đầu, dựa trên tập thuộc tính lõi hay xuất phát từ tập rỗng.

Các thuộc tính điều kiện trong bảng quyết định có thể được chia làm 03 nhóm có tính chất như sau:

- Tính độc lập (Independent): Bao gồm các thuộc tính điều kiện không tương quan với các thuộc tính điều kiện khác nhưng tương quan với thuộc tính quyết định.

- Tính dư thừa (Redundant): Bao gồm các thuộc tính điều kiện có tương quan với các thuộc tính điều kiện khác nhưng không tương quan với thuộc tính quyết định.

- Tính không phù hợp: Bao gồm các thuộc tính điều kiện không tương quan với các thuộc tính điều kiện khác mà cũng không tương quan với thuộc tính quyết định.

Bên cạnh phương pháp rút gọn thuộc tính truyền thống đã được phát triển hơn ba thập kỉ vừa qua, trong những năm gần đây nhiều nhà nghiên cứu đề xuất cách tiếp cận rút gọn thuộc tính theo tiếp cận tôpô đại số, cách tiếp cận này nhận được nhiều sự quan tâm của cộng đồng các nhà nghiên cứu lý thuyết về RS, tôpôlogy và tôpô đại số do một số công trình công bố đã chỉ ra sự tương quan khá gần gũi giữa mô hình RS cổ điển và không gian tôpô, đặc biệt về ý bài toán rút gọn thuộc tính có liên quan đến khái niệm bất biến của không gian tôpô dưới góc nhìn của đại số trừu tượng.

## 1.2. Các khái niệm cơ bản

### 1.2.1. Hệ thông tin và mô hình RS truyền thống

Hệ thông tin là công cụ biểu diễn tri thức dưới dạng một bảng dữ liệu gồm  $A$  cột ứng với  $A$  thuộc tính và  $U$  hàng ứng với  $U$  đối tượng. Một cách hình thức, hệ thông tin được định nghĩa như sau.

**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  $f : U \times A \rightarrow V_a$  là hàm thông tin,  $\forall a \in A, u \in U$  ta có  $f(u, a) \in V_a$ .

Một lớp đặc biệt của các hệ thông tin có vai trò quan trọng trong nhiều ứng dụng là bảng quyết định. Bảng quyết định là một hệ thông tin  $DT$  với tập thuộc tính  $A$  được chia thành hai tập khác rỗng rời nhau  $C$  và  $D$ , lần lượt được gọi là tập thuộc tính điều kiện và tập thuộc tính quyết định. Tức là  $DT = (U, C, D, f)$  với  $C \cap D = \emptyset$ .

Xét bảng quyết định  $DT = (U, C, D, f)$  với giả thiết  $\forall u \in U, \forall d \in D$  đầy đủ giá trị, nếu tồn tại  $u \in U$  và  $c \in C$  sao cho  $c(u)$  thiếu giá trị thì  $DT$  được gọi là bảng quyết định không đầy đủ, trái lại  $DT$  được gọi là bảng quyết định đầy đủ.

**Định nghĩa 1.2** (Quan hệ bất khả phân biệt được [14]). Xét bảng quyết định  $DT = (U, C, D, f)$ . Khi đó, quan hệ của các đối tượng trong  $U$  trên tập thuộc tính  $B \subseteq C$  ký hiệu bởi  $ID(B)$ , được định nghĩa bởi:

$$ID(B) = \{(u, v) \in U \mid \forall a \in B, a(u) = a(v)\} \quad (1.1)$$

Rõ ràng  $ID(B)$  là một quan hệ tương đương trên  $U$ . Nếu  $(u, v) \in IND(P)$  thì hai đối tượng  $u$  và  $v$  không phân biệt được bởi các thuộc tính trong  $P$ . Quan hệ tương đương  $ID(B)$  xác định một phân hoạch trên  $U$ , ký hiệu là  $U/ID(B)$  hay  $U/B$ . Ký hiệu lớp tương đương trong phân hoạch  $U/P$  chứa đối tượng  $u$  là  $[u]_B$ , khi đó  $[u]_B = \{v \in U \mid (u, v) \in ID(B)\}$ .

**Định nghĩa 1.3** (Phân hoạch của thuộc tính [18], [54]). 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  $\forall u \in U, [u]_P = [u]_Q$ .

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

**Định nghĩa 1.4** (Mô hình RS truyền thống [14], [18], [54]). Trong mô hình RS cổ điển truyền thống, để biểu diễn tập  $X \subseteq U$  trên cơ sở tri thức của tập thuộc tính  $B$  theo khái niệm RS, 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:

$$\underline{B}(X) = \{u \in U \mid [u]_B \subseteq X\} \quad (1.2)$$

$$\overline{B}(X) = \{u \in U \mid [u]_B \cap X \neq \emptyset\} \quad (1.3)$$

Khi đó,  $\underline{B}(X)$  là tập các phần tử trong  $U$  chắc chắn thuộc  $X$  (xác định thuộc), còn  $\overline{B}(X)$  là tập các phần tử của  $U$  có thể thuộc  $X$  dựa trên tập thuộc tính  $B$ . Trên cơ sở đó, các tập *không xác định* và *tập xác định không thuộc* được định nghĩa như sau:

$$BN_B(X) = \overline{B}(X) - \underline{B}(X) \quad (1.4)$$

$$U - \overline{B}(X) \quad (1.5)$$

Trong đó:  $BN_B(X)$  được gọi là *miền biên của  $X$  theo  $B$*  là tập các đối tượng có thể thuộc hoặc không thuộc (không xác định) trong  $X$ , còn  $U - \overline{B}(X)$  là *miền ngoài của  $X$*  là tập các đối tượng chắc chắn không thuộc (không xác định) trong  $X$ . Trong trường hợp  $BN_B(X) = \emptyset$  thì  $X$  được gọi là tập xác định, ngược lại  $X$  được gọi là RS (Rough Set - RS).

**Mệnh đề 1.1** (Các tính chất cơ bản của RS [14], [18]). Cho bảng quyết định  $DS =$



$(U, C, D, f)$ , với  $X, Y \subseteq U$  và  $A \subseteq C$ . Khi đó:

- 1)  $\underline{A}(\emptyset) = \bar{A}(\emptyset)$ ,  $\underline{A}(U) = \bar{A}(U)$ ;
- 2)  $\underline{A}(X) \subseteq X \subseteq \bar{A}(X)$ ;
- 3)  $\underline{A}(X \cup Y) \supseteq \underline{A}(X) \cup \underline{A}(Y)$ ;
- 4)  $\underline{A}(X \cap Y) = \underline{A}(X) \cap \underline{A}(Y)$ ;
- 5)  $\bar{A}(X \cup Y) = \bar{A}(X) \cup \bar{A}(Y)$ ;
- 6)  $\bar{A}(X \cap Y) \subseteq \bar{A}(X) \cap \bar{A}(Y)$ ;
- 7)  $\underline{A}(U - X) \subseteq U - \bar{A}(X)$ ;
- 8)  $\bar{A}(U - X) \subseteq U - \underline{A}(X)$ ;
- 9)  $\underline{A}(\underline{A}(X)) = \bar{A}(\underline{A}(X)) = \underline{A}(X)$ ;
- 10)  $\bar{A}(\bar{A}(X)) = \underline{A}(\bar{A}(X)) = \bar{A}(X)$

### 1.2.2. RS mờ trực cảm

**Định nghĩa 1.5** (Tập mờ [55]). Cho  $U$  là tập hữu hạn khác rỗng các đối tượng, tập mờ  $A$  xác định trên  $U$  có dạng:  $A = \{\langle x, \mu_A \rangle \mid x \in U\}$ . Với  $\mu_A : U \rightarrow [0, 1]$  thỏa mãn  $0 \leq \mu_A(x) \leq 1$ . Trong đó  $\mu_A(x)$  được gọi là độ thành viên của phần tử  $x$  trong  $A$ .

**Mệnh đề 1.2** (Tính chất và phép toán cơ bản của FS[55]). Cho  $A, B \subseteq F(U)$  với  $F(U)$  là họ các tập mờ trên  $U$ . Khi đó:

- 1)  $A = B$  nếu  $A(x) = B(x)$ ,  $\forall x \in U$
- 2)  $A \subseteq B$  nếu  $A(x) \leq B(x)$ ,  $\forall x \in U$
- 3)  $(A \cup B)(x) = \max\{A(x), B(x)\}$ ,  $\forall x \in U$
- 4)  $(A \cap B)(x) = \min\{A(x), B(x)\}$ ,  $\forall x \in U$
- 5)  $\bar{A}(x) = 1 - A(x)$ ,  $\forall x \in U$

Để phát triển các phép toán trong môi trường dữ liệu mờ, các toán tử logic rõ cũng được mở rộng cho môi trường dữ liệu mờ. Bảng 1.1 mô tả chi tiết một số toán tử *chuẩn* T-norm và *đổi chuẩn* T-conorm tương ứng cho phép toán hợp và giao. Bảng 1.2 mô

tả các chi tiết các toán tử kéo theo I-norm và I-conorm. Trong đó :  $x = \sqrt{1 - a^2}, y = \sqrt{1 - b^2}, p = \sqrt{2a - a^2}, q = \sqrt{2b - b^2}$ .

**Bảng 1.1:** Các toán tử T-chuẩn và T-đối chuẩn

T-norm	T-conorm
$T_m(a, b) = \min\{a, b\}$	$S_m(a, b) = \max\{a, b\}$
$T_p(a, b) = ab$	$S_p(a, b) = a + b - ab$
$T_L(a, b) = \max\{a + b - 1, 0\}$	$S_L(a, b) = \min\{a + b, 1\}$
$T_{\cos}(a, b) = \max\{ab - xy, 0\}$	$S_{\cos}(a, b) = \min\{a + b - ab + pq, 1\}$

**Bảng 1.2:** Các toán tử kéo theo chuẩn và đối chuẩn

I-norm	I-conorm
$\theta_m(a, b) = \begin{cases} 1, & a \leq b \\ b, & a > b \end{cases}$	$\sigma_m(a, b) = \begin{cases} 0, & a \geq b \\ b, & a < b \end{cases}$
$\theta_p(a, b) = \begin{cases} 1, & a = 0 \\ \min\{1, \frac{b}{a}\}, & \text{otherwise} \end{cases}$	$\sigma_p(a, b) = \begin{cases} 1, & a = 0; \\ \max\{0, \frac{b-a}{1-a}\}, & \text{otherwise.} \end{cases}$
$\theta_L(a, b) = \min\{b - a + 1, 1\}$	$\sigma_L(a, b) = \min\{0, b - a\}$
$\theta_{\cos}(a, b) = \begin{cases} 1, & a \leq b \\ ab + xy, & a > b \end{cases}$	$\sigma_{\cos}(a, b) = \begin{cases} 0, & a > b \\ a + b - ab - pq, & a \leq b \end{cases}$

Dựa trên nền tập mờ, để biểu diễn sự tương quan của các đối tượng trong cùng một tập được trơn hơn, phản ánh rõ nét độ tương tự giữa các đối tượng. Đặc biệt là quan hệ giữa các đối tượng thuộc  $\mathbb{R}$ . Khi đó, quan hệ tương đương của mô hình RS truyền thống được mở rộng trên nền tập mờ bằng quan hệ tương đương mờ theo định nghĩa sau đây:

**Định nghĩa 1.6** (Quan hệ tương đương mờ [56]). Xét  $R$  là quan hệ tương đương trên  $U$  không rỗng. Khi đó  $R$  được gọi là quan hệ tương đương mờ nếu các tiêu chuẩn sau đây được thỏa mãn:

- (C1) Có tính phản xạ: nếu  $R(x, x) = 1$  với mọi  $x \in U$ ;
- (C2) Có tính đối xứng nếu  $R(x, y) = R(y, x)$  với mọi  $x, y \in U$ ;
- (C3) Có tính bắc cầu nếu  $T(R(x, y), R(y, z)) \leq R(x, z)$  với mọi  $x, y, z \in U$ .

**Định nghĩa 1.7** (Ma trận quan hệ). Cho  $R$  là một quan hệ tương mờ trên  $U$ , khi đó quan hệ giữa các đối tượng trong  $U$  theo quan hệ  $R$  có thể được biểu diễn bởi ma trận quan hệ  $M = [i, j]_{|U| \times |U|}$ . Đây là ma trận vuông có số lượng phần tử  $|U| \times |U|$ . Trong đó  $|U|$  là số các đối tượng trong  $U$  và  $i, j$  là chỉ số của phần tử tại hàng  $i$  và cột  $j$  trên ma trận  $M$ . Khi đó mỗi giá trị tại hàng  $i$  cột  $j$  cho biết độ tương tự giữa đối tượng  $i$  và  $j$  trong  $U$ .

Trong đó  $T$  và  $I$  tương ứng với toán tử  $T$  – norm và toán tử kéo theo  $I$  – norm được chọn trong Bảng 1.1 và Bảng 1.2.

Để có thể mô tả thông tin đầy đủ hơn trong các trường hợp dữ liệu phức tạp như thông tin về phiếu tín nhiệm của một ứng viên cần được đánh giá. Trong đó có các thành phần về tỉ lệ tín nhiệm và tỉ lệ bất tín nhiệm. Khi đó khái niệm về RS mờ trực cảm ra đời và được định nghĩa như sau:

**Định nghĩa 1.8** (Tập mờ trực cảm [57]). Cho  $U$  là tập không rỗng các đối tượng, tập mờ trực cảm  $X$  trên  $U$  được xác định bởi:

$$X = \{ \langle x, \mu_X(x), \nu_X(x) \rangle \mid x \in U \} \quad (1.6)$$

Trong đó,  $\mu_X(x) \in [0, 1]$  là mức độ thành viên của  $x \in U$  với  $X$  và  $\nu_X(x) \in [0, 1]$  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:

$$X = \{ \langle x, \mu_Y(x), 1 - \mu_Y(x) \rangle \mid x \in U \} \quad (1.7)$$

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 1.9** (Phép toán cơ bản của IFS [36]). Xét  $P$  và  $Q$  là các tập mờ trực cảm xác định trên  $U$ . Khi đó hợp và giao của của  $P$  và  $Q$  được xác định như sau:

$$P \cup Q = \{ \langle x, \vee(\mu_P(x), \mu_Q(x)), \wedge(\nu_P(x), \nu_Q(x)) \rangle \} \quad (1.8)$$

$$P \cap Q = \{ \langle x, \wedge (\mu_P(x), \mu_Q(x)), \vee (v_P(x), v_Q(x)) \rangle \} \quad (1.9)$$

**Mệnh đề 1.3** (Quan hệ của hai IFS [36]). Xét  $P$  và  $Q$  là các tập mờ trực cảm xác định trên  $U$ :

- 1)  $P \subseteq Q$  khi và chỉ khi  $\mu_P(x) \leq \mu_Q(x)$  và  $v_Q(x) \geq v_P(x)$  với mọi  $x \in U$
- 2)  $P = Q$  khi và chỉ khi  $P \subseteq Q$  và  $Q \subseteq P$ .

Dựa trên nền tập mờ trực cảm, để biểu diễn sự tương quan của các đối tượng trong cùng một tập được chặt hơn. Đặc biệt là quan hệ giữa các đối tượng thuộc  $\mathbb{R}$  trên các miền giá trị có chứa nhiều. Khi đó, quan hệ tương đương của mô hình RS truyền thống được mở rộng trên nền tập mờ trực cảm bằng quan hệ tương đương mờ trực cảm theo định nghĩa sau đây:

**Định nghĩa 1.10** (Quan hệ mờ trực cảm [36]). Xét quan hệ  $R$  xác định trên  $U$  không rỗng. Khi đó  $R$  được gọi là quan hệ tương đương mờ trực cảm nếu:

- 1) Có tính phản xạ:  $\mu_{R(x,x)} = 1$  và  $v_{R(x,x)} = 0$  với mọi  $x \in U$ ;
- 2) Có tính đối xứng:  $\mu_{R(x,y)} = \mu_{R(y,x)}$  và  $v_{R(x,y)} = v_{R(y,x)}$  với mọi  $x, y \in U$ ;
- 3) Có tính bắc cầu:  $\mu_{R(x,z)} \geq \bigvee_{y \in U} [\mu_{R(x,y)} \wedge \mu_{R(y,z)}]$  và  $v_{R(x,z)} \leq \bigwedge_{y \in U} [v_{R(x,y)} \vee v_{R(y,z)}]$  với mọi  $x, y \in U$ .

**Định nghĩa 1.11** (Lực lượng của một tập mờ trực cảm [36]). Cho tập mờ trực cảm  $X$  xác định trên  $U$ , với  $U$  là tập không rỗng các đối tượng. Khi đó lực lượng của  $X$  được xác định như sau:

$$|X| = \sum_{i=1}^{|U|} \frac{1 + \mu_i - v_i}{2} \quad (1.10)$$

Dựa trên quan hệ mờ trực cảm và các toán tử logic mờ, mô hình RS mờ trực cảm được mở rộng theo định nghĩa sau:

**Định nghĩa 1.12** (Mô hình RS 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:

$$\underline{A}(x) = \bigwedge_{y \in U} I(R(x,y), A(y)) \quad (1.11)$$

$$\bar{A}(x) = \bigvee_{y \in U} T(R(x, y), A(y)) \quad (1.12)$$

### 1.2.3. Không gian tôpô

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

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

(T2)  $\tau$  có tính đóng dưới phép toán hợp bất kì.

(T3)  $\tau$  có tính đóng dưới phép toán giao hữu hạn.

Cặp  $(U, \tau)$  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 1.13** (Cơ sở (base) [58]). Cho  $U$  là tập không rỗng các đối tượng. Khi đó cơ sở (base) của tôpô  $\tau$  trên  $U$  là họ các tập con của  $U$  kí hiệu là  $B$  sao cho:

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

(2) Với 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 1.14** (Cơ sở con (subbase) [58]). Cho không gian tôpô  $(U, \tau)$ . Khi đó  $S \subseteq \tau$  được gọi là cơ sở con (subbase) của tôpô  $\tau$  nếu giao hữu hạn các tập con của  $S$  tạo thành cơ sở  $B$  của tôpô  $\tau$ .

**Định nghĩa 1.15** (Tôpô Hausdorff [37]). Cho không gian xấp xỉ  $(U, \tau)$ , 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  $V_x \cap V_y = \emptyset$ .

Để xây dựng tôpô, cách truyền thống là dựa trên cơ sở (base). Trong đó, mỗi phần tử của tôpô tương ứng là hợp các phần tử trong một tập con của cơ sở (base). Trong đó cơ sở base được tạo ra từ cơ sở con (subbase). Khi đó họ các tập con  $\beta \subseteq \tau$  được gọi là cơ sở của  $\tau$  khi và chỉ khi mọi tập mở không rỗng của  $U$  có thể được biểu diễn bởi hợp các phần tử của  $\beta$ . Họ các tập con  $S \subseteq \tau$  được gọi là cơ sở con khi và chỉ khi giao hữu hạn các phần tử của  $S$  tạo thành một cơ sở của  $\tau$ .

Để xác định các đối tượng có thể thuộc vào tập mục tiêu  $A$  với  $A \subseteq U$  dựa trên khái niệm tập đóng  $F$  với  $F \subseteq U$  được xác định như sau:

$$\bar{A} = \cap \{F \subseteq U : A \subseteq F\} \quad (1.13)$$

Để xác định các đối tượng chắc chắn thuộc vào tập mục tiêu  $A$  với  $A \subseteq U$  dựa trên khái niệm tập mở  $G$  với  $G \subseteq U$  được xác định như sau:

$$A^\circ = \cup \{G \subseteq U : G \subseteq A\} \quad (1.14)$$

**Mệnh đề 1.4** (Cấu trúc tôpô theo tiếp cận RS [38]). *Cho bảng quyết định  $DT = (U, C, D, f)$  và quan hệ tương đương  $R$  xác định trên  $U$ . Khi đó  $\tau = \{X \subseteq U \mid \underline{R}(X) = \bar{R}(X)\}$  là một tôpô trên  $U$ .*

**Định nghĩa 1.16** (Tôpô mờ trực cảm IFT). [20] *Cho  $\tau$  là họ các tập mờ trực cảm xác định trên tập không rỗng  $U$ . Khi đó  $\tau$  được gọi là tôpô mờ trực cảm nếu:*

- (T<sub>1</sub>)  $0_{IF}, 1_{IF} \in \tau$
- (T<sub>2</sub>)  $G_1 \cap G_2 \in \tau : G_1, G_2 \in \tau$
- (T<sub>3</sub>)  $\cup G_i \in \tau : \{G_i : G_i \in \tau, i \in I\}$

Khi đó, cặp  $(U, \tau)$  được gọi là không gian tôpô mờ trực cảm. Trong đó,  $0_{IF}$  và  $1_{IF}$  lần lượt là các tập mờ trực cảm nhỏ nhất và lớn nhất trên  $U$ .

#### 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 reduct 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 reduct nào của bảng quyết định. Thuộc tính rút gọn là thuộc tính xuất hiện trong một reduct

nào đó của bảng quyết định.

### 1.3. Một số công thức tính toán độ thành viên

Bảng quyết định 1.3 được biểu diễn bởi bộ  $DT = (U, C, D, f)$ . Trong đó  $C \cap D = \emptyset$ ,  $U$  là một tập không rỗng các đối tượng,  $C$  là tập không rỗng các thuộc tính điều kiện và  $D$  là thuộc tính quyết định. Hàm thông tin  $f_c$  xác định một giá trị trong  $V_c$  tương ứng với mỗi  $u \in U$  và  $c \in C$ , trong đó  $V_c$  là miền giá trị của thuộc tính  $c$  và  $V_c$  thuộc  $\mathbb{R}$ . Hàm thông tin  $f_D$  xác định một giá trị trong  $V_D$  với mỗi  $u \in U$ , trong đó  $V_D$  là miền giá trị của  $D$  và  $V_D$  thuộc  $\mathbb{N}$ .

**Bảng 1.3:** Mô tả cấu trúc bảng quyết định số

U	$a$	$b$	$c$	$d$	$e$	$f$	D
$u_1$	1.0	0.4	0.8	0.2	1.0	0.0	0
$u_2$	1.0	0.4	0.2	0.4	0.2	0.8	1
$u_3$	0.8	0.6	1.0	0.0	0.6	0.4	0
$u_4$	0.2	0.6	0.8	0.2	0.0	1.0	1
$u_5$	0.2	0.8	0.8	0.2	0.0	1.0	1
$u_6$	0.2	0.8	0.2	0.8	0.0	1.0	0

#### 1.3.1. Chuẩn hóa dữ liệu

Đối với các bảng quyết định có thuộc tính điều kiện miền giá trị số, các thuộc tính thường được chuẩn hóa để tăng hiệu quả cho các mô hình. Theo các kết quả khảo sát của [56] Sau đây là một số phương pháp chuẩn hóa dữ liệu được sử dụng phổ biến.

(1) Min-max normalization:

$$F(f_{c_k}(x_i)) = \frac{f_{c_k}(x_i) - \min_{c_k}}{\max_{c_k} - \min_{c_k}} (\max'_{c_k} - \min'_{c_k}) + \min'_{c_k} \quad (1.15)$$

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:

$$F(f_{c_k}(x_i)) = \frac{f_{c_k}(x_i) - \bar{c}_k}{\sigma_{c_k}} \quad (1.16)$$

Trong đó,  $\bar{c}_k$  và  $\sigma_{c_k}$  kí hiệu là giá trị trung bình và độ lệch chuẩn của thuộc tính  $c_k$ .

(3) Chuẩn hóa về thang đo hệ 10:

$$F(f_{c_k}(x_i)) = \frac{f_{c_k}(x_i)}{10^I} \quad (1.17)$$

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

### 1.3.2. Độ đo độ tương tự

Cho bảng quyết định  $DT = (U, C, D, f)$  với  $B \subseteq C$  và quan hệ tương đương mờ  $R$ . Khi đó  $R$  sẽ chia  $U$  thành các lớp tương đương mờ theo  $B$  gọi là phân hoạch mờ của  $B$  trên  $U$  kí hiệu là  $U/R_B$ . Trong đó  $U/R_B = \{[x_1]_{R_B}, [x_2]_{R_B}, \dots, [x_n]_{R_B}\}$ , với:  $[x_i]_{R_B} = (r_{i1}^B, r_{i2}^B, \dots, r_{in}^B)$ . Rõ ràng,  $[x_i]_{R_B}$  là một tập mờ trên  $R_B$ . ta có  $[x_i]_{R_B}(x_j) = R_B(x_i, x_j) = r_{ij}^B$ . Nếu  $R_B(x_i, x_j) = 1$ , nghĩa là  $x_j$  chắc chắn thuộc  $[x_i]_{R_B}$ ; Nếu  $R_B(x_i, x_j) = 0$ , thì  $x_j$  chắc chắn không thuộc  $[x_i]_{R_B}$ .

Lực lượng của tập mờ  $[x_i]_{R_B}$  được xác định bởi  $|[x_i]_{R_B}| = \sum_{j=1}^n R_B(x_i, x_j)$ . Ta có  $1 \leq |[x_i]_{R_B}| \leq n$  với  $n = |U|$ . Để xác định độ tương tự  $r_{ij}^B$ . Sau đây là một số công thức tính độ tương tự được sử dụng phổ biến.

(1) Khoảng cách [59] được xác định bởi

$$r_{ij}^B = 1 - \frac{1}{\sqrt[p]{C}} \Delta_p^B(x_i, x_j) \quad (1.18)$$

Trong đó  $h = |B|$ ,  $\Delta_p^B(x_i, x_j) = \sqrt[p]{\sum_{k=1}^h (f_{c_k}(x_i) - f_{c_k}(x_j))^p}$ . Khi  $p = 1, p = 2$ , và  $p = \infty, \Delta_p^B(x_i, x_j)$  tương ứng với khoảng cách của Manhattan, Euclidean, và Chebyshev.  $C$  là hệ số khoảng cách sao cho  $r_{ij}^B$  thuộc đoạn  $[0, 1]$ .



(2) Độ tương quan [60] được xác định bởi:

$$r_{ij}^B = \frac{\sum_{k=1}^h |f_{c_k}(x_i) - \bar{x}_i| |f_{c_k}(x_j) - \bar{x}_j|}{\sqrt{\sum_{k=1}^h (f_{c_k}(x_i) - \bar{x}_i)^2} \sqrt{\sum_{k=1}^h (f_{c_k}(x_j) - \bar{x}_j)^2}} \quad (1.19)$$

Trong đó:  $\bar{x}_i = \frac{1}{h} \sum_{c=1}^h f_{c_k}(x_i)$ ,  $\bar{x}_j = \frac{1}{h} \sum_{c=1}^h f_{c_k}(x_j)$ .

(3) Độ phân li [61] được xác định bởi:

$$r_{ij}^B = \bigwedge_{c=1}^h r_{ij}^{c_k} = \min_{c=1}^h r_{ij}^{c_k} \quad (1.20)$$

(4) Nhân đại số [62] được xác định bởi:

$$r_{ij}^B = \prod_{c=1}^h r_{ij}^{c_k} \quad (1.21)$$

(5) Phương pháp T-norm [63] được xác định bởi

$$r_{ij}^B = T r_{ij}^{c_k} \quad (1.22)$$

(6) Hàm Kernel [64], [65]:

(i) Gaussian kernel:

$$r_{ij}^B = \exp\left(-\frac{\|x_i - y_i\|_B^2}{2\delta^2}\right) \quad (1.23)$$

(ii) Exponential kernel:

$$r_{ij}^B = \exp\left(-\frac{\|x_i - y_i\|_B}{\delta}\right) \quad (1.24)$$

(iii) Rational quadratic kernel:

$$r_{ij}^B = 1 - \frac{\|x_i - y_i\|_B^2}{\|x_i - y_i\|_B \|x_i - y_i\|_B^2 + \delta} \quad (1.25)$$

(iv) Spherical kernel:

$$r_{ij}^B = 1 - \frac{3}{2} \frac{\|x_i - y_i\|_B}{\delta} + \frac{1}{2} \left(\frac{\|x_i - y_i\|_B}{\delta}\right)^3 \quad (1.26)$$

với  $\|x_i - y_i\|_B < \delta$

(v) Circular kernel:

$$r_{ij}^B = \frac{2}{\pi} \arccos\left(\frac{\|x_i - y_i\|_B}{\delta}\right) - \frac{2}{\pi} \frac{\|x_i - y_i\|_B}{\delta} \sqrt{1 - \left(\frac{\|x_i - y_i\|_B}{\delta}\right)^2} \quad (1.27)$$

với  $\|x_i - y_i\|_B < \delta$

Trong đó  $\delta$  là tham số của hàm kernel.

(7) Một số phương pháp khác [66]:

$$r_{ij}^B = \frac{1}{|B|} |\{c_k \in B \mid f_{c_k}(x_i) = f_{c_k}(x_j)\}| \quad (1.28)$$

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

$$r_{ij}^{c_k} = \begin{cases} 1, & \text{if } f_{c_k}(x_i) = f_{c_k}(x_j) \\ 0, & \text{otherwise.} \end{cases} \quad (1.29)$$

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

$$r_{ij}^{c_k} = F(x_i, x_j) \quad (1.30)$$

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

Sau đây là một số ví dụ của hàm  $F$

$$(1) r_{ij}^{c_k} = 1 - |f_{c_k}(x_i) - f_{c_k}(x_j)|.$$

$$(2) r_{ij}^{c_k} = \max\left(\min\left(\frac{f_{c_k}(x_j) - f_{c_k}(x_i) + \sigma_{c_k}}{\sigma_{c_k}}, \frac{f_{c_k}(x_i) - f_{c_k}(x_j) + \sigma_{c_k}}{\sigma_{c_k}}\right), 0\right)$$

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

## 1.4. Phương pháp đánh giá reduct

### 1.4.1. Các tiêu chí đánh giá

Các thuật toán giảm thuộc tính theo tiếp cận độ đo hiện nay thường được đánh giá dựa trên ba tiêu chí gồm có: *số lượng phần tử* của reduct thu được, *khả năng phân lớp*

của reduct trên mô hình được huấn luyện và *chi phí tính toán* của thuật toán.

Tập rút gọn thu được từ thuật toán có số lượng phần tử 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 reduct thu được hiệu quả. Chi phí tính toán càng nhanh cho biết khả năng rút gọn dữ liệu của thuật toán trên các tập dữ liệu lớn.

Mục tiêu chung của các thuật toán giảm thuộc tính là cố gắng đạt được cả ba tiêu chí trên, tuy nhiên trong thực tế với các bộ dữ liệu nhiều và phức tạp. Tiêu chí số lượng phần tử và khả năng phân lớp chính xác của reduct 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 reduct.

#### ***1.4.2. Mô hình và dữ liệu đánh giá***

Theo khảo sát của các tác giả trong công trình [56] cho thấy các thuật toán phân lớp được sử dụng phổ biến trong đánh giá khả năng 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 giảm 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***

Để đánh giá hiệu quả về khả năng phân lớp chính xác của reduct thu được từ các thuật toán. Nhiều chỉ số đánh giá đã được đề xuất, trong đó các chỉ số này đều sử dụng ma trận nhầm lẫn kết hợp với phương pháp đánh giá chéo [56].

Tiếp cận đánh giá chéo là phương pháp đánh giá quan trọng trong các mô hình của học máy, trong đó tập dữ liệu ban đầu được chia thành hai phần chính là tập dữ liệu huấn luyện và tập dữ liệu kiểm thử. Có hai loại đánh giá chéo phổ biến gồm có: phương pháp k-fold, phương pháp left-one-out. Trong hai loại phương pháp đánh giá chéo này, phương pháp 10-fold được sử dụng rộng rãi nhất trong các nghiên cứu về mô hình học máy. Phương pháp này thực hiện việc chia ngẫu nhiên tập dữ liệu ban đầu thành 10 phần trong đó 9 phần được huấn luyện và 1 phần được dùng để đánh giá. Thực hiện việc thay đổi tỉ lệ này với 10 lần khác nhau ta được phương pháp đánh giá chéo 10-fold.

**Bảng 1.4:** Ma trận nhầm lẫn nhị phân

Actual class	Predicted class	
	Positive	Negative
Positive	TP	FN
Negative	FP	TN

Ma trận nhầm lẫn là một công cụ hiệu quả trong việc phân tích hiệu năng của các bộ phân lớp dữ liệu. Với mô hình phân lớp dữ liệu nhị phân, ma trận nhầm lẫn có thể được biểu diễn trong Bảng 1.4. Sau đây là một số chỉ số để đánh giá khả năng phân lớp dựa trên ma trận nhầm lẫn.

(1) Accuracy:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}. \quad (1.31)$$

(2) Error:

$$Error = \frac{FP + FN}{TP + TN + FP + FN} \quad (1.32)$$

(3) Precision:

$$Precision = \frac{TP}{TP + FP} \quad (1.33)$$

(4) Recall:

$$Recall = \frac{TP}{TP + FN} \quad (1.34)$$

(5)  $F$  measure ( $F$ ):

$$F = \frac{2 \times \text{precision} \times \text{recall}}{\text{precision} + \text{recall}} \quad (1.35)$$

(6)  $F_\beta$  measure ( $F_\beta$ ):

$$F_\beta = \frac{(1 + \beta^2) \times \text{precision} \times \text{recall}}{\beta^2 \times \text{precision} + \text{recall}} \quad (1.36)$$

## 1.5. Một số phương pháp giảm thuộc tính

### 1.5.1. Phương pháp giảm thuộc tính theo tiếp cận ma trận phân biệt

Vào năm 1992, Skowron và Rauszer lần đầu tiên giới thiệu phương pháp giảm thuộc tính theo tiếp cận ma trận phân biệt trên nền RS [67]. Khi đó ma trận phân biệt có số lượng phần tử  $n \times n$  với  $n = |U|$ , kí hiệu là  $M(DS) = (c_{ij})_{n \times n}$  được xác định bởi:

$$c_{ij} = \begin{cases} \{c \in C \mid c(x_i) \neq c(x_j)\}, \omega(x_i, x_j) \\ \emptyset, \text{ otherwise.} \end{cases} \quad (1.37)$$

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

- (1)  $x_i \in POS_C(D) \wedge x_j \notin POS_C(D)$ ;
- (2)  $x_i \notin POS_C(D) \wedge x_j \in POS_C(D)$ ;
- (3)  $x_i, x_j \in POP_C(D) \wedge (x_i, x_j) \notin ind(D)$ .

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

$$f(C, D) = \bigwedge \{ \bigvee c_{ij} \mid c_{ij} \neq \emptyset \} \quad (1.38)$$

Khi đó tập thuộc tính lõi được xác định bởi:

$$\text{core}_C(D) = \{c \mid c_{ij} = \{c\}\} \quad (1.39)$$

Năm 2008, Tsang và các cộng sự [22] giới thiệu ma trận phân biệt mờ cho mô hình giảm thuộc tính dựa trên công thức xây dựng ma trận quan hệ sau:

$$c_{ij} = \begin{cases} \{c \in C \mid 1 - R_c(x_i, x_j) \geq \lambda_i\}, \lambda_j < \lambda_i; \\ \emptyset, \text{ otherwise,} \end{cases} \quad (1.40)$$

Trong đó:  $\lambda_i = R_C[x_i]_D(x_i)$ , và  $\lambda_j = R_C[x_i]_D(x_j)$ .

Năm 2009 Jensen và các cộng sự [68] định nghĩa lại ma trận của Tsang dựa theo công thức:

$$c_{ij} = \{c_\mu \mid \mu = N(R_c(x_i, x_j))\} \quad (1.41)$$

Tuy nhiên, các ma trận phân biệt này còn sinh ra quá nhiều reduct ứng viên do đó các tác giả [69] đề xuất ma trận phân biệt các thuộc tính điều kiện ràng buộc bởi thuộc tính quyết định như sau:

$$c_{ij} = \begin{cases} \{c \in C \mid T(R_c(x_i, x_j), \lambda_i) = 0\}, \text{ if } x_j \notin [x_i]_D; \\ \emptyset, \text{ otherwise} \end{cases} \quad (1.42)$$

trong đó  $\lambda_i = R_{\theta C}[x_i]_D(x_i)$ .

Chen và các cộng sự [70] kết hợp ma trận phân biệt với RS mờ để giảm thuộc tính trong bảng quyết định hybrid. Ma trận phân biệt được xác định bởi:

$$c_{ij} = \begin{cases} \{c \in C \mid \varphi_C(x_i) - R_D(c)(x_i, x_j) \leq \varepsilon\}, \text{ if } x_j \notin [x_i]_D \text{ and } \varphi_C(x_i) \neq 0; \\ \emptyset, \text{ otherwise,} \end{cases} \quad (1.43)$$

Trong đó  $\varphi_C(x_i) = \min_{D(x) \neq D(y)} R_D(c)(x_i, x_j)$ . Nếu  $(x, y) \in R_D(c)$ , thì  $R_D(c)(x_i, x_j) = 1$ ; ngược lại  $R_D(c)(x_i, x_j) = 0$ . Cho đến nay, có khá nhiều phương pháp giảm thuộc tính theo tiếp cận ma trận phân biệt được đề xuất trong các công trình [71]–[74].

### 1.5.2. Phương pháp giảm thuộc tính theo tiếp cận độ đo

Hầu hết các phương pháp giảm thuộc tính theo tiếp cận độ đo hiện nay đều sử dụng mô hình filter thuộc tính để xác định reduct. Ba thành phần quan trọng nhất để xây dựng mô hình filter thuộc tính gồm có:

1) Phương pháp tìm kiếm: Hầu hết các thuật toán giảm thuộc tính hiện nay đều sử dụng tiếp cận tìm kiếm tham lam. Trong đó kỹ thuật tìm kiếm tham lam theo chiều tiến thường xuất phát từ reduct ban đầu, sau đó bổ sung lần lượt vào reduct từng thuộc tính quan trọng nhất đối với reduct. Đối với tiếp cận tìm kiếm tham lam lùi, xuất phát từ tập thuộc tính ban đầu, lần lượt loại bỏ đi các thuộc tính dư thừa, không liên quan đến việc ra quyết định.

2) Phương pháp đánh giá: Hầu hết các thuật toán giảm thuộc tính theo tiếp cận độ đo đều sử dụng phương pháp bảo toàn độ đo để xác định reduct, dựa trên các chiến lược chọn lọc thuộc tính quan trọng nhất và loại bỏ đi thuộc tính dư thừa cho đến khi thông tin ra quyết định của reduct là tương đương với tập thuộc tính gốc thì thuật toán kết thúc.

Sau đây là một số độ đo được sử dụng để đánh giá độ quan trọng của thuộc tính và định nghĩa reduct trong các mô hình giảm thuộc tính theo tiếp cận độ đo hiện nay.

#### 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 RS.

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:

$$\text{POS}_B(D) = \bigcup_{X_i \in U/D} \underline{R}_B X_i \quad (1.44)$$

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

$$\gamma_B(D) = \frac{|P_B(D)|}{|U|} = \frac{\sum_{x \in U} P_B S_B(D)(x)}{|U|} \quad (1.45)$$

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:

$$\text{Sig}_1(a, B, D) = \gamma_B(D) - \gamma_{B-a}(D) \quad (1.46)$$

$$\text{Sig}_2(a, B, D) = \gamma_{B \cup a}(D) - \gamma_B(D) \quad (1.47)$$

Trong đó công thức 1.46 phù hợp với kỹ thuật tìm kiếm tham lam lùi còn công thức 1.47 phù hợp với kỹ thuật tìm kiếm tham lam tiến.

Trên cơ sở đó, các phương pháp giảm 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 1.5.

#### 1.5.2.2. độ đo độ chắc chắn

Độ đo độ chắc chắn là một độ đo quan trọng được dùng để đánh giá sự chắc chắn của thông tin trong bảng quyết định. dựa trên khái niệm Entropy thông tin của Shanon, một số độ đo độ chắc chắn được mở rộng cho bài toán giảm thuộc tính gồm có:

(1) Entropy điều kiện [57]: 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:

$$FE(B) = -\frac{1}{|U|} \sum_{i=1}^{|U|} \log_2 \frac{|[x_i]_{R_B}|}{|U|} \quad (1.48)$$

- entropy kết hợp:

$$FE(B, E) = -\frac{1}{|U|} \sum_{i=1}^{|U|} \log_2 \frac{|[x_i]_{R_B} \cap [x_i]_{R_E}|}{|U|} \quad (1.49)$$



**Bảng 1.5:** Tổng hợp phương pháp giảm thuộc tính theo độ phụ thuộc

STT	Tài liệu tham chiếu	Kiểu dữ liệu	Tiếp cận	Tập nền	Tiêu chuẩn đánh giá
1	[75]–[86]	Hybrid	NRS	Classical	accuracy, size, computation time
2	[27], [32], [87]–[92]	Number	NRS	FS	accuracy, size, computation time
3	[93]	Number	NRS	IFS	accuracy, size, computation time
4	[94]	Hybrid	PRS	Classical	accuracy, size, computation time
5	[17], [22]–[29], [59], [87], [89], [95]–[101]	Number	FRS	FS	accuracy, size, computation time
6	[34]–[36], [93], [102]–[108]	Number	IFRS	FS	accuracy, size, computation time

- entropy có điều kiện:

$$FE(E | B) = -\frac{1}{|U|} \sum_{i=1}^n \log_2 \frac{|[x_i]_{R_E} \cap [x_i]_{R_B}|}{|[x_i]_{R_B}|} \quad (1.50)$$

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:

$$\text{Sig}(a, B) = FE(B) - FE(B - \{a\}) \quad (1.51)$$

$$\text{Sig}(a, B, D) = FE(D | B - \{a\}) - FE(D | B) \quad (1.52)$$

(2) Mutual information [93]: Vào năm 2008, An và các cộng sự sử dụng khái niệm cực đại hóa thông tin liên quan và cực tiểu hóa thông tin dư thừa (mRMR) kết hợp với

khái niệm entropy thông tin để đề xuất thuật toán mRMR:

$$FMI(E;B) = FE(E) - FE(E | B) = -\frac{1}{|U|} \sum_{i=1}^{|U|} \log_2 \frac{|[x_i]_{R_B}| \cdot |[x_i]_{R_E}|}{|U| \cdot |[x_i]_{R_B} \cap [x_i]_{R_E}|}. \quad (1.53)$$

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:

$$Sig(a, B, D) = FMI(B \cup \{a\}; D) - FMI(B; D) \quad (1.54)$$

Trên cơ sở đó, các phương pháp giảm 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 1.6.

**Bảng 1.6:** Tổng hợp phương pháp giảm 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], [109], [110]	Number	Entropy thông tin	IFS	accuracy, size, computation time
2	[31], [53], [88]	Number	Entropy điều kiện	FS	accuracy, size, computation time
3	[111]	Hybrid	Entropy kết hợp	Classical	accuracy, size, computation time
4	[112]	Number	Entropy bù	FS	accuracy, size, computation time

### 1.5.2.3. độ đo khoảng cách

Độ đo khoảng cách là độ đo quan trọng được sử dụng để đo lường độ khác biệt giữa hai phần tử hai tập hợp. Dựa trên tính chất đơn điệu của độ đo, một số độ đo được sử dụng để mở rộng cho bài toán giảm thuộc tính gồm có:

(1) Khoảng cách Jacard [19]: Cho bảng quyết định  $DT = (U, C, D, f)$ . Với mọi  $X, Y \subseteq U$ , khoảng cách Jacard được xác định như sau:

$$D(X, Y) = 1 - \frac{|X \cap Y|}{|X \cup Y|} \quad (1.55)$$

(2) Khoảng cách tri thức [48]: 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  $K(P) = \{[u]_P : u \in U\}$  và  $K(Q) = \{[u]_Q : u \in U\}$ . Khi đó, khoảng cách tri thức giữa  $P$  và  $Q$  theo tiếp cận Jacard được xác định như sau:

$$d_J(K(P), K(Q)) = 1 - \frac{1}{|U|^2} \sum_{u=1}^{|U|} \frac{|[u]_P \cap [u]_Q|}{|[u]_P \cup [u]_Q|} \quad (1.56)$$

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:

$$SIG_B(a) = d_J(K(B), K(B \cup D)) - d_J(K(B \cup \{a\}), K(B \cup \{a\} \cup D)) \quad (1.57)$$

Trên cơ sở đó, các phương pháp giảm 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 1.7.

**Bảng 1.7:** Tổng hợp phương pháp giảm thuộc tính theo khoảng cách

STT	Tài liệu tham chiếu	Kiểu dữ liệu	Tiếp cận	Tập nền	Tiêu chuẩn đánh giá
1	[24], [33], [78], [113], [114]	Hybrid	KD	Classical, FS, IFS	accuracy, size, computation time
2	[29], [115], [116]	Number	GD	FS	accuracy, size, computation time
3	[29]	Number	PD	FRS	accuracy, size, computation time

### 1.5.3. Phương pháp giảm thuộc tính theo tiếp cận tôpô

Khái niệm không gian tôpô trên nền RS lần đầu tiên được đưa ra bởi Pawlack và cộng sự [14], [18], [54]. Trong đó, các phân hoạch được xem như là một cơ sở (subbase) của tôpô. Khi đó, Pawlack và các cộng sự cũng đã khẳng định cấu trúc đại số của RS với tôpô là tương đương trên nền tập rõ với các phép toán xấp xỉ dưới của RS tương đương với toán tử miền trong (INT) của tôpô và phép toán xấp xỉ trên của RS tương đương với toán tử miền ngoài (CLS) của tôpô [18], [54].

**Bảng 1.8:** Tổng hợp phương pháp xây dựng tôpô theo tiếp cận RS

STT	Tài liệu tham chiếu	Cơ sở tính toán
1	[18], [20], [37], [39], [41], [117], [118]	Không gian xấp xỉ
2	[37]–[39], [41], [47], [48], [117]–[120]	Tập xấp xỉ trên và tập xấp xỉ dưới
3	[20], [39], [45], [47], [58], [100], [119], [121], [122]	Không gian mẫu và quan hệ của các phép toán

Trên cơ các đề xuất ban đầu của Pawlack và các cộng sự, Lashin và các cộng sự đưa ra khái niệm biểu diễn tri thức trên không gian tôpô và phương pháp loại bỏ các tri thức dư thừa thông qua bảo toàn cấu trúc tôpô ban đầu [37]. Sau đó Zhu đề xuất khái niệm không gian tôpô trên họ các phủ của RS để giảm thuộc tính trong bảng quyết định không đầy đủ [123], tuy nhiên các nghiên cứu vẫn dừng lại ở giai đoạn đề xuất mà chưa có ứng dụng trên các bộ dữ liệu thực và đánh giá tính hiệu quả của mô hình đề xuất [124]. Gần đây có công trình nghiên cứu của Xie và các cộng sự có đề xuất hướng giảm thuộc tính dựa trên ma trận phân biệt sử dụng cấu trúc tôpô, trong đó tính chất phân biệt được mở rộng dựa trên cấu trúc tôpô thay vì tính chất của thuộc tính điều kiện hay thuộc tính quyết định [52], [94]. Shami và các cộng sự đề xuất xây dựng mô hình RS mới dựa trên không gian tôpô [44], [51]. Nhìn chung, sự tương quan của không gian tôpô và mô hình RS đều dựa trên cấu trúc của các phân hoạch sinh bởi các quan hệ và các phép toán xấp xỉ của RS, xem Bảng 1.8. Do đó, về cơ bản hiện nay có ba loại phương pháp xây dựng không gian tôpô như sau:

*Phương pháp sinh tôpô từ không gian xấp xỉ:* Đối với mô hình RS truyền thống, không gian xấp xỉ là các phân hoạch hay tập các lớp tương đương [18], [117]. Do đó, các phân hoạch này luôn là các base của một tôpô. Khi đó để nghiên cứu cấu trúc tôpô, các nhà nghiên cứu thường phân tích cấu trúc của một base là đủ thông tin cần phân tích [37]. Do đó, khi một quan hệ xây dựng không gian xấp xỉ có tính chất khác nhau thì base cũng sẽ khác nhau. Đối với các phủ sinh bởi các quan hệ dung sai thì các phủ này được coi là các subbase, khi đó dựa trên các tính chất của base sinh từ subbase, có thể đề xuất các phương pháp xây dựng base từ các phủ này [41]. Trên nền FS và IFS, các nhà nghiên cứu cũng đề xuất phương pháp xây dựng tôpô từ các không gian xấp xỉ mờ [20] và xấp xỉ mờ trực cảm tương ứng [39], [118].

*Phương pháp sinh tôpô từ cấu trúc RS:* Mô hình RS có hai phép toán cơ bản là xấp xỉ trên và xấp xỉ dưới để đánh giá mức độ thô của một tập trong một không gian xấp xỉ. Pawlack và cộng sự [14], [18] cũng đã chỉ ra sự tương quan giữa một tập với các tập xấp xỉ trên và dưới có quan hệ thứ tự, trong đó tập xấp xỉ trên luôn lớn hơn tập mục tiêu và tập mục tiêu luôn lớn tập xấp xỉ dưới. Do đó, kết hợp các tập xấp xỉ dưới và tập xấp xỉ trên luôn thỏa mãn là một cấu trúc tôpô [37]–[39], [41], [47], [48], [117]–[120].

*Phương pháp sinh tôpô dựa trên không gian mẫu:* Dựa trên các phép toán của FRS và không gian mẫu của FS, các nhà nghiên cứu đề xuất các phương pháp khác nhau để xây dựng tôpô mờ [39], [45], [58], [100], trong đó các tác giả đưa ra nhiều cấu trúc tôpô mờ khác nhau dựa trên tính chất của quan hệ xây dựng không gian xấp xỉ mờ và tính chất của phép toán xấp xỉ mờ trên, xấp xỉ mờ dưới tương ứng. Dựa trên các phép toán của IFRS và không gian mẫu IFS, các nhà nghiên cứu cũng mở rộng để xây dựng tôpô mờ trực cảm [20], [47], [119], [121], [122]. Hơn nữa, trong các công trình này, các tác giả còn chỉ rõ trường hợp nào hai tôpô bằng nhau và trường hợp nào hai tôpô bao thuộc lẫn nhau [20], [122]

Dựa trên khái niệm cơ sở  $\beta$  của không gian tôpô  $(U, \tau)$ . Lashin và các cộng sự [37] đã sử dụng khái niệm quan hệ dư thừa để định nghĩa reduct theo tiếp cận tôpô như

sau:

**Định nghĩa 1.17** (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 reduct của  $C$  khi và chỉ khi:

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

Dựa trên các kết quả khảo sát về các phương pháp giảm thuộc tính theo tiếp cận tôpô và RS cho thấy các tiếp cận giảm thuộc tính trong bảng quyết định hiện nay còn gặp nhiều thách thức về *chi phí tính toán* của thuật toán trên các tập dữ liệu số lượng phần tử lớn, khả năng cải thiện *số lượng phần tử* trên các bộ dữ liệu có số chiều lớn và khả năng cải thiện *nhiều* trên các bộ dữ liệu xấu. Sau đây là các phân tích về nhược điểm của từng tiếp cận, từ đó đưa ra câu hỏi nghiên cứu và phương hướng giải quyết được thực hiện trong các Chương nghiên cứu tiếp theo luận án.

1) Theo các tiếp cận độ đo được đề xuất trên nền tập mờ trực cảm hiện nay còn khoảng trống nghiên cứu cải thiện độ chính xác cho các tập dữ liệu nhiễu thông qua bài toán giảm thuộc tính. Các độ đo dựa trên mô hình VPRS của tập mờ và tập mờ trực cảm đã cho thấy tính hiệu quả của việc điều chỉnh các ngưỡng xấp xỉ để phù hợp với từng bộ dữ liệu. Các kết quả thực nghiệm trên các thuật toán này cho thấy khả năng phân lớp chính xác của reduct thu được được cải thiện đáng kể, đặc biệt là trên các bộ dữ liệu nhiễu. Tuy nhiên, chi phí tính toán trên tiếp cận này vẫn còn gặp nhiều thách thức. Do đó cần phải phát triển tiếp cận độ đo mới và phương pháp xác định thuộc tính nhiễu hiệu quả nhằm giảm chi phí tính toán của thuật toán.

2) Tiếp cận độ đo là tiếp cận truyền thống để đánh giá mức độ tương quan và mức độ phụ thuộc giữa các thuộc tính. Bản chất của các độ đo này dựa trên sự thay đổi về tỉ lệ tổng thành phần của các giá trị địa phương trên tổng thành phần của các giá trị toàn cục. Do đó để so sánh sự khác biệt giữa các tập mục tiêu cần xét, ta cần so sánh sự chênh lệch giữa các tỉ lệ này. Tuy nhiên, cách tính tỉ lệ trong hầu hết các độ đo hiện nay đều dựa trên khái niệm lực lượng của tập hợp. Tiếp cận tính toán này là không đủ

chặt để kết luận sự tương đương giữa hai tập giá trị cần xét, đặc biệt là trên các tập giá trị mờ dẫn tới số lượng phần tử reduct của các thuật toán hiện nay vẫn còn chưa hiệu quả. Do đó cần phải phát triển độ đo mới để đánh giá độ tương tự của hai tập chặt chẽ hơn.

3) Hầu hết các nghiên cứu về cấu trúc tôpô rút gọn theo tiếp cận RS hiện nay còn chưa đầy đủ, chưa rõ ràng về nền tảng lý thuyết. Do đó, các phương pháp giảm thuộc tính cho các tập dữ liệu thực hiện nay còn rất hạn chế về số lượng cũng như chất lượng mô hình lý thuyết. Do đó, cần phải phát triển khung nền tảng lý thuyết tôpô cho bài toán giảm thuộc tính. Trong đó cần phải chỉ rõ, phương pháp sinh tôpô theo tiếp cận nào là hiệu quả cho bài toán giảm thuộc tính và mở rộng các phép toán đại số trên tôpô để phát triển các phương pháp giảm thuộc tính hiệu quả trên các bộ dữ liệu có số chiều lớn.

## 1.6. Kết luận Chương 1

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

- Trình bày khái quát các khái niệm cơ bản về mô hình RS cổ điển truyền thống và tập mờ trực cảm. Ý nghĩa và vai trò của tập mờ trực cảm trong việc cải thiện nhiều, các độ đo cơ bản là cơ sở kiến thức quan trọng được sử dụng trong Chương 2 của luận án.

- Trình bày khái quát các khái niệm cơ bản về không gian tôpô, các nghiên cứu liên quan đến phương pháp xây dựng tôpô trên không gian xấp xỉ mờ là cơ sở kiến thức quan trọng được sử dụng trong Chương 3 của luận án.

- Trình bày khái quát các nghiên cứu liên quan đến phương pháp xây dựng tôpô theo tiếp cận RS, các phương pháp xây dựng các phép toán cho cấu trúc tôpô đại số là các kiến thức quan trọng được sử dụng trong Chương 4 của luận án.

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

### 2.1. Mở đầu

Để giảm thuộc tính trực tiếp trên các bảng quyết định miền giá trị số liên tục. Khoảng hơn mười năm trở lại đây, các nhà nghiên cứu thường mở rộng mô hình RS truyền thống sang mô hình RS lân cận [75], [80], [81], [85], [125], mô hình RS mờ [33], [56], [59], [99], [126]. Bên cạnh đó, mô hình RS xác suất [29], [127], [128], mô hình tập thô biến thiên về độ chính xác [89], [129] cũng được các nhà nghiên cứu mở rộng cho các trường hợp dữ liệu nhiễu. Tuy nhiên khả năng cải thiện nhiều của các tiếp cận trên vẫn còn chưa hiệu quả. Gần đây, mô hình RS mờ trực cảm [27], [34], [93] được nhiều nhà nghiên cứu quan tâm trong cải thiện chất lượng phân lớp của các reduct trên các bộ dữ liệu nhiễu.

Trên cơ sở khung kiến trúc [34] của IFRS, nhiều mô hình suy rộng từ RS mờ trực cảm được đề xuất như mô hình RS lân cận mờ trực cảm [27], [34], [93], mô hình RS biến thiên mờ trực cảm [108], mô hình RS xác suất mờ trực cảm [128]. Từ đó, nhiều phương pháp giảm thuộc tính cho bảng quyết định miền giá trị số được đề xuất [36], [102], [103], [128], [130].

Gần đây các phương pháp giảm thuộc tính theo tiếp cận IFRS do Tan và các công sự đề xuất [36], [131] đã chứng minh được sự cải thiện nhiều cho reduct hiệu quả trên một số bộ dữ liệu so với các mô hình mở rộng khác của IFRS. Tuy nhiên không gian xấp xỉ mờ trực cảm các tác giả đề xuất còn quá phức tạp và chưa mô tả đầy đủ thông tin quan hệ của các đối tượng theo góc nhìn của tập mờ trực cảm. Hơn nữa độ đo đánh giá độ quan trọng của các tác giả đề xuất còn quá phức tạp, chiếm nhiều chi phí tính



toán trong thực tế.

Trong phần này luận án trình bày về phương pháp giảm thuộc tính theo tiếp cận RS mờ trực cảm. Trong đó, đề xuất độ đo đánh giá độ quan trọng của thuộc tính trên cơ sở mở rộng độ đo khoảng cách trên không gian xấp xỉ mờ trực cảm. Trên cơ sở đó đề xuất thuật toán giảm thuộc tính theo phương pháp lai ghép filter - wrapper hai bước. Trong đó bước  $W_\delta$  được sử dụng để xác định reduct ứng viên mức  $\delta$ , bước  $W_A$  được sử dụng để xác định reduct thực sự của thuật toán. Các kết quả thực nghiệm cho thấy thuật toán đề xuất cho các reduct hiệu quả về số lượng phần tử và khả năng phân lớp trên hầu hết các tập dữ liệu so với tập dữ liệu gốc. Hơn nữa một số tập dữ liệu nhiều cũng được cải thiện tốt hơn so với các thuật toán được so sánh.

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  $a \geq b$  thì  $a - b \geq \min(a, c) - \min(b, c)$
- 2) Nếu  $a \leq b$  thì  $a - b \leq \max(a, c) - \max(b, c)$

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

- 1) Nếu  $\tilde{X} \subseteq \tilde{Y}$  thì  $|\tilde{Y}| - |\tilde{Y} \cap \tilde{Z}| \geq |\tilde{X}| - |\tilde{X} \cap \tilde{Z}|$
- 2) Nếu  $\tilde{X} \subseteq \tilde{Y}$  thì  $|\tilde{Z}| - |\tilde{Z} \cap \tilde{X}| \geq |\tilde{Z}| - |\tilde{Z} \cap \tilde{Y}|$
- 3)  $|\tilde{X}| - |\tilde{X} \cap \tilde{Y}| + |\tilde{Z}| - |\tilde{Z} \cap \tilde{X}| \geq |\tilde{Z}| - |\tilde{Z} \cap \tilde{Y}|$

*Chứng minh.* Sau đây ta lần lượt chứng minh từng tính chất như sau:

- 1) Vì  $\tilde{X} \subseteq \tilde{Y}$ , do đó với mọi  $u \in U$  ta có  $\mu_{\tilde{Y}}(u) \geq \mu_{\tilde{X}}(u)$  và  $\nu_{\tilde{Y}}(u) \leq \nu_{\tilde{X}}(u)$ . Áp dụng bổ đề 2.1 ta có:

$$\begin{aligned}
(1): & \mu_{\tilde{Y}}(u) - \mu_{\tilde{X}}(u) \geq \min(\mu_{\tilde{Y}}(u), \mu_{\tilde{Z}}(u)) - \min(\mu_{\tilde{X}}(u), \mu_{\tilde{Z}}(u)) \\
\Leftrightarrow & \sum_{i=1}^{|U|} \mu_{\tilde{Y}}(i) - \sum_{i=1}^{|U|} \mu_{\tilde{X}}(i) \geq \sum_{i=1}^{|U|} \min(\mu_{\tilde{Y}}(i), \mu_{\tilde{Z}}(i)) - \sum_{i=1}^{|U|} \min(\mu_{\tilde{X}}(i), \mu_{\tilde{Z}}(i)) \\
(2): & v_{\tilde{Y}}(u) - v_{\tilde{X}}(u) \leq \max(v_{\tilde{Y}}(u), v_{\tilde{Z}}(u)) - \max(v_{\tilde{X}}(u), v_{\tilde{Z}}(u)) \\
\Leftrightarrow & \sum_{i=1}^{|U|} v_{\tilde{Y}}(i) - \sum_{i=1}^{|U|} v_{\tilde{X}}(i) \leq \sum_{i=1}^{|U|} \max(v_{\tilde{Y}}(i), v_{\tilde{Z}}(i)) - \sum_{i=1}^{|U|} \max(v_{\tilde{X}}(i), v_{\tilde{Z}}(i))
\end{aligned}$$

Từ (1) và (2) ta có:  $|\tilde{Y}| - |\tilde{X}| \geq |\tilde{Y} \cap \tilde{Z}| - |\tilde{X} \cap \tilde{Z}| \Leftrightarrow |\tilde{Y}| - |\tilde{Y} \cap \tilde{Z}| \geq |\tilde{X}| - |\tilde{X} \cap \tilde{Z}|$

2) Vì  $\tilde{X} \subseteq \tilde{Y}$ , do đó với mọi  $u \in U$  ta có  $\mu_{\tilde{Y}}(u) \geq \mu_{\tilde{X}}(u)$  và  $v_{\tilde{Y}}(u) \leq v_{\tilde{X}}(u)$ . Áp dụng bổ đề 2.1 ta có:

$$\begin{aligned}
(3): & \mu_{\tilde{Y}}(u) \geq \mu_{\tilde{X}}(u) \Leftrightarrow \min(\mu_{\tilde{Y}}(u), \mu_{\tilde{Z}}(u)) \geq \min(\mu_{\tilde{X}}(u), \mu_{\tilde{Z}}(u)) \\
\Leftrightarrow & \mu_{\tilde{Z}}(u) - \min(\mu_{\tilde{X}}(u), \mu_{\tilde{Z}}(u)) \geq \mu_{\tilde{Z}}(u) - \min(\mu_{\tilde{Y}}(u), \mu_{\tilde{Z}}(u)) \\
\Leftrightarrow & \sum_{i=1}^{|U|} \mu_{\tilde{Z}}(i) - \sum_{i=1}^{|U|} \min(\mu_{\tilde{X}}(i), \mu_{\tilde{Z}}(i)) \geq \sum_{i=1}^{|U|} \mu_{\tilde{Z}}(i) - \sum_{i=1}^{|U|} \min(\mu_{\tilde{Y}}(i), \mu_{\tilde{Z}}(i)) \\
(4): & v_{\tilde{Y}}(u) \leq v_{\tilde{X}}(u) \Leftrightarrow \max(v_{\tilde{Y}}(u), v_{\tilde{Z}}(u)) \leq \max(v_{\tilde{X}}(u), v_{\tilde{Z}}(u)) \\
\Leftrightarrow & v_{\tilde{Z}}(u) - \max(v_{\tilde{X}}(u), v_{\tilde{Z}}(u)) \leq v_{\tilde{Z}}(u) - \max(v_{\tilde{Y}}(u), v_{\tilde{Z}}(u)) \\
\Leftrightarrow & \sum_{i=1}^{|U|} v_{\tilde{Z}}(i) - \sum_{i=1}^{|U|} \max(v_{\tilde{X}}(i), v_{\tilde{Z}}(i)) \leq \sum_{i=1}^{|U|} v_{\tilde{Z}}(i) - \sum_{i=1}^{|U|} \max(v_{\tilde{Y}}(i), v_{\tilde{Z}}(i))
\end{aligned}$$

Từ (3) và (4) ta có:  $|\tilde{Z}| - |\tilde{Z} \cap \tilde{X}| \geq |\tilde{Z}| - |\tilde{Z} \cap \tilde{Y}|$

3) Vì  $\tilde{X} \cap \tilde{Z} \subseteq \tilde{X}$  và  $\tilde{X} \cap \tilde{Y} \subseteq \tilde{Y}$ . Từ 1) và 2) ta có:

$$\begin{aligned}
(5): & |\tilde{X}| - |\tilde{X} \cap \tilde{Y}| \geq |\tilde{X} \cap \tilde{Z}| - |\tilde{X} \cap \tilde{Z} \cap \tilde{Y}| \\
(6): & |\tilde{Z}| - |\tilde{Z} \cap \tilde{X} \cap \tilde{Y}| \geq |\tilde{Z}| - |\tilde{Z} \cap \tilde{Y}|
\end{aligned}$$

Từ (5) và (6) ta có:  $|\tilde{X}| - |\tilde{X} \cap \tilde{Y}| + |\tilde{Z}| - |\tilde{Z} \cap \tilde{X}| \geq |\tilde{Z}| - |\tilde{Z} \cap \tilde{Y}|$ . đpcm.  $\square$

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

*Chứng minh.* Vì  $|\tilde{X} \cup \tilde{Y}| \geq |\tilde{X} \cap \tilde{Y}|$  do đó  $\tilde{d}(\tilde{X}, \tilde{Y}) \geq 0$ . Để chứng minh  $\tilde{d}(\tilde{X}, \tilde{Y})$  là độ đo khoảng cách, ta cần chứng minh  $\tilde{d}(\tilde{X}, \tilde{Y})$  thỏa mãn bất đẳng thức tam giác.

Thật vậy ta có:

$$(7): \left| \tilde{X} \right| - \left| \tilde{X} \cap \tilde{Y} \right| + \left| \tilde{Z} \right| - \left| \tilde{Z} \cap \tilde{X} \right| \geq \left| \tilde{Z} \right| - \left| \tilde{Z} \cap \tilde{Y} \right|$$

$$(8): \left| \tilde{X} \right| - \left| \tilde{X} \cap \tilde{Z} \right| + \left| \tilde{Y} \right| - \left| \tilde{Y} \cap \tilde{X} \right| \geq \left| \tilde{Y} \right| - \left| \tilde{Y} \cap \tilde{Z} \right|$$

Từ (7) và (8) ta có:

$$(9): \left( \left| \tilde{X} \right| + \left| \tilde{Y} \right| - 2 \left| \tilde{X} \cap \tilde{Y} \right| \right) + \left( \left| \tilde{X} \right| + \left| \tilde{Z} \right| - 2 \left| \tilde{X} \cap \tilde{Z} \right| \right) \geq \left| \tilde{Y} \right| + \left| \tilde{Z} \right| - 2 \left| \tilde{Y} \cap \tilde{Z} \right|.$$

Hơn nữa với mọi  $x, y \in \mathbb{R}$  ta luôn có  $\max(x, y) = x + y - \min(x, y)$  và  $\min(x, y) = x + y - \max(x, y)$  do đó với mọi  $u \in U$ :

$$(10): \max(\mu_X(u), \mu_Y(u)) = \mu_X(u) + \mu_Y(u) - \min(\mu_X(u), \mu_Y(u))$$

$$(11): \min(\nu_X(u), \nu_Y(u)) = \nu_X(u) + \nu_Y(u) - \max(\nu_X(u), \nu_Y(u))$$

Từ (10) và (11) ta có:

$$(12): \left| \tilde{X} \cup \tilde{Y} \right| = \left| \tilde{X} \right| + \left| \tilde{Y} \right| - \left| \tilde{X} \cap \tilde{Y} \right|.$$

Từ (9) và (12) ta có:  $\left( \left| \tilde{X} \cup \tilde{Y} \right| - \left| \tilde{X} \cap \tilde{Y} \right| \right) + \left( \left| \tilde{X} \cup \tilde{Z} \right| - \left| \tilde{X} \cap \tilde{Z} \right| \right) \geq \left| \tilde{Y} \cup \tilde{Z} \right| - \left| \tilde{Y} \cap \tilde{Z} \right|$  hay  $\tilde{d}(\tilde{X}, \tilde{Y}) + \tilde{d}(\tilde{X}, \tilde{Z}) \geq \tilde{d}(\tilde{Y}, \tilde{Z})$ . Do đó  $\tilde{d}(\tilde{X}, \tilde{Y})$  là một độ đo khoảng cách.  $\square$

### 2.2.2. Khoảng cách giữa hai phân hoạch mờ trực cảm

**Định nghĩa 2.1** (Khoảng cách giữa hai phân hoạch mờ trực cảm). Cho bảng quyết định  $DT = (U, C, D, f)$  và hai phân hoạch  $[\tilde{X}], [\tilde{Y}]$  tương ứng của  $X, Y \subseteq C$ . Khi đó khoảng cách giữa  $[\tilde{X}], [\tilde{Y}]$  được xác định bởi:

$$\tilde{d}([\tilde{X}], [\tilde{Y}]) = \frac{1}{|U|^2} \sum_{i=1}^{|U|} \left( \left| [i]_{[\tilde{X}]} \cup [i]_{[\tilde{Y}]} \right| - \left| [i]_{[\tilde{X}]} \cap [i]_{[\tilde{Y}]} \right| \right) \quad (2.1)$$

**Mệnh đề 2.3** (Độ đo khoảng cách mờ trực cảm). Cho bảng quyết định  $DT = (U, C, D, f)$  và hai phân hoạch  $[\tilde{X}], [\tilde{Y}]$  tương ứng của  $X, Y \subseteq C$ . Khi đó  $\tilde{d}([\tilde{X}], [\tilde{Y}])$  là khoảng cách giữa hai phân hoạch mờ trực cảm.

*Chứng minh.* Thật vậy, ta luôn có  $\tilde{d}([\tilde{X}], [\tilde{Y}]) \geq 0$  và  $\tilde{d}([\tilde{X}], [\tilde{Y}]) = \tilde{d}([\tilde{Y}], [\tilde{X}])$ .

Khi đó để  $\tilde{d}([\tilde{X}], [\tilde{Y}])$  là một khoảng cách, ta cần chứng minh bất đẳng thức tam giác

$$\tilde{d}([\tilde{X}], [\tilde{Y}]) + \tilde{d}([\tilde{X}], [\tilde{Z}]) \geq \tilde{d}([\tilde{Y}], [\tilde{Z}]).$$

Từ mệnh đề 2.2, với mọi  $u \in U$  ta có  $\tilde{d}([\tilde{X}], [\tilde{Y}]) + \tilde{d}([\tilde{X}], [\tilde{Z}]) \geq \tilde{d}([\tilde{Y}], [\tilde{Z}])$ . Khi

$$\text{đó: } \tilde{d}([\tilde{X}], [\tilde{Y}]) + \tilde{d}([\tilde{X}], [\tilde{Z}])$$

$$= \frac{1}{|U|^2} \sum_{i=1}^{|U|} \left( \left| [i]_{[\tilde{X}]} \cup [i]_{[\tilde{Y}]} \right| - \left| [i]_{[\tilde{X}]} \cap [i]_{[\tilde{Y}]} \right| \right)$$

$$+ \frac{1}{|U|^2} \sum_{i=1}^{|U|} \left( \left| [i]_{[\tilde{X}]} \cup [i]_{[\tilde{Z}]} \right| - \left| [i]_{[\tilde{X}]} \cap [i]_{[\tilde{Z}]} \right| \right)$$

$$= \frac{1}{|U|^2} \sum_{i=1}^{|U|} \tilde{d}([i]_{[\tilde{X}]}, [i]_{[\tilde{Y}]}) + \frac{1}{|U|^2} \sum_{i=1}^{|U|} \tilde{d}([i]_{[\tilde{X}]}, [i]_{[\tilde{Z}]}) \geq \frac{1}{|U|^2} \sum_{i=1}^{|U|} \tilde{d}([i]_{[\tilde{Y}]}, [i]_{[\tilde{Z}]})$$

$$= \tilde{d}([\tilde{Y}], [\tilde{Z}]). \text{ Ta có đpcm.} \quad \square$$

**Định nghĩa 2.2** (Khoảng cách phân hạt mờ trực cảm). Cho bảng quyết định  $DT = (U, C, D, f)$  và  $[\tilde{X}], [\tilde{Y}]$  tương ứng là các phân hoạch của  $X$  và  $Y = X \cup D$  với  $X \subseteq C$ . Khi đó khoảng cách giữa  $[\tilde{X}], [X \cup D]$  được xác định bởi:

$$\tilde{d}([\tilde{X}], [\tilde{Y}]) = \frac{1}{|U|^2} \sum_{i=1}^{|U|} \left( \left| [i]_{[\tilde{X}]} \right| - \left| [i]_{[\tilde{X}]} \cap [i]_{[\tilde{D}]} \right| \right) \quad (2.2)$$

**Mệnh đề 2.4** (Độ đo phân hạt). Cho bảng quyết định  $DT = (U, C, D, f)$  và  $[\tilde{X}], [\tilde{Y}]$  tương ứng là các phân hoạch của  $X$  và  $Y = X \cup D$  với  $X \subseteq C$ . Khi đó  $\tilde{d}([\tilde{X}], [\tilde{Y}])$  là một độ đo khoảng cách.

*Chứng minh.* Từ mệnh đề 2.3 ta có:  $\tilde{d}([\tilde{X}], [X \cup D])$

$$= \frac{1}{|U|^2} \sum_{i=1}^{|U|} \left( \left| [i]_{[\tilde{X}]} \cup [i]_{[X \cup D]} \right| - \left| [i]_{[\tilde{X}]} \cap [i]_{[X \cup D]} \right| \right)$$

$$= \frac{1}{|U|^2} \sum_{i=1}^{|U|} \left( \left| [i]_{[\tilde{X}]} \cup \left( [i]_{[\tilde{X}]} \cap [i]_{[\tilde{D}]} \right) \right| - \left| [i]_{[\tilde{X}]} \cap [i]_{[\tilde{D}]} \right| \right)$$

$$= \frac{1}{|U|^2} \sum_{i=1}^{|U|} \left( \left| [i]_{[\tilde{X}]} \right| - \left| [i]_{[\tilde{X}]} \cap [i]_{[\tilde{D}]} \right| \right) \quad \square$$

**Mệnh đề 2.5** (Tính chất phản đơn điệu của độ đo phân hạt). Cho bảng quyết định  $DT = (U, C, D, f)$  và  $[\tilde{X}], [\tilde{Y}]$  tương ứng là các phân hoạch của  $X$  và  $Y$  với  $X \subseteq Y \subseteq C$ .

Khi đó  $\tilde{d} \left( [\tilde{X}], [X \cup D] \right) \geq \tilde{d} \left( [\tilde{Y}], [Y \cup D] \right)$  với mọi  $u \in U$ .

*Chứng minh.* Vì  $X \subseteq Y$ , do đó  $[\tilde{Y}] \leq [\tilde{X}]$  nghĩa là  $[u]_{[\tilde{Y}]} \subseteq [u]_{[\tilde{X}]}$  với  $1 \leq u \leq |U|$ . Khi

đó  $\left| [u]_{[\tilde{Y}]} \right| \leq \left| [u]_{[\tilde{X}]} \right|$ . Với mọi  $u \in U$  ta có:

$$(13): \left| [u]_{[\tilde{Y}]} \right| - \left| [u]_{[\tilde{Y}]} \cap [u]_{[\tilde{D}]} \right| = \sum_{k=1}^{|U|} \mu_{[u]_{[\tilde{Y}]}}(k) - \sum_{k=1}^{|U|} \min \left\{ \mu_{[u]_{[\tilde{Y}]}}(k), \mu_{[u]_{[\tilde{D}]}}(k) \right\}$$

$$(14): \left| [u]_{[\tilde{X}]} \right| - \left| [u]_{[\tilde{X}]} \cap [u]_{[\tilde{D}]} \right| = \sum_{k=1}^{|U|} \mu_{[u]_{[\tilde{X}]}}(k) - \sum_{k=1}^{|U|} \min \left\{ \mu_{[u]_{[\tilde{X}]}}(k), \mu_{[u]_{[\tilde{D}]}}(k) \right\}$$

Với mọi  $k \in [u]_{[\tilde{D}]}$ , nếu  $\mu_{[u]_{[\tilde{D}]}}(k) = 1$  thì

$$(15): \left| [u]_{[\tilde{X}]} \right| - \left| [u]_{[\tilde{X}]} \cap [u]_{[\tilde{D}]} \right| = 0 = \left| [u]_{[\tilde{Y}]} \right| - \left| [u]_{[\tilde{Y}]} \cap [u]_{[\tilde{D}]} \right|$$

Với mọi  $k \notin [u]_{[\tilde{D}]}$  nếu  $\mu_{[u]_{[\tilde{D}]}}(k) = 0$  thì

$$(16): \left| [u]_{[\tilde{X}]} \right| - \left| [u]_{[\tilde{X}]} \cap [u]_{[\tilde{D}]} \right| \geq \left| [u]_{[\tilde{Y}]} \right| - \left| [u]_{[\tilde{Y}]} \cap [u]_{[\tilde{D}]} \right| \Leftrightarrow \left| [u]_{[\tilde{X}]} \right| \geq \left| [u]_{[\tilde{Y}]} \right|$$

Từ (15) và (16) ta có:

$$\left| [u]_{[\tilde{X}]} \right| - \left| [u]_{[\tilde{X}]} \cap [u]_{[\tilde{D}]} \right| \geq \left| [u]_{[\tilde{Y}]} \right| - \left| [u]_{[\tilde{Y}]} \cap [u]_{[\tilde{D}]} \right|$$

$$\Leftrightarrow \frac{1}{|U|^2} \sum_{u=1}^{|U|} \left( \left| [u]_{[\tilde{X}]} \right| - \left| [u]_{[\tilde{X}]} \cap [u]_{[\tilde{D}]} \right| \right) \geq \frac{1}{|U|^2} \sum_{u=1}^{|U|} \left( \left| [u]_{[\tilde{Y}]} \right| - \left| [u]_{[\tilde{Y}]} \cap [u]_{[\tilde{D}]} \right| \right)$$

$$\Leftrightarrow \tilde{d} \left( [\tilde{X}], [X \cup D] \right) \geq \tilde{d} \left( [\tilde{Y}], [Y \cup D] \right). \text{ đpcm} \quad \square$$

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

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

**Định nghĩa 2.3** (Ma trận  $\delta$  equal). Cho bảng quyết định  $DT = (U, C, D, f)$  và hai ma trận quan hệ mờ trực cảm  $\tilde{M}_B = [b_{ij}]_{n \times n}$ ,  $\tilde{M}_C = [c_{ij}]_{n \times n}$  xác định trên  $B$  và  $C$  với  $B \subseteq C$ ,  $n = |U|$ . Khi đó  $\tilde{M}_B$  và  $\tilde{M}_C$  được gọi là  $\delta$  - equal khi và chỉ khi:

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

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

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  $\tilde{M}_B \stackrel{\delta}{=} \tilde{M}_C$ .

**Định nghĩa 2.4** (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 định nghĩa bởi công thức sau:

$$SIG_B(a) = \tilde{d} \left( [\tilde{B}], [B \tilde{\cup} D] \right) - \tilde{d} \left( [B \tilde{\cup} \{a\}], [B \cup \{a\} \cup D] \right) \quad (2.3)$$

**Định nghĩa 2.5** (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 reduct nếu:

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

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

Dựa theo các thành phần cơ bản của mô hình giảm thuộc tính đề xuất đã được định nghĩa bên trên, sau đây luận án đề xuất thuật toán giảm thuộc tính theo phương pháp lai ghép filter-wrapper hai bước.

Thuật toán đề xuất bao gồm có hai giai đoạn, giai đoạn filter -  $W_\delta$  và giai đoạn  $W_A$ . Trong đó, bước filter sử dụng định nghĩa 2.4 và định nghĩa 2.5 để xác định reduct ứng viên mức  $\delta$ . Kết hợp với với mô hình phân lớp *Model*, bước  $W_\delta$  để xác định reduct ứng viên tốt nhất trong toàn bộ các giá trị  $\delta$ . Kết thúc giai đoạn filter -  $W_\delta$ , chuyển sang giai đoạn  $W_A$ . Giai đoạn  $W_A$  truy vết tập con nào của reduct ứng viên mức  $\delta$  có khả năng phân lớp cao nhất với mô hình phân lớp *Model*. Kết thúc giai đoạn  $W_A$  ta thu được reduct thực sự của thuật toán. Sau đây là các bước chi tiết của thuật toán đề xuất.

Trong đó:  $R_W^A$  là kí hiệu cho reduct thu được,  $R_F^\delta$  là tập thuộc tính lọc được tại mức  $\delta$ ,  $R_W^\delta$  là reduct ứng viên có khả năng phân lớp cao nhất tại mức  $\delta$ . ACC là hàm đánh giá khả năng phân lớp chính xác của reduct cho trước trên mô hình phân lớp *Model*.

---

**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  $Model$ ,  $\Delta = \{0.1, 0.2, \dots, 0.9\}$

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  $[c]$ ;
5: end for
6: for all  $\delta \in \Delta$  do
7:    $R_F^\delta \leftarrow \emptyset$ ;
8:   while  $[R_F^\delta \cup D] \neq [C \cup D]$  do
9:      $c_m \in C - R_F^\delta \mid SIG_{R_F^\delta}(c_m) = \underset{c \in C - R_F^\delta}{Max} \{SIG_{R_F^\delta}(c)\}$ ;           {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 ( $W_\delta$ )}
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  $R_W^A$ ;

```

---

Mỗi giá trị  $\delta$  có bước nhảy là 0.1. Sau đây là phần đánh giá độ phức tạp của thuật toán đề xuất.

Trước tiên, luận án kí hiệu  $|U|$  là số các đối tượng và  $|C|$  là số các thuộc tính trong bảng quyết định  $DT = (U, C, D, f)$ ,  $|R^\delta|$  là số lượng phần tử của tập thuộc tính rút gọn ứng viên,  $|\delta|$  là số lượng các giá trị  $\delta$  cần xét. Gọi  $\mathbb{T}$  là chi phí tính toán của mô hình phân lớp  $Model$ . Khi đó, độ phức tạp của thuật toán được trình bày trong Bảng 2.1 như sau:

Để minh họa quá trình hoạt động của thuật toán đề xuất, sau đây là phần trình

**Bảng 2.1:** Độ phức tạp của thuật toán IFD

Dòng lệnh	Độ phức tạp
Dòng 3 - 5	$O( C  U ^2)$
Dòng 8 - 11	$O( C ^2 U ^2)$
Dòng 13	$O(\mathbb{T})$
Dòng 6-14	$O(\mathbb{T} \Delta  C ^2 U ^2)$
Dòng 16	$O(\mathbb{T} R_W^\delta )$
Tổng	$O( C  U ^2) + O(\mathbb{T} \Delta  C ^2 U ^2) + O(\mathbb{T} R_W^\delta )$

bày ví dụ số minh họa trên bảng quyết định số như sau. Cho bảng quyết định  $DT = (U, C, D, f)$  được trình bày như trong bảng 1.3. Các bước của thuật toán được thực hiện tuần tự như sau:

*Giai đoạn khởi tạo:*

- Đặt:  $R_W^A \leftarrow \emptyset$ ;  $R_W^\delta \leftarrow \emptyset$ ;

- Tính các phân hoạch  $[\tilde{a}]$ ,  $[\tilde{b}]$ ,  $[\tilde{c}]$ ,  $[\tilde{d}]$ ,  $[\tilde{e}]$ ,  $[\tilde{f}]$ ,  $[\tilde{D}]$ ,  $[\tilde{C}]$ ,  $[\tilde{B}]$ . Trong đó:  $[\tilde{B}]$  là ma trận quan hệ mờ trực cảm thô nhất.

*Giai đoạn Filter -  $W_\delta$*

Lần lượt thử các giá trị  $\delta$  từ 0.1  $\rightarrow$  0.9 với mỗi bước nhảy là 0.1. Khi đó giá trị  $\delta$  nào cho reduct ứng viên có độ chính xác cao nhất, ta sẽ chọn reduct ứng viên đó cho giai đoạn  $W_A$  của thuật toán. Giả sử reduct ứng viên có độ chính xác cao nhất tại giá trị  $\delta = 0.8$ , sau đây là chi tiết các bước thực hiện tại giá trị  $\delta = 0.8$ .

$$R_F^\delta = \emptyset$$

Vì  $[R_F^\delta \cup D] \neq [C \cup D]$ , do đó:

- Tính:

$$SIG_{R_F^\delta}(a) = \tilde{d}\left(\tilde{R}_F^\delta, [R_F^\delta \cup D]\right) - \tilde{d}\left([R_F^\delta \cup \{a\}], [R_F^\delta \cup \{a\} \cup D]\right) = 0.5 - 0.31 = 0.19$$

$$SIG_{R_F^\delta}(b) = \tilde{d}\left(\tilde{R}_F^\delta, [R_F^\delta \cup D]\right) - \tilde{d}\left([R_F^\delta \cup \{b\}], [R_F^\delta \cup \{b\} \cup D]\right) = 0.5 - 0.43 = 0.07$$



$$SIG_{R_F^\delta}(c) = \tilde{d}\left([R_F^\delta], [R_F^\delta \cup D]\right) - \tilde{d}\left([R_F^\delta \cup \{c\}], [R_F^\delta \cup \{c\} \cup D]\right) = 0.5 - 0.36 = 0.14$$

$$SIG_{R_F^\delta}(d) = \tilde{d}\left([R_F^\delta], [R_F^\delta \cup D]\right) - \tilde{d}\left(P_{R_F^\delta \cup \{d\}}^\approx, [R_F^\delta \cup \{d\} \cup D]\right) = 0.5 - 0.38 = 0.12$$

$$SIG_{R_F^\delta}(e) = \tilde{d}\left([R_F^\delta], [R_F^\delta \cup D]\right) - \tilde{d}\left([R_F^\delta \cup \{e\}], [R_F^\delta \cup \{e\} \cup D]\right) = 0.5 - 0.26 = 0.24$$

$$SIG_{R_F^\delta}(f) = \tilde{d}\left([R_F^\delta], [R_F^\delta \cup D]\right) - \tilde{d}\left([R_F^\delta \cup \{f\}], [R_F^\delta \cup \{f\} \cup D]\right) = 0.5 - 0.26 = 0.24$$

- Chọn  $e$  vì  $SIG_{R_F^\delta}(e)$  lớn nhất, do đó  $R_F^\delta = \{e\}$ .

Vì  $\min\left(1 - \sup_{i,j=1}^6 |\mu(b_{ij}) - \mu(c_{ij})|, 1 - \sup_{i,j=1}^6 |\nu(b_{ij}) - \nu(c_{ij})|\right) = 0.4 < \delta$  hay  $[\{R_F^\delta - b\} \cup D] \neq (\delta)[C \cup D]$ , do đó tiếp tục vòng lặp ta có:

- Tính:

$$SIG_{R_F^\delta}(a) = \tilde{d}\left([R_F^\delta], [R_F^\delta \cup D]\right) - \tilde{d}\left([R_F^\delta \cup \{a\}], [R_F^\delta \cup \{a\} \cup D]\right) = 0.26 - 0.23 = 0.03$$

$$SIG_{R_F^\delta}(b) = \tilde{d}\left([R_F^\delta], [R_F^\delta \cup D]\right) - \tilde{d}\left([R_F^\delta \cup \{b\}], [R_F^\delta \cup \{b\} \cup D]\right) = 0.26 - 0.25 = 0.01$$

$$SIG_{R_F^\delta}(c) = \tilde{d}\left([R_F^\delta], [R_F^\delta \cup D]\right) - \tilde{d}\left([R_F^\delta \cup \{c\}], [R_F^\delta \cup \{c\} \cup D]\right) = 0.26 - 0.18 = 0.08$$

$$SIG_{R_F^\delta}(d) = \tilde{d}\left([R_F^\delta], [R_F^\delta \cup D]\right) - \tilde{d}\left(P_{R_F^\delta \cup \{d\}}^\approx, [R_F^\delta \cup \{d\} \cup D]\right) = 0.26 - 0.20 = 0.06$$

$$SIG_{R_F^\delta}(f) = \tilde{d}\left([R_F^\delta], [R_F^\delta \cup D]\right) - \tilde{d}\left([R_F^\delta \cup \{f\}], [R_F^\delta \cup \{f\} \cup D]\right) = 0.26 - 0.26 = 0.00$$

- Chọn  $c$  vì  $SIG_{R_F^\delta}(c)$  lớn nhất, do đó  $R_F^\delta = R_F^\delta \cup \{c\} = \{e, c\}$ .

Vì  $\min\left(1 - \sup_{i,j=1}^6 |\mu(b_{ij}) - \mu(c_{ij})|, 1 - \sup_{i,j=1}^6 |\nu(b_{ij}) - \nu(c_{ij})|\right) = 0.76 < \delta$  hay  $[R_F^\delta \cup D] \neq \overset{\delta}{[C \cup D]}$ , do đó tiếp tục vòng lặp ta có:

- Tính:

$$SIG_{R_F^\delta}(a) = \tilde{d}\left([R_F^\delta], [R_F^\delta \cup D]\right) - \tilde{d}\left([R_F^\delta \cup \{a\}], [R_F^\delta \cup \{a\} \cup D]\right) = 0.18 - 0.15 =$$

0.03

$$SIG_{R_F^\delta}(b) = \tilde{d} \left( [R_F^\delta], [R_F^\delta \cup D] \right) - \tilde{d} \left( [R_F^\delta \cup \{b\}], [R_F^\delta \cup \{b\} \cup D] \right) = 0.18 - 0.18 =$$

0.00

$$SIG_{R_F^\delta}(d) = \tilde{d} \left( [R_F^\delta], [R_F^\delta \cup D] \right) - \tilde{d} \left( P_{R_F^\delta \cup \{d\}}^\approx, [R_F^\delta \cup d \cup D] \right) = 0.18 - 0.18 = 0.00$$

$$SIG_{R_F^\delta}(f) = \tilde{d} \left( [R_F^\delta], [R_F^\delta \cup D] \right) - \tilde{d} \left( [R_F^\delta \cup \{f\}], [R_F^\delta \cup \{f\} \cup D] \right) = 0.18 - 0.18 =$$

0.00

- Chọn  $a$  vì  $SIG_{R_F^\delta}(a)$  lớn nhất, do đó

$$R_F^\delta = R_F^\delta \cup \{a\} = \{e, c, a\}.$$

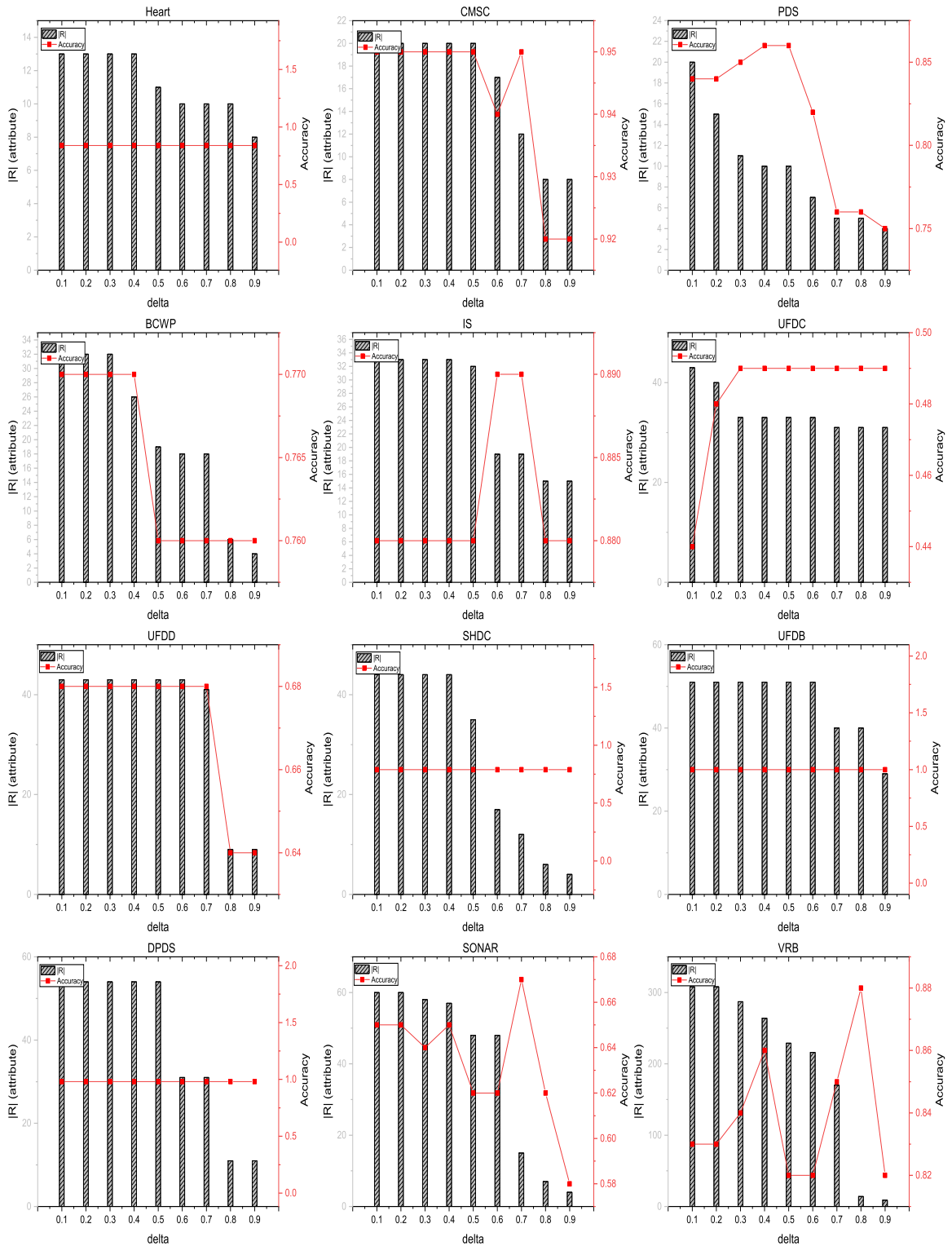
Vì  $\min \left( 1 - \sup_{i,j=1}^6 |\mu(b_{ij}) - \mu(c_{ij})|, 1 - \sup_{i,j=1}^6 |\nu(b_{ij}) - \nu(c_{ij})| \right) = 0.8 = \delta$  hay  $[R_F^\delta \cup D] \stackrel{\delta}{=} [C \cup D]$ . Theo giả thiết ban đầu, reduct ứng viên  $R_F^{0.8}$  có độ chính xác cao nhất trong các ngưỡng  $\delta$  do đó  $R_W^\delta = \{e, c, a\}$ . Kết thúc giao đoạn filter -  $W_\delta$ . Thuật toán chuyển đến giai đoạn  $W_A$  của thuật toán IFD.

*Giai đoạn  $W_A$ :*

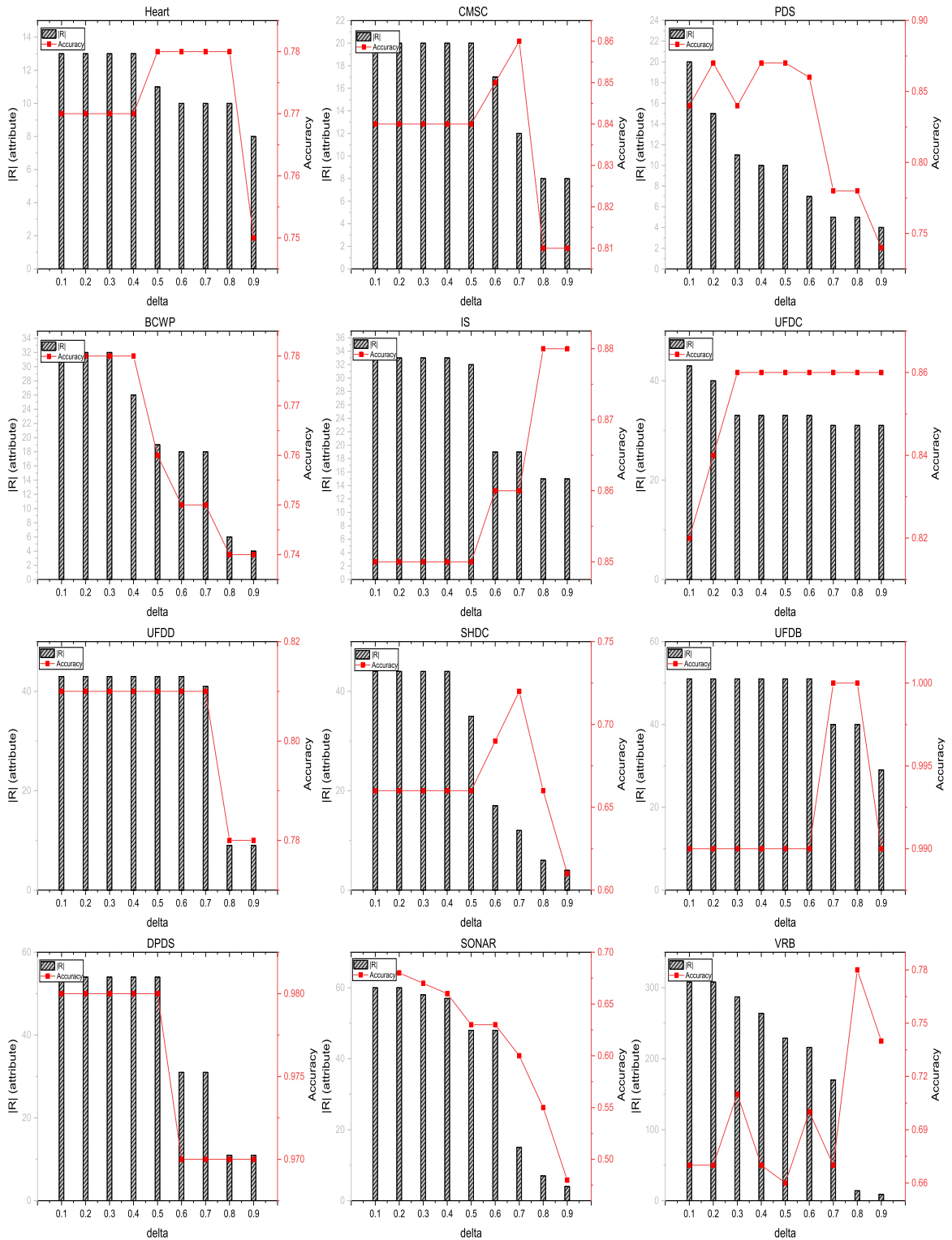
Tập rút gọn ứng viên  $R_F^\delta$  được chia thành các tập con  $\{e, c\}$  tương ứng với khoảng  $[1 : 2]$  và tập con  $\{e, c, a\}$  tương ứng với khoảng  $[1 : 3]$  của tập thuộc tính  $R_W^\delta$ . Chọn tập con tập thuộc tính có khả năng phân lớp cao nhất trên mô hình phân lớp *Model*. Giả sử  $\{e, c\}$  là tập thuộc tính con có khả năng phân lớp cao nhất, khi đó  $R_W^A = \{e, c\}$ .

### 2.3.2. Thực nghiệm và đánh giá thuật toán

Trong các kết quả nghiên cứu của A.Tan và các cộng sự đã chỉ ra phương pháp giảm thuộc tính theo tiếp cận RS mờ trực cảm hiệu quả hơn tiếp cận RS mờ truyền thống về khả năng phân lớp. Do đó, 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] sử dụng độ đo miền dương mờ trực cảm (Intuitionistic Fuzzy POS - IFPOS[36]) và thuật toán [15] sử dụng độ đo Entropy mờ trực cảm (Intuitionistic Fuzzy Information Entropy - IFIE[15]).



**Hình 2.1:** Tác động của  $\delta$  tới số lượng phần tử và khả năng phân lớp trên mô hình phân lớp SVM



**Hình 2.2:** Tác động của  $\delta$  tới số lượng phần tử và khả năng phân lớp trên mô hình phân lớp KNN

**Bảng 2.2:** Bảng mô tả các tập dữ liệu thực nghiệm

STT	Dataset	Mô tả.	U	C	D
1	Heart	Statlog (Heart)	270	13	2
2	CMSC	Climate Model Simulation Crashes Data Set	540	18	2
3	PDS	Parkinsons Data Set	196	22	2
4	BCWP	Breast Cancer Wisconsin (Prognostic)	198	33	2
5	IS	Ionosphere	351	34	2
6	UFDC	Ultrasonic flowmeter diagnostics (C)	181	43	4
7	UFDD	Ultrasonic flowmeter diagnostics (D)	181	43	4
8	SHDC	SPECTF Heart Data Set	267	44	2
9	UFDB	Ultrasonic flowmeter diagnostics (B)	92	51	3
10	DPDS	Divorce Predictors data set	170	54	2
11	Sona	Connectionist Bench	208	60	2
12	VRB	Voice Rehabilitation(Binary)	126	310	2

### 2.3.2.1. Môi trường thực nghiệm

Các thuật toán được cài đặt bằng ngôn ngữ lập trình Python và chạy trên nền hệ điều hành Window 10 với cấu hình phần cứng là bộ xử lý Core I5, ram 8GB. Cùng với 12 tập dữ liệu thử nghiệm được tải về từ UCI được mô tả chi tiết trong Bảng 2.2. Trong đó  $|U|$  là số lượng mẫu,  $|C|$  là số thuộc tính điều kiện và  $|D|$  là số phân lớp của thuộc tính quyết định trong mỗi tập dữ liệu (dataset).

Các tập dữ liệu thử nghiệm đều là các dataset có thuộc tính điều kiện miền giá trị số liên tục. Do đó, trước khi thực hiện thuật toán giảm thuộc tính, miền giá trị của các thuộc tính sẽ được chuẩn hóa về đoạn  $[0,1]$ . Độ tương tự (hàm thuộc) của  $x, y \in U$  theo thuộc tính  $a$  được xác định bởi công thức  $\mu = R_a(x, y) = 1 - |a(x) - a(y)|$ , độ khác biệt (hàm không thuộc) được tính theo công thức  $\nu = \frac{1-\mu}{1+\lambda\mu}$  với  $\lambda > 0$ , khi  $\lambda = 0$  công thức này suy biến về tập mờ tuyến thống, giá trị của  $\lambda$  càng tăng thì giá trị của  $\nu$  càng giảm. Tuy nhiên, với sự đa dạng của dữ liệu, việc phân bố dữ liệu trong các tập dữ liệu là rất khác nhau. Do đó để chọn được hệ số  $\lambda$  dùng chung cho toàn bộ các tập dữ liệu là rất khó, đặc biệt là trên các **bộ dữ liệu có khả năng phân lớp ban đầu thấp**.

Do đó, chương này đề xuất công thức tính giá trị  $\lambda$  như sau:

$$\lambda_a = \begin{cases} 1 \Leftrightarrow \sigma_a = 0 \\ \frac{\beta_a}{\sigma_a} \Leftrightarrow \sigma_a > 0 \end{cases} \quad (2.4)$$

Trong đó  $\sigma_a = \sqrt{\frac{1}{n-1} \sum_1^n (a(y_i) - \bar{a})^2}$  là độ lệch chuẩn của miền giá trị thuộc tính  $a$  và  $\beta_a = \frac{|P_{\{a\} \cup \{d\}}^F|}{|P_{\{d\}}^F|}$  là độ nhất quán của thuộc tính  $a$  trong bảng quyết định.

Để đánh giá độ chính xác phân lớp của reduct. Chương này sử dụng hai mô hình phân lớp dữ liệu số là SVM và k-NN(k=|D|). Độ đo đánh giá và phương pháp đánh giá độ chính xác trên các mô hình là độ đo *Accuracy* và phương pháp đánh giá chéo *10-fold* được sử dụng chung cho toàn bộ các reduct thu được từ các thuật toán.

### 2.3.2.2. Kịch bản thực nghiệm

Nhằm khẳng định phương pháp giảm thuộc tính đề xuất là hiệu quả hơn về khả năng phân lớp so với một số phương pháp giảm thuộc tính khác của A.Tan và các cộng sự đề xuất, chương này tiến hành thực nghiệm thuật toán đề xuất IFD theo các kịch bản như sau:

1) Lựa chọn giá trị  $\delta$  tốt nhất cho thuật toán. Thuật toán đề xuất có hai bước wrapper là wrapper\_delta (W\_delta) và wrapper thuộc tính (W\_A), trong đó bước W\_delta được thực hiện trước bước W\_A. Như vậy, với mỗi bộ dữ liệu khác nhau sẽ có một giá trị  $\delta$  khác nhau sao cho reduct thu được sẽ được tối ưu cả về khả năng phân lớp và số lượng phần tử.

2) Đánh giá reduct của thuật toán đề xuất IFD với các thuật toán IFPOS[36], IFIE[15]. Trong đó các tiêu chí được sử dụng để so sánh và đánh giá bao gồm khả năng phân lớp (accuracy), số lượng phần tử của reduct ( $|R|$ ) và chi phí tính toán của thuật toán (second).

**Bảng 2.3:** Mô tả sự tương quan về số lượng phần tử và khả năng phân lớp chính xác của reduct tại hai giai đoạn wrapper trên mô hình phân lớp SVM

ID	Dataset	U	R			Accuracy		
			C	$W_\delta$	$W_A$	C	$W_\delta$	$W_A$
1	Heart	270	13	8	7	0.84	0.84	0.84
2	CMSC	540	20	12	11	0.95	0.95	0.95
3	PDS	195	22	10	9	0.84	0.86	0.86
4	BCWP	198	32	26	25	0.77	0.77	0.77
5	IS	351	34	19	16	0.88	0.89	0.89
6	UFDC	181	43	31	26	0.44	0.49	0.52
7	UFDD	180	43	41	27	0.68	0.68	0.68
8	SHDC	267	44	4	2	0.79	0.79	0.79
9	UFDB	92	51	29	2	1	1	1
10	DPDS	170	54	11	5	0.98	0.98	0.98
11	Sonar	208	60	15	11	0.65	0.67	0.7
12	VRB	126	310	14	11	0.83	0.88	0.88

**Bảng 2.4:** Mô tả sự tương quan về số lượng phần tử và khả năng phân lớp chính xác của reduct tại hai giai đoạn wrapper trên mô hình phân lớp KNN

ID	Dataset	U	R			Accuracy		
			C	$W_\delta$	$W_A$	C	$W_\delta$	$W_A$
1	heart	270	13	10	9	0.77	0.78	0.78
2	CMSC	540	20	12	11	0.84	0.86	0.86
3	PDS	195	22	10	7	0.85	0.87	0.87
4	BCWP	198	32	26	21	0.78	0.78	0.79
5	IS	351	34	15	5	0.85	0.88	0.92
6	UFDC	181	43	31	29	0.82	0.86	0.86
7	UFDD	180	43	41	25	0.81	0.81	0.84
8	SHDC	267	44	12	9	0.66	0.72	0.72
9	UFDB	92	51	29	2	0.99	0.99	1
10	DPDS	170	54	11	7	0.98	0.97	0.97
11	sonar	208	60	48	31	0.68	0.63	0.69
12	VRB	126	310	14	12	0.68	0.78	0.82

### 2.3.2.3. Lựa chọn giá trị $\delta$

Trước tiên, Để minh họa sự tác động của giá trị  $\delta$  tới số lượng phần tử của reduct thu được từ thuật toán đề xuất. Chương này xây dựng biểu đồ 2.1 và biểu đồ 2.2 để

minh họa sự tác động của  $\delta$  tới số lượng phần tử và khả năng phân lớp chính xác của reduct thu được trên mỗi mô hình phân lớp dữ liệu.

Thông quan sự biến động về số lượng phần tử và khả năng phân lớp của các tập thuộc tính con thu được từ sự thay đổi của các giá trị  $\delta$ , ta có thể thấy khi giá trị  $\delta$  càng tăng, số lượng phần tử của reduct càng giảm. Điều đó cho thấy sự thay đổi tuyến tính về số lượng phần tử của reduct so với giá trị thay đổi của  $\delta$ . Tuy nhiên khả năng phân lớp trên cả hai mô hình phân lớp lại không tuyến tính với sự thay đổi của giá trị  $\delta$ .

#### 2.3.2.4. Đánh giá reduct của các thuật toán

Bảng 2.3 và Bảng 2.4 mô tả số lượng phần tử của các reduct ứng viên tại giai đoạn  $W_\delta$  và reduct tại giai đoạn  $W_A$  tương ứng với hai mô hình phân lớp k-NN và SVM. Sau giai đoạn  $W_\delta$  ta thu được reduct ứng viên mức delta, sau giai đoạn  $W_A$  ta thu được reduct thực sự của thuật toán IFD. Ta có thể thấy số lượng phần tử của reduct thực sự nhỏ hơn đáng kể so với số lượng phần tử của reduct ứng viên. Đặc biệt trên mô hình phân lớp KNN, reduct thực sự không những cải thiện cả về số lượng phần tử mà độ chính xác của reduct thu được cũng hiệu quả hơn so với reduct ứng viên của thuật toán.

Quan sát Bảng 2.5 ta có thể thấy số lượng phần tử trung bình của reduct trên toàn bộ các tập dữ liệu không chênh lệch quá nhiều so với thuật toán giảm thuộc tính theo tiếp cận IFPOS[36] nhưng lại tốt hơn so với thuật toán theo tiếp cận IFIE[15]. Trên hai mô hình phân lớp SVM và KNN ta có thể thấy số lượng phần tử reduct được wrapper theo mô hình phân lớp SVM là tốt hơn mô hình phân lớp KNN.

Mặc dù số lượng phần tử trung bình của các reduct thu được từ các thuật toán không chênh lệch nhau đáng kể nhưng quan sát trên các bộ dữ liệu có số lượng phần tử lớn như **CMSC** và bộ dữ liệu có số chiều lớn như **VRB** ta có thể thấy số lượng phần tử reduct thu được từ thuật toán đề xuất là tốt hơn hẳn so với hai thuật toán còn lại. Để có thể quan sát trực quan hơn sự khác biệt về reduct thu được từ các thuật toán trên



**Bảng 2.5:** Mô tả số lượng phần tử thu được của reduct 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

tung tập dữ liệu, chúng ta có thể quan sát Hình 2.3 và Hình 2.4 để biết thêm thông tin chi tiết.

**Bảng 2.6:** So sánh khả năng phân lớp của các reduct 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 2.6 và Bảng 2.7 mô tả sự chênh lệch về khả năng phân lớp của các reduct tương ứng với các mô hình phân lớp SVM và KNN. Chúng ta có thể thấy trên mô hình phân lớp SVM, thuật toán đề xuất cho reduct có khả năng phân lớp trung bình trên

**Bảng 2.7:** So sánh khả năng phân lớp của các reduct 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	<b>UFDC</b>	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	<b>SHDC</b>	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	<b>sonar</b>	208	68±0.8	69±0.1	62±0.9	60±0.9
12	<b>VRB</b>	126	68±0.6	82±0.7	81±0.7	65±0.2

các bộ dữ liệu là tương đương với thuật toán theo tiếp cận IFPOS[36] và tốt hơn thuật toán theo tiếp cận IFE là 3%.

Thuật toán đề xuất IFD và thuật toán IFPOS[36] đều cho reduct có độ chính xác cao hơn so với tập dữ liệu gốc (Raw). Mặc dù, độ chính xác trung bình trên các bộ dữ liệu rút gọn của hai thuật toán này là như nhau nhưng khả năng cải thiện nhiều trên thuật toán đề xuất IFD là tốt hơn thuật toán IFPOS[36]. Cụ thể với bộ dữ liệu **UFDC** có khả năng phân lớp ban đầu là 0.44 (44%), thuật toán IFD cho ra reduct có độ chính xác 0.52 (52%) và thuật toán IFPOS[36] cho reduct có độ chính xác 0.49 (49%). Trên bộ dữ liệu Sona cả hai thuật toán đều cải thiện nhiều tốt như nhau tăng từ 0.65 (65%) lên 0.7 (70%). Đối với mô hình phân lớp KNN, thuật toán đề xuất IFD cho reduct có khả năng phân lớp trung bình trên toàn bộ dữ liệu trội hơn so với hai thuật toán còn lại. Với các **bộ dữ liệu có khả năng phân lớp ban đầu thấp** như **SHDC**, **Sona**, **VRB** đã được cải thiện hiệu quả về khả năng phân lớp với thuật toán IFD. Tuy nhiên, Bảng 2.8 cho thấy chi phí tính toán của thuật toán đề xuất IFD còn thấp hơn so với thuật toán IFPOS[36] và thấp hơn đáng kể so với thuật toán IFIE[15].

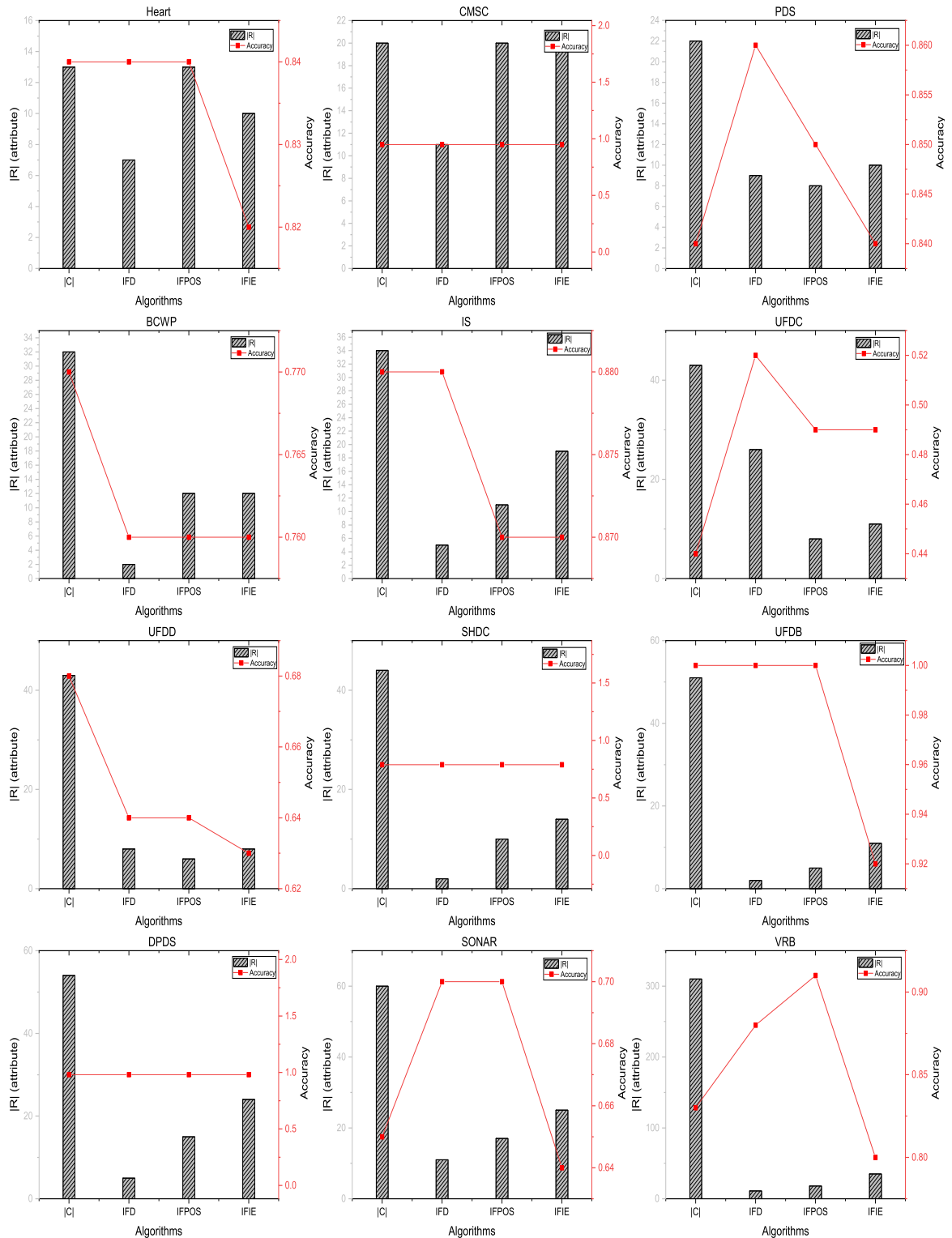
Nhìn chung, tiêu chí khả năng phân lớp và số lượng phần tử của reduct của hai

**Bảng 2.8:** Mô tả chi phí tính toán của các thuật toán

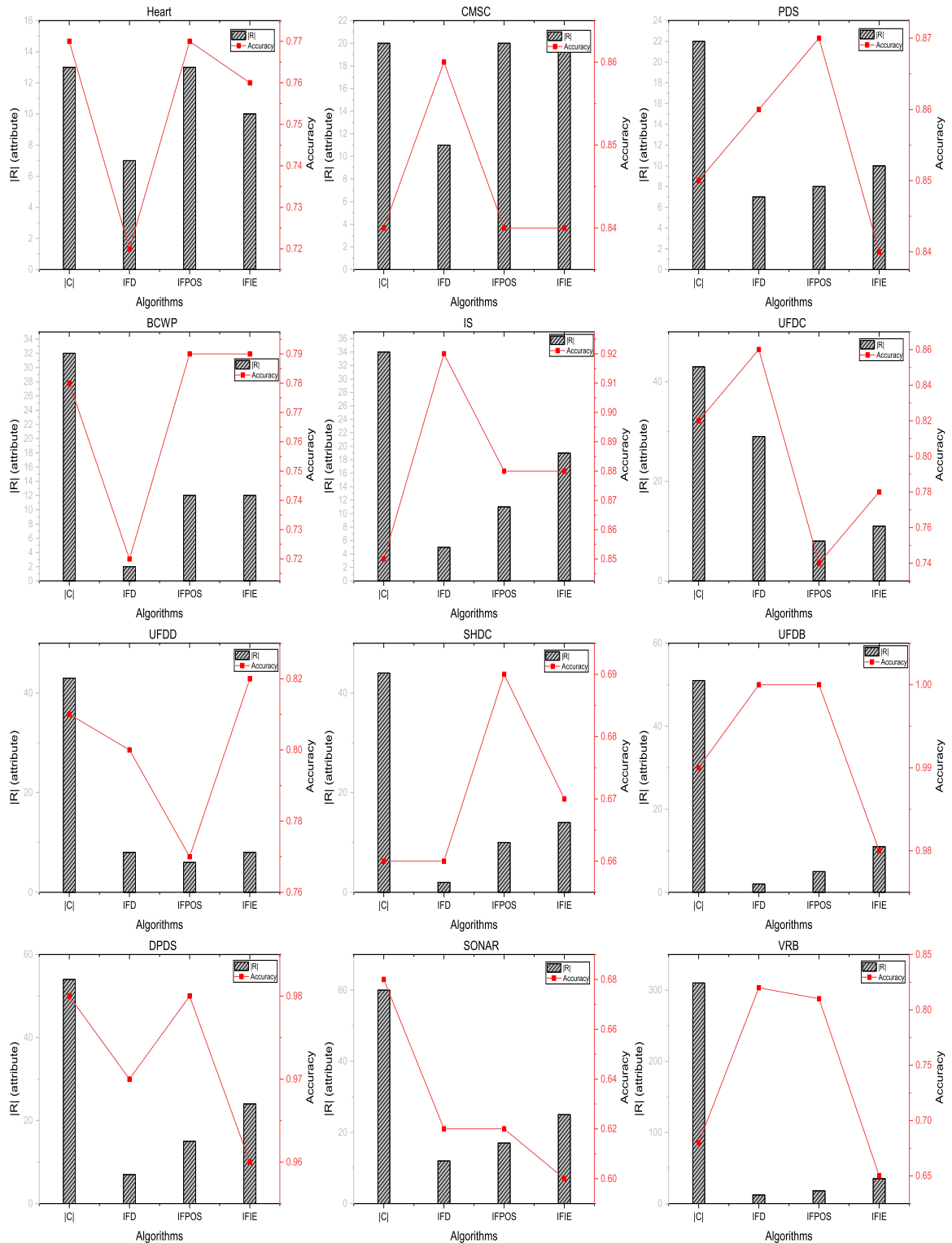
ID	Dataset	U	C	Computation time (second)			
				IFD-SVM	IFD-KNN	IFPOS[36]	IFIE[15]
1	heart	270	13	7.89	7.91	1.58	2.06
2	CMSC	540	18	53.33	54.25	36.56	12.28
3	PDS	195	22	7.67	7.63	6.18	2.46
4	BCWP	198	33	13.09	12.3	8.26	4.28
5	IS	351	34	44.94	43.94	24.34	12.16
6	UFDC	181	43	14.83	13.43	13.46	4.98
7	UFDD	180	43	13.08	13.64	10.86	4.22
8	SHDC	267	44	29.75	32.52	18.32	9.6
9	UFDB	92	51	4.5	5.02	2.78	2.44
10	DPDS	170	54	13.83	16.14	12.84	8.68
11	sonar	208	60	27.12	29.1	23.26	13.3
12	VRB	126	310	52.86	49.83	57.34	57.18

thuật toán IFD và IFPOS[36] là tốt nhất. Trong khi đó, hầu hết các thuật toán giảm thuộc tính theo tiếp cận độ đo dựa trên RS và RS mở rộng lâu nay vẫn được chứng minh là tiếp cận bảo toàn thông tin tốt nhất so với các tiếp cận độ đo khác. Điều đó càng khẳng định phương pháp giảm thuộc tính theo tiếp cận độ đo khoảng cách mờ trực cảm IFD đề xuất là hiệu quả và đáng được quan tâm. Sau đây là các phân tích về nguyên nhân ảnh hưởng tới chi phí tính toán của thuật toán IFD, số lượng phần tử và khả năng phân lớp chính xác của reduct thu được bởi thuật toán IFD

- Kích thước của reduct: Các kết quả so sánh trong Bảng 2.5 cũng như sự chênh lệch của hai giai đoạn wrapper đã được phân tích bên trên cho thấy giai đoạn  $W_A$  có vai trò quan trọng trong việc giảm thuộc tính trong khi vẫn giữ được khả năng phân lớp tốt nhất của giai đoạn  $W_S$ . Hơn nữa quan sát các Hình 2.1 và 2.2 chúng ta có thể thấy rõ về sự tương quan giữa số lượng phần tử và khả năng phân lớp chính xác của reduct. Hầu hết các tập dữ liệu được phân tích cho thấy reduct có số lượng phần tử lớn hơn chưa chắc có khả năng phân lớp cao hơn so với reduct có số lượng phần tử nhỏ hơn. Đây chính là nguyên nhân chương này đưa giai đoạn  $W_A$  vào sau giai đoạn  $W_S$  trong phương pháp giảm thuộc tính đề xuất.



**Hình 2.3:** Sự tương quan về số lượng phần tử và khả năng phân lớp chính xác của reduct thu được từ các thuật toán trên mô hình phân lớp SVM



**Hình 2.4:** Sự tương quan về số lượng phần tử và khả năng phân lớp chính xác của reduct thu được từ các thuật toán trên mô hình phân lớp KNN

- Khả năng phân lớp chính xác của reduct: Bên cạnh việc sử dụng tập nền IFS để xây dựng độ đo đánh giá độ quan trọng của thuộc tính. Giai đoạn  $W_\delta$  của thuật toán giảm thuộc tính có ảnh hưởng quan trọng đến khả năng phân lớp chính xác của reduct thực sự thu được từ thuật toán. Với mỗi giá trị  $\delta$  khác nhau, reduct ứng viên thu được có thể có số lượng phần tử và khả năng phân lớp khác nhau. Thông qua giai đoạn  $W_\delta$ , thuật toán đề xuất đã lọc bỏ đi phần lớn các thuộc tính không cần thiết, giảm thời gian tính toán cho giai đoạn  $W_A$  của thuật toán đề xuất.

- Khả năng cải thiện nhiều: Cải thiện nhiều tốt cũng chính là tăng khả năng phân lớp cho reduct. Bên cạnh yếu tố về tập nền IFS có khả năng cải thiện nhiều như đã được phân tích trong các công trình nghiên cứu của A.Tan và các cộng sự [15], [36] thì cách thức xây dựng công thức tính độ thuộc và độ không thuộc theo tiếp cận độ nhất quán của thuộc tính trong phần thực nghiệm của Chương cũng ảnh hưởng quan trọng tới việc cải thiện nhiều, tăng khả năng phân lớp cho tập dữ liệu.

- Thời gian tính toán: Chi phí tính toán của thuật toán giảm thuộc tính đề xuất IFD còn hạn chế về mặt thời gian tính toán so với các phương pháp giảm thuộc tính khác. Nguyên nhân chính là sự ảnh hưởng của hai giai đoạn wrapper của thuật toán đề xuất. giai đoạn  $W_\delta$  bị phụ thuộc vào độ phức tạp của độ đo đề xuất và số lượng giá trị  $\delta$  cần xét. Đây là sự tương quan tuyến tính. Giai đoạn  $W_A$  có thời gian tính toán phụ thuộc vào mô hình phân lớp và số lượng phần tử của reduct ứng viên. Nếu reduct ứng viên có số lượng phần tử nhỏ thì giai đoạn  $W_A$  sẽ thực hiện nhanh và ngược lại.

**Bảng 2.9:** Mô tả reduct thu được từ các thuật toán

ID	Dataset	Reducts			
		IFD-SVM	IFD-KNN	IFPOS[36]	IFIE[15]
1	heart	[12, 6, 1, 2, 8, 5, 11]	[12, 6, 1, 2, 8, 5, 11, 10, 9]	[0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12]	[6, 1, 8, 12, 5, 10, 2, 11, 0, 3]
Tiếp theo trang sau					

**Bảng 2.9 – Tiếp theo trang trước**

ID	Dataset	Reducts			
		IFD-SVM	IFD-KNN	IFPOS[36]	IFIE[15]
2	CMSC	[0, 2, 1, 19, 3, 18, 17, 4, 16, 15, 14]	[0, 2, 1, 19, 3, 18, 17, 4, 16, 15, 14]	[2, 3, 15, 0, 5, 1, 6, 4, 7, 9, 13, 8, 10, 11, 12, 14, 16, 17, 18, 19]	[0, 1, 4, 6, 5, 7, 10, 8, 9, 3, 11, 17, 12, 2, 13, 14, 15, 16, 18, 19]
3	PDS	[0, 2, 16, 1, 3, 21, 10, 20, 19]	[0, 2, 16, 1, 3, 21, 10]	[18, 0, 10, 16, 17, 2, 19, 20]	[16, 2, 17, 1, 0, 10, 3, 19, 6, 20]
4	BCWP	[0, 1, 31, 6, 30, 29, 11, 28, 27, 26, 25, 12, 24, 23, 22, 19, 21, 20, 18, 17, 16, 15, 14, 13, 10]	[0, 1, 31, 6, 30, 29, 11, 28, 27, 26, 25, 12, 24, 23, 22, 19, 21, 20, 18, 17, 16]	[0, 4, 2, 5, 6, 19, 11, 9, 12, 17, 31, 20]	[0, 31, 6, 19, 1, 11, 22, 5, 20, 8, 12, 18]
5	IS	[14, 0, 2, 4, 7, 8, 33, 11, 32, 31, 30, 24, 29, 28, 27, 26]	[14, 0, 2, 4, 7]	[0, 4, 2, 5, 27, 30, 7, 3, 9, 16, 17]	[14, 0, 27, 28, 7, 31, 18, 23, 26, 4, 2, 17, 3, 5, 9, 10, 24, 12, 19]
Tiếp theo trang sau					

**Bảng 2.9 – Tiếp theo trang trước**

ID	Dataset	Reducts			
		IFD-SVM	IFD-KNN	IFPOS[36]	IFIE[15]
6	UFDC	[3, 23, 27, 8, 18, 42, 41, 40, 39, 38, 37, 36, 35, 34, 0, 33, 32, 31, 30, 2, 29, 28, 26, 25, 24, 22]	[3, 23, 27, 8, 18, 42, 41, 40, 39, 38, 37, 36, 35, 34, 0, 33, 32, 31, 30, 2, 29, 28, 26, 25, 24, 22, 21, 20, 19]	[7, 9, 25, 5, 27, 0, 39, 11]	[3, 23, 8, 25, 0, 2, 27, 15, 29, 33, 31]
7	UFDD	[33, 5, 21, 27, 42, 41, 40, 39, 38, 37, 36, 35, 34, 32, 31, 30, 29, 28, 26, 25, 24, 23, 22, 20, 19, 18, 17]	[33, 5, 21, 27, 42, 41, 40, 39, 38, 37, 36, 35, 34, 32, 31, 30, 29, 28, 26, 25, 24, 23, 22, 20, 19]	[25, 27, 17, 3, 11, 42]	[5, 27, 21, 42, 39, 31, 0, 41]
8	SHDC	[25, 1]	[25, 1, 2, 21, 43, 42, 41, 29, 40]	[40, 29, 1, 2, 13, 25, 3, 4, 18, 9]	[43, 3, 18, 1, 12, 14, 21, 9, 29, 36, 23, 2, 24, 41]
Tiếp theo trang sau					



**Bảng 2.9 – Tiếp theo trang trước**

ID	Dataset	Reducts			
		IFD-SVM	IFD-KNN	IFPOS[36]	IFIE[15]
9	UFDB	[41, 14]	[41, 14]	[41, 14, 13, 16, 12]	[35, 6, 12, 39, 31, 0, 19, 22, 23, 3, 43]
10	DPDS	[39, 34, 0, 53, 15]	[39, 34, 0, 53, 15, 52, 51]	[10, 32, 30, 0, 48, 15, 6, 39, 35, 3, 19, 43, 42, 27, 46]	[44, 45, 3, 51, 34, 5, 30, 46, 41, 6, 48, 2, 36, 52, 43, 0, 38, 21, 42, 1, 31, 47, 53, 27]
11	sonar	[19, 16, 22, 25, 34, 28, 44, 59, 58, 57, 35]	[19, 16, 22, 25, 34, 28, 44, 59, 58, 57, 35, 56, 55, 54, 53, 31, 52, 51, 50, ...]	[0, 11, 15, 36, 26, 19, 21, 9, 53, 23, 24, 28, 6, 30, 32, 35, 44]	[19, 25, 16, 22, 34, 27, 29, 31, 36, 53, 44, 0, 5, 9, 11, 17, 20, 7, 24, 18, 26, 28, ...]
12	VRB	[59, 70, 1, 69, 34, 41, 62, 79, 57, 309, 83]	[59, 70, 1, 69, 34, 41, 62, 79, 57, 309, 83, 308]	[54, 84, 79, 83, 91, 41, 3, 16, 46, 52, 70, 34, 55, 59, 62, 64, 90, 69]	[58, 70, 59, 60, 62, 92, 55, 69, 127, 57, 107, 121, 138, 25, ...]

## 2.4. Kết luận Chương 2

Chương 2, luận án trình bày về một phương pháp giảm thuộc tính theo tiếp cận IFRS. Các đóng góp chính của Chương này gồm có:

- Đề xuất độ đo khoảng cách mờ trực cảm là cơ sở để xây dựng độ đo đánh giá độ quan trọng của thuộc tính
- Đề xuất thuật toán tìm reduct trong bảng quyết định số với định nghĩa mới về reduct theo tiếp cận  $\delta$  - equal.

Bên cạnh đó, phương pháp xây dựng hàm thành viên và hàm không thành viên cho không gian xấp xỉ mờ trực cảm theo tiếp cận độ nhất quán của thuộc tính do tác giả đề xuất cũng là nhân tố quan trọng ảnh hưởng tới khả năng chọn lọc thuộc tính cho reduct.

Các kết quả thực nghiệm cho thấy thuật toán đề xuất cho các reduct hiệu quả về số lượng phần tử và khả năng phân lớp trên hầu hết các tập dữ liệu so với các thuật toán theo tiếp cận IFRS khác. Tuy nhiên chi phí tính toán của thuật toán đề xuất còn hạn chế do phải đánh đổi về số lượng phần tử và khả năng phân lớp cho reduct, đặc biệt là mục tiêu nâng cao chất lượng phân lớp cho các bộ dữ liệu có khả năng phân lớp ban đầu thấp.

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

### **3.1. Mở đầu**

Topology là một nhánh toán học mà các khái niệm của nó xuất hiện phổ biến trong các lĩnh vực toán học khác và trong các ứng dụng cuộc sống. Theo góc nhìn của lý thuyết RS, cấu trúc tôpô là nền tảng toán học quan trọng trong quá trình trích rút, phân tích và xử lý thông tin [37]. Do đó, việc kết hợp lý thuyết RS và tôpô sẽ tăng khả năng xử lý trên các tập dữ liệu phức tạp, trong đó các bộ dữ liệu có số lượng phần tử lớn, số chiều cao, chứa nhiều và không đầy đủ xuất hiện ngày càng nhiều trong cuộc sống. Do đó hướng giảm thuộc tính theo tiếp cận tôpô trên nền RS ngày càng nhận được nhiều quan tâm từ các nhà nghiên cứu [38], [40]–[42].

Năm 2005, Lashin và các cộng sự lần đầu tiên giới thiệu khái niệm tôpô rút gọn theo tiếp cận rough set [37]. Từ đó, phương pháp xây dựng cấu trúc tôpô theo tiếp cận RS được nhiều nhà nghiên cứu quan tâm và đề xuất. Hiện nay có hai phương pháp xây dựng tôpô theo tiếp cận RS gồm có, các phương pháp xây dựng tôpô từ không gian xấp xỉ của RS [38], [40]–[42], các phương pháp xây dựng tôpô từ các phép toán xấp xỉ của RS [43].

Năm 2014, Yu và các cộng sự [38] đã chỉ hai phân hoạch khác nhau có thể có cùng cấu trúc tôpô do đó giảm thuộc tính theo tiếp cận bảo toàn cấu trúc có cho tập rút gọn có số lượng phần tử tối ưu hơn so với tiếp cận độ đo truyền thống. Bên cạnh đó, khả năng cải thiện nhiều của tập mờ trực cảm trên các bộ dữ liệu có khả năng phân lớp ban đầu thấp như đã trình bày trong Chương 2.

Trên cơ sở đó, Chương này đề xuất phương pháp giảm 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 giảm thuộc tính trên cơ sở tính chất tương đồng của hai tôpô thông và định nghĩa reduct theo cấu trúc tôpô đơn vị. Cuối cùng là thực nghiệm và so sánh các thuật toán đề xuất với một số thuật toán theo tiếp cận IFRS trên các bộ dữ liệu được tải về từ UCI.

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** (Quan hệ mờ trực cảm). [58] Cho bảng quyết định  $DT = (U, C, D, f)$  và quan hệ mờ trực cảm  $IFR$  xác định trên  $U$ . Khi đó  $IFR$  được gọi là quan hệ tương đương mờ trực cảm nếu các điều kiện sau đây thỏa mãn:

- (1) Tính phản xạ:  $IFR(x, x) = 1_{IF}$  với mọi  $x \in U$ .
- (2) Tính đối xứng:  $IFR(x, y) = IFR(y, x)$  với mọi  $x, y \in U$ .
- (3) Tính bắc cầu:  $R(x, y) \wedge R(y, z) \leq R(x, z)$  với mọi  $x, y, z \in U$ .

Khi đó,  $IFR$  được gọi là quan hệ ưu tiên mờ trực cảm nếu các tính chất (1) và (3) thỏa mãn.

**Định nghĩa 3.2** (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:

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

$$\nu_y = 1 - \mu_y$$

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 3.3** (Ma trận quan hệ ưu tiên mờ trực cảm). Cho bảng quyết định  $DT = (U, C, D, f)$  và quan hệ ưu tiên mờ trực cảm  $IFR_a^{\geq}$  tương ứng với thuộc tính  $a \in C$  xác định trên  $U$ . Khi đó  $IFR_a^{\geq}$  có thể được biểu diễn bởi ma trận quan hệ  $M_a^{\geq} = [i, j]_{n \times n}$ .

**Ví dụ 3.1.** Xét bảng quyết định cho trong Bảng 1.3, với  $\delta = 0.5$ , ta có:

$$M_a^{\geq} = \begin{bmatrix} (1,0) & (1,0) & (0.8,0.2) & (0.2,0.8) & (0.2,0.8) & (0.2,0.8) \\ (1,0) & (1,0) & (0.8,0.2) & (0.2,0.8) & (0.2,0.8) & (0.2,0.8) \\ (0,1) & (0,1) & (1,0) & (0.4,0.6) & (0.4,0.6) & (0.4,0.6) \\ (0,1) & (0,1) & (0,1) & (1,0) & (1,0) & (1,0) \\ (0,1) & (0,1) & (0,1) & (1,0) & (1,0) & (1,0) \\ (0,1) & (0,1) & (0,1) & (1,0) & (1,0) & (1,0) \end{bmatrix}$$

**Định nghĩa 3.4** (Hợp hai ma trận). Cho bảng quyết định  $DT = (U, C, D, f)$  và hai ma trận quan hệ ưu tiên mờ trực cảm  $M_a^{\geq}, M_b^{\geq}$  tương ứng của  $a, b \in C$  xác định trên  $U$ . Khi đó hợp của hai ma trận được định nghĩa bởi

$$M_a^{\geq} \cup M_b^{\geq}[i, j] = \max(M_a^{\geq}[i, j], M_b^{\geq}[i, j]) \quad (3.2)$$

**Định nghĩa 3.5** (Giao hai ma trận). Cho bảng quyết định  $DT = (U, C, D, f)$  và hai ma trận quan hệ ưu tiên mờ trực cảm  $M_a^{\geq}, M_b^{\geq}$  tương ứng của  $a, b \in C$  xác định trên  $U$ . Khi đó giao của hai ma trận được định nghĩa bởi

$$M_a^{\geq} \cap M_b^{\geq}[i, j] = \min(M_a^{\geq}[i, j], M_b^{\geq}[i, j]) \quad (3.3)$$

**Định nghĩa 3.6** (Cổ 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:

$$S_a = \{S_a^L, S_a^R\} \quad (3.4)$$

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  $(M_a^{\geq})^T$  trên thuộc tính  $a \in C$ , với  $(M_a^{\geq})^T$  là ma trận chuyển vị của ma trận  $M_a^{\geq}$ .

**Ví dụ 3.2.** Tiếp theo Ví dụ 3.1, thực hiện phép lấy đối xứng ta có:

$$(M_a^{\geq})^T = \begin{bmatrix} (1,0) & (1,0) & (0,1) & (0,1) & (0,1) & (0,1) \\ (1,0) & (1,0) & (0,1) & (0,1) & (0,1) & (0,1) \\ (0.8,0.2) & (0.8,0.2) & (1,0) & (0,1) & (0,1) & (0,1) \\ (0.2,0.8) & (0.2,0.8) & (0.4,0.6) & (1,0) & (1,0) & (1,0) \\ (0.2,0.8) & (0.2,0.8) & (0.4,0.6) & (1,0) & (1,0) & (1,0) \\ (0.2,0.8) & (0.2,0.8) & (0.4,0.6) & (1,0) & (1,0) & (1,0) \end{bmatrix}$$

**Định nghĩa 3.7** (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:

$$S_p \cap S_q = \{S_p^L \cap S_q^L, S_p^R \cap S_q^R\} \quad (3.5)$$

**Định nghĩa 3.8** (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:

$$S_p \cup S_q = \{S_p^L \cup S_q^L, S_p^R \cup S_q^R\} \quad (3.6)$$

**Định nghĩa 3.9** (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:

$$B_a = S_a^L \cap S_a^R \quad (3.7)$$

**Định nghĩa 3.10** (Tập mờ trực cảm IFT). Cho bảng quyết định  $DT = (U, C, D, f)$  và IF-base  $B_a$  tương ứng với  $a \in C$ . Khi đó IFT  $\mathcal{T}_a$  được định nghĩa bởi:

$$\mathcal{T}_a = \{c : c = \cup\{b \in B\}, B \subseteq B_a\} \quad (3.8)$$

**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 3.7. Khi đó,  $B_a$  là một cơ sở của  $\mathcal{T}_a$ .

*Chứng minh.* : Ta có hai điều phải chứng minh

(1) Theo định nghĩa 3.9 ta có  $B_a = S_a^L \cap S_a^R$  là một tập các quan hệ mờ trực cảm chỉ có tính chất phản xạ và đối xứng do IF-subbase trái  $S_a^L$  và IF-subbase phải  $S_a^R$  là các thành phần đối xứng nhau.

(2) Theo định nghĩa 3.10 về cấu trúc tôpô IFT trên cơ sở IF-base và định nghĩa 1.13 về cơ sở của một tôpô ta thấy rõ ràng  $B_a$  là một IF-base.

Từ (1) và (2) ta có điều phải chứng minh (đpcm) □

**Ví dụ 3.3.** Tiếp theo ví dụ 3.2, thực hiện phép toán giao của hai ma trận đối xứng ta

$$\text{có: } B_a = \begin{bmatrix} (1,0) & (1,0) & (0,1) & (0,1) & (0,1) & (0,1) \\ (1,0) & (1,0) & (0,1) & (0,1) & (0,1) & (0,1) \\ (0,1) & (0,1) & (1,0) & (0,1) & (0,1) & (0,1) \\ (0,1) & (0,1) & (0,1) & (1,0) & (1,0) & (1,0) \\ (0,1) & (0,1) & (0,1) & (1,0) & (1,0) & (1,0) \\ (0,1) & (0,1) & (0,1) & (1,0) & (1,0) & (1,0) \end{bmatrix}$$

**Mệnh đề 3.2** (So sánh hai IF-tôpô). Cho bảng quyết định  $DT = (U, C, D, f)$  và hai tôpô  $\mathcal{T}_p, \mathcal{T}_q$  tương ứng của  $p, q \in C$ . Khi đó,  $\mathcal{T}_p \prec \mathcal{T}_q$  nếu  $B_p \prec B_q$ .

*Chứng minh.* Theo định nghĩa cấu trúc IF-tôpô trên IF-base ta có đpcm. □

**Định nghĩa 3.11** (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:

$$S_{\{p\} \cup \{q\}} = S_p \cap S_q \tag{3.9}$$

**Mệnh đề 3.3** (So sánh hai IF-subbase). Cho bảng quyết định  $DT = (U, C, D, f)$  và hai IF-subbases  $S_P, S_Q$  tương ứng với  $P, Q \subseteq C$ . Khi đó  $S_Q \prec S_P$  nếu  $P \subseteq Q$ .

*Chứng minh.* Theo định nghĩa 3.8 và phép toán giao trên tập nền IFS ta có đpcm.  $\square$

**Định nghĩa 3.12** (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:

$$B_a[i, j] = \begin{cases} 1_{IF} & \text{if } i = j \\ 0_{IF} & \text{if } other \end{cases}$$

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

**Định nghĩa 3.13** (Khoảng cách 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, độ khác biệt giữa  $S_p$  và  $S_q$  được định nghĩa bởi:

$$\begin{aligned} \zeta(S_p, S_q) &= \frac{1}{|U|^2} \sum_{i=1}^{|U|} (|S_p^L[i] \cup S_q^L[i]| - |S_p^L[i] \cap S_q^L[i]|) \\ &+ \frac{1}{|U|^2} \sum_{i=1}^{|U|} (|S_p^R[i] \cup S_q^R[i]| - |S_p^R[i] \cap S_q^R[i]|) \end{aligned} \quad (3.10)$$

**Mệnh đề 3.4** (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:

$$\zeta(S_p, S_q) = \frac{2}{|U|^2} \sum_{i=1}^{|U|} (|S_p^L[i] \cup S_q^L[i]| - |S_p^L[i] \cap S_q^L[i]|) \quad (3.11)$$

Là độ khác biệt giữa  $S_p$  và  $S_q$

*Chứng minh.* Vì  $S_p^L, S_p^R$ , và  $S_q^L, S_q^R$  là đối xứng lẫn nhau, do đó  $|S_p^L| = |S_p^R|$  và  $|S_q^L| = |S_q^R|$ . Ta có đpcm.  $\square$

**Mệnh đề 3.5** (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:

$$\zeta(S_C, S_{C \cup D}) = \frac{2}{|U|^2} \sum_{i=1}^{|U|} (|S_D^L[i] - S_D^L[i] \cap S_C^L[i]|) \quad (3.12)$$

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

*Chứng minh.* Theo mệnh đề 3.4 ta có:

$$\begin{aligned} \zeta(S_D, S_{D \cup C}) &= \frac{2}{|U|^2} \sum_{i=1}^{|U|} (|S_D^L[i] \cup S_{D \cup C}^L[i]| - |S_D^L[i] \cap S_{D \cup C}^L[i]|) \\ &= \frac{2}{|U|^2} \sum_{i=1}^{|U|} (|S_D^L[i] \cup (S_D^L[i] \cap S_C^L[i])| - |S_D^L[i] \cap (S_D^L[i] \cap S_C^L[i])|) \\ &= \frac{2}{|U|^2} \sum_{i=1}^{|U|} (|S_D^L[i]| - |S_D^L[i] \cap S_C^L[i]|) \end{aligned}$$

Ta có đpcm. □

**Mệnh đề 3.6** (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  $\zeta(S_D, S_{D \cup C}) \leq \zeta(S_D, S_{D \cup B})$ :

*Chứng minh.* Theo mệnh đề 3.3, vì  $B \subseteq C$ ,  $S_C^L \leq S_B^L$ , nghĩa là với mọi  $x \in U$ , nếu  $[x]_C^L \subseteq [x]_B^L$  thì  $|[x]_C^L| \leq |[x]_B^L|$ . Khi đó  $\zeta(S_D, S_{D \cup C}) \leq \zeta(S_D, S_{D \cup B})$ . □

### 3.4. Giảm thuộc tính trong bảng quyết định theo tiếp cận tô pô mờ trực cảm

#### 3.4.1. Đề xuất thuật toán tìm reduct trong bảng quyết định theo phương pháp filter, sử dụng cấu trúc tô pô mờ trực cảm

**Định nghĩa 3.14** (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:

$$Sig_R(a) = \zeta(S_D, S_{D \cup R \cup \{a\}}) - \zeta(S_D, S_{D \cup R}) \quad (3.13)$$

**Mệnh đề 3.7** (Tính tồn tại của reduct). Cho bảng quyết định  $DT = (U, C, D, f)$  và hai IF-bases  $B_R$  và  $B_C$  tương ứng với  $R \subseteq C$ . Khi đó, nếu  $B_R = B_I$  thì  $B_C = B_I$ .

*Chứng minh.* Vì  $R \subseteq C$ ,  $S_C = S_{R \cup \{C-R\}} = S_R \cap S_{\{C-R\}}$ , nghĩa là  $B_C = B_R \cap B_{\{C-R\}}$ . Khi đó  $B_R = B_I \rightarrow B_C = B_I \cap B_{\{C-R\}} = B_I$ .  $\square$

Dựa trên mệnh đề 3.9 ta có thể khẳng định, nếu một bảng quyết định tồn tại một tập con  $R$  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, reduct theo tiếp cận tôpô có thể được định nghĩa như sau:

**Định nghĩa 3.15** (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 reduct 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  $\delta$  được chọn mặc định là 0.5 cho toàn bộ các ví dụ minh họa và thực nghiệm các thuật toán. Sau đây là phần đề xuất thuật toán F\_IFT tìm reduct theo phương pháp filter.

Tiếp theo sẽ là phần đánh giá độ phức tạp của thuật toán. Trước tiên, kí hiệu  $|U|, |C|$  lần lượt là số lượng các đối tượng và số lượng các thuộc tính của bảng quyết định  $DT = (U, C, D, f)$ .

- (1) Độ phức tạp tại các bước 4-6 là  $\mathcal{O}(|C||U|^2)$ ;
- (2) Độ phức tạp tại các bước 8-10 là  $\mathcal{O}(|C-R||U|^2)$ , độ phức tạp tại bước 11 là  $\mathcal{O}(|C-R|)$ , độ phức tạp tại bước 13 là  $\mathcal{O}(|U|^2)$ . Do đó, độ phức tạp tại các bước 7-14 là  $\mathcal{O}(|R||C-R||U|^2)$ ;

Từ (1) và (2) ta có độ phức tạp của thuật toán F\_IFT algorithm là  $\mathcal{O}(|R||C-R||U|^2)$ .

**Ví dụ 3.4.** Để minh họa quá trình hoạt động của thuật toán đề xuất, sau đây là phần trình bày ví dụ số cho thuật toán F\_IFT.

---

**Thuật toán 3.1** Giảm thuộc tính theo phương pháp filter sử dụng tiếp cận tô pô mờ trực cảm (F\_IFT)

---

**Input:** Bảng quyết định  $DT = (U, C, D, f)$  và  $\delta = 0.5$

**Output:** Tập rút gọn  $R$

```

1:  $R \leftarrow \emptyset$ ;
2:  $B_R$  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 3.4}
6: end for
7: while  $B_R \neq B_I$  do
8:   for all  $c \in C - R$  do
9:     calculate  $Sig_R(c)$ ;                             {theo công thức 3.13}
10:  end for
11:  select  $c_m \in C - R : Sig_R(c_m) = \underset{c \in C - R}{Max} \{Sig_R(c)\}$ ;
12:   $R \leftarrow R \cup \{c_m\}$ ;
13:  update  $B_R$ ;                                       {theo công thức 3.7}
14: end while
15: return  $R$ ;

```

---

Cho bảng quyết định  $DT = (U, C, D, f)$  được trình bày như trong Bảng 1.3 trong đó  $U = \{u_1, u_2, u_3, u_4, u_5, u_6\}$  và  $C = \{a, b, c, d, e, f\}$ .

*Giai đoạn khởi tạo:*

$ST \leftarrow \emptyset$ ;  $R_W \leftarrow \emptyset$ ;  $R_F \leftarrow \emptyset$ ;  $R \leftarrow \emptyset$ ;  $B_R$  là cơ sở mờ trực cảm thô nhất;  $B_I$  là cơ sở mờ trực cảm mịn nhất;

*Giai đoạn tính các IF-subbase ban đầu  $S_a, S_b, S_c, S_d, S_e, S_f$ :*

*Giai đoạn filter:*

Vì  $B_R \neq B_I$  do đó thực hiện tính toán độ quan trọng của từng thuộc tính theo IF-subbase như sau:

$$Sig_R(a) = \zeta(S_D, S_{DURU\{a\}}) - \zeta(S_D, S_{DUR}) = 0.23$$

$$Sig_R(b) = \zeta(S_D, S_{DURU\{b\}}) - \zeta(S_D, S_{DUR}) = 0.21$$

$$Sig_R(c) = \zeta(S_D, S_{DURU\{c\}}) - \zeta(S_D, S_{DUR}) = 0.22$$

$$Sig_R(d) = \zeta(S_D, S_{DURU\{d\}}) - \zeta(S_D, S_{DUR}) = 0.19$$

$$Sig_R(e) = \zeta(S_D, S_{DURU\{e\}}) - \zeta(S_D, S_{DUR}) = 0.21$$

$$Sig_R(f) = \zeta(S_D, S_{DURU\{f\}}) - \zeta(S_D, S_{DUR}) = 0.21$$

Vì  $Sig_R(a)$  lớn nhất nên  $R = R \cup \{a\} = \{a\}$  và cập nhật lại  $B_R$ . Vì  $B_R \neq B_I$ , khi đó tiếp tục tính độ quan trọng của các thuộc tính còn lại ta có:

$$Sig_R(b) = \zeta(S_D, S_{DURU\{b\}}) - \zeta(S_D, S_{DUR}) = 0.08$$

$$Sig_R(c) = \zeta(S_D, S_{DURU\{c\}}) - \zeta(S_D, S_{DUR}) = 0.04$$

$$Sig_R(d) = \zeta(S_D, S_{DURU\{d\}}) - \zeta(S_D, S_{DUR}) = 0.02$$

$$Sig_R(e) = \zeta(S_D, S_{DURU\{e\}}) - \zeta(S_D, S_{DUR}) = 0.01$$

$$Sig_R(f) = \zeta(S_D, S_{DURU\{f\}}) - \zeta(S_D, S_{DUR}) = 0.05$$

Vì:  $Sig_R(b)$  lớn nhất nên  $R = R \cup \{a\} = \{a, b\}$ , và cập nhật lại  $B_R$ . Vì  $B_R \neq B_I$  khi đó tiếp tục tính độ quan trọng của các thuộc tính còn lại ta có:

$$Sig_c = \zeta(S_D, S_{DURU\{c\}}) - \zeta(S_D, S_{DUR}) = 0.01$$

$$Sig_d = \zeta(S_D, S_{DURU\{d\}}) - \zeta(S_D, S_{DUR}) = 0.01$$

$$Sig_e = \zeta(S_D, S_{DURU\{e\}}) - \zeta(S_D, S_{DUR}) = 0.01$$

$$Sig_f = \zeta(S_D, S_{DURU\{f\}}) - \zeta(S_D, S_{DUR}) = 0.01$$

Vì  $Sig_R(c) = Sig_R(d) = Sig_R(e) = Sig_R(f)$  do đó ta chọn  $c$ . Khi đó  $R = R \cup \{c\} = \{a, b, c\}$ . Cập nhật lại  $B_R$  ta có  $B_R = B_I$ . Khi đó vòng lặp kết thúc. Ta có  $R = \{a, b, c\}$  là reduct của thuật toán F\_IFT algorithm.

### 3.4.2. Đề xuất thuật toán tìm reduct trong bảng quyết định theo phương pháp lai ghép filter - wrapper, sử dụng cấu trúc tập mờ trực cảm

Về cơ bản, phương pháp chọn lọc thuộc tính được sử dụng trong thuật toán F\_IFT vẫn theo tiếp cận độ đo, do đó khả năng phân loại giữa các thuộc tính vẫn còn thấp. Khi đó, trong quá trình đánh giá, có thể xuất hiện nhiều thuộc tính có cùng độ quan trọng như nhau dẫn tới bỏ sót các thuộc tính ứng viên có thể tốt hơn trong thực tế.

Để giải quyết hạn chế của thuật toán F\_IFT, sau đây là phần đề xuất thuật toán lai ghép filter - wrapper FW\_IFT tìm reduct với cấu trúc dữ liệu **Stack** được sử dụng.

Trong đó các tập thuộc tính ứng viên tại giai đoạn **filter** sẽ được đẩy vào **Stack** để sinh các reduct ứng viên cho giai đoạn **wrapper** của thuật toán.

---

**Thuật toán 3.2** Phương pháp giảm thuộc tính lai ghép filter - wrapper sử dụng tiếp cận tô pô 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 3.4}
6: end for
7: for all  $c \in C - R_F$  do
8:   calculate  $Sig_{R_F}(c)$ ;                        {theo công thức 3.13}
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 reduct 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  $R$ ;

```

---

Trong đó  $ST$  là cấu trúc dữ liệu ngăn xếp Stack với các phép toán PUSH để đẩy dữ liệu vào và POP để đẩy dữ liệu ra khỏi Stack.  $R_F$  là reduct ứng viên tại bước filter của thuật toán,  $R_W$  là danh sách chứa các reduct ứng viên cho giai đoạn wrapper của thuật toán.

Tiếp theo sẽ là phần đánh giá độ phức tạp của thuật toán. Trước tiên, kí hiệu  $|U|, |C|$  lần lượt là số lượng các đối tượng và số lượng các thuộc tính của bảng quyết định  $DT = (U, C, D, f)$ .

(1) Độ phức tạp tại các bước 4-6 là  $\mathcal{O}(|C||U|^2)$ ;

(2) Độ phức tạp tại các bước 7-9 là  $\mathcal{O}(|C - R_F||U|^2)$ ;

(3) Độ phức tạp tại các bước 10-12 là  $\mathcal{O}(|C - R_F|)$ ;

(4) Độ phức tạp tại các bước 13-21 là  $\mathcal{O}(|ST||C - R_F||U|^2)$ ;

(5) From (1), (2), (3), và (4), ta có độ phức tạp của thuật toán FW\_IFT tại giai đoạn filter là  $\mathcal{O}(|ST||C - R_F||U|^2)$ ;

(6) Giả sử độ phức tạp của mô hình phân lớp *Model* là  $\mathcal{O}(|T|)$ . Khi đó, độ phức tạp tại bước 22-26 là  $\mathcal{O}(|R_W||T|)$ ;

Từ (5) và (6), ta có độ phức tạp của thuật toán FW\_IFT là:  $\mathcal{O}(|ST||C - R_F||U|^2) + \mathcal{O}(|R_W||T|)$ .

**Ví dụ 3.5.** Để minh họa quá trình hoạt động của thuật toán đề xuất, sau đây sẽ là phần trình bày ví dụ số cho thuật toán FW\_IFT.

Cho bảng quyết định  $DT = (U, C, D, f)$  được trình bày như trong Bảng 1.3 trong đó  $U = \{u_1, u_2, u_3, u_4, u_5, u_6\}$  và  $C = \{a, b, c, d, e, f\}$ .

*Giai đoạn khởi tạo*

$ST \leftarrow \emptyset; R_W \leftarrow \emptyset; R_F \leftarrow \emptyset; R \leftarrow \emptyset; B_{R_F}$  là cơ sở mờ trực cảm thô nhất;  $B_I$  là cơ sở mờ trực cảm mịn nhất;

*Giai đoạn tính các IF-subbase ban đầu  $S_a, S_b, S_c, S_d, S_e, S_f$ ;*

Tính độ quan trọng của từng thuộc tính trong  $C - R_F$  ta có:

$$Sig_{R_F}(a) = \zeta(S_D, S_{D \cup R_U\{a\}}) - \zeta(S_D, S_{D \cup R}) = 0.23$$

$$Sig_{R_F}(b) = \zeta(S_D, S_{D \cup R_U\{b\}}) - \zeta(S_D, S_{D \cup R}) = 0.21$$

$$Sig_{R_F}(c) = \zeta(S_D, S_{D \cup R_U\{c\}}) - \zeta(S_D, S_{D \cup R}) = 0.22$$

$$Sig_{R_F}(d) = \zeta(S_D, S_{D \cup R_U\{d\}}) - \zeta(S_D, S_{D \cup R}) = 0.19$$

$$Sig_{R_F}(e) = \zeta(S_D, S_{D \cup R_U\{e\}}) - \zeta(S_D, S_{D \cup R}) = 0.21$$

$$Sig_{R_F}(f) = \zeta(S_D, S_{DURU\{f\}}) - \zeta(S_D, S_{DUR}) = 0.21$$

Vì  $Sig_{R_F}(a)$  là lớn nhất nên thực hiện  $ST.PUSH(R_F \cup \{a\})$  do đó  $ST = \{a\}$

*Giai đoạn filter*

Vì  $ST = \{a\} \neq \emptyset$  nên thực hiện  $R_F = ST.POP = \{a\}$ . Khi đó  $ST = \emptyset$  và  $B_{R_F} \neq B_I$  nên tính độ quan trọng của các thuộc tính còn lại trong  $C - R_F$  với  $R_F$  ta có:

$$Sig_{R_F}(b) = \zeta(S_D, S_{DURU\{b\}}) - \zeta(S_D, S_{DUR}) = 0.08$$

$$Sig_{R_F}(c) = \zeta(S_D, S_{DURU\{c\}}) - \zeta(S_D, S_{DUR}) = 0.04$$

$$Sig_{R_F}(d) = \zeta(S_D, S_{DURU\{d\}}) - \zeta(S_D, S_{DUR}) = 0.02$$

$$Sig_{R_F}(e) = \zeta(S_D, S_{DURU\{e\}}) - \zeta(S_D, S_{DUR}) = 0.01$$

$$Sig_{R_F}(f) = \zeta(S_D, S_{DURU\{f\}}) - \zeta(S_D, S_{DUR}) = 0.05$$

Vì  $Sig_{R_F}(b)$  là lớn nhất nên thực hiện  $ST.PUSH(R_F \cup \{a\})$  do đó  $ST = \{a, b\}$

Vì  $ST = \{a, b\} \neq \emptyset$  nên thực hiện  $R_F = ST.POP = \{a\}$ . Khi đó  $ST = \emptyset$  và  $B_{R_F} \neq B_I$  nên tính độ quan trọng của các thuộc tính còn lại trong  $C - R_F$  với  $R_F$  ta có:

$$Sig_c = \zeta(S_D, S_{DURU\{c\}}) - \zeta(S_D, S_{DUR}) = 0.01$$

$$Sig_d = \zeta(S_D, S_{DURU\{d\}}) - \zeta(S_D, S_{DUR}) = 0.01$$

$$Sig_e = \zeta(S_D, S_{DURU\{e\}}) - \zeta(S_D, S_{DUR}) = 0.01$$

$$Sig_f = \zeta(S_D, S_{DURU\{f\}}) - \zeta(S_D, S_{DUR}) = 0.01$$

Vì  $Sig_{R_F}(c) = Sig_{R_F}(d) = Sig_{R_F}(e) = Sig_{R_F}(f)$  nên ta thực hiện các lệnh:

$ST.PUSH(R_F \cup \{c\}); ST.PUSH(R_F \cup \{d\});$

$ST.PUSH(R_F \cup \{e\}); ST.PUSH(R_F \cup \{f\});$

Khi đó  $ST = \{\{a, b, c\}; \{a, b, d\}; \{a, b, e\}; \{a, b, f\}\}$

Vì  $ST \neq \emptyset$  nên thực hiện  $R_F = ST.POP = \{a, b, f\}$ .

Khi đó  $ST = \{\{a, b, c\}; \{a, b, d\}; \{a, b, e\}\}$ .

Vì  $B_{R_F} = B_I$  do đó  $R_W = R_W \cup R_F = \{\{a, b, f\}\}$

Vì  $ST \neq \emptyset$  nên thực hiện  $R_F = ST.POP = \{a, b, e\}$ .

Khi đó  $ST = \{\{a, b, c\}; \{a, b, d\}\}$ .

Vì  $B_{R_F} = B_I$  do đó  $R_W = R_W \cup R_F = \{\{a, b, f\}; \{a, b, e\}\}$

Vì  $ST \neq \emptyset$  nên thực hiện  $R_F = ST.POP = \{a, b, d\}$ .

Khi đó  $ST = \{\{a, b, c\}\}$ .

Vì  $B_{R_F} = B_I$  do đó  $R_W = R_W \cup R_F = \{\{a, b, f\}; \{a, b, e\}; \{a, b, d\}\}$

Vì  $ST \neq \emptyset$  nên thực hiện  $R_F = ST.POP = \{a, b, c\}$  do đó  $ST = \emptyset$ . Vì  $B_R = B_I$  do đó:

$R_W = R_W \cup R_F = \{\{a, b, f\}; \{a, b, e\}; \{a, b, d\}; \{a, b, c\}\}$

Vì  $ST = \emptyset$  nên kết thúc giai đoạn filter, chuyển sang giai đoạn wrapper.

#### *Giai đoạn wrapper*

Thực hiện đánh giá khả năng phân lớp của từng reduct ứng viên  $r \in R_W$ . Xác định ứng viên nào có khả năng phân lớp cao nhất trên mô hình *Model*. Giả sử ứng viên  $r = \{a, b, e\}$  có khả năng phân lớp cao nhất, khi đó reduct thức sự của thuật toán FW\_IFT thu được là  $R = \{a, b, e\}$ .

### **3.4.3. 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 F\_IFT và FW\_IFT trên một số bộ dữ liệu của UCI. Mục tiêu của thực nghiệm nhằm củng cố giả thiết phương pháp giảm thuộc tính theo tiếp cận tôpô cho reduct tối ưu hơn tiếp cận độ đo truyền thống [38]. Trong đó thuật toán F\_IFT sẽ được so sánh với các thuật toán của A. Tan [15], [36] và Thang [113]. Thuật toán FW\_IFT sẽ được so sánh với thuật toán FW\_IFD [113].

#### *3.4.3.1. Kế hoạch thực nghiệm*

Các thuật toán được cài đặt bằng ngôn ngữ lập trình Python và chạy trên nền hệ điều hành Window 10 với cấu hình phần cứng là bộ xử lý Core I5, ram 8GB. Cùng với 12 tập dữ liệu thử nghiệm được tải về từ UCI được mô tả chi tiết trong Bảng 3.1. Trong đó  $|U|$  là số lượng mẫu,  $|C|$  là số thuộc tính điều kiện và  $|D|$  là số phân lớp của



**Bảng 3.1:** Mô tả dữ liệu thực nghiệm

ID	Data	Describe	$ U $	$ C $	$ D $
1	Wine	Wine	178	13	3
2	Heart	Statlog (Heart)	270	13	2
3	Wdbc	Breast Cancer Wisconsin (Diagnostic)	569	30	2
4	Wpbc	Breast Cancer Wisconsin (Prognostic)	198	33	2
5	Iono	Ionosphere	351	34	2
6	UFDC	Ultrasonic flowmeter diagnostics (C)	181	43	4
7	Sona	Connectionist Bench	208	60	2
8	Libras	Libras Movement	360	90	15
9	Musk	Musk	476	166	2
10	LVB	Voice Rehabilitation(Binary)	126	310	2
11	LVG	Voice Rehabilitation(Gender)	126	310	2
12	PD	Parkinson's Disease Classification	756	754	2

thuộc tính quyết định trong mỗi tập dữ liệu (dataset).

Các tập dữ liệu thử nghiệm đều là các dataset có thuộc tính điều kiện miền giá trị số liên tục. Do đó, trước khi thực hiện thuật toán giảm thuộc tính, miền giá trị của các thuộc tính sẽ được chuẩn hóa về đoạn  $[0,1]$ . Độ tương tự và độ không tương tự của quan hệ ưu tiên mờ trực cảm được tính giống phần thực nghiệm của chương 2 luận án.

Để đánh giá khả năng phân lớp chính xác của reduct. Chương này sử dụng hai mô hình phân lớp dữ liệu số là SVM và k-NN( $k=|D|$ ). Độ đo đánh giá và phương pháp đánh giá độ chính xác trên các mô hình là độ đo *Accuracy* và phương pháp đánh giá chéo *10-fold* được sử dụng chung cho toàn bộ các reduct thu được từ các thuật toán.

#### 3.4.3.2. Kịch bản thực nghiệm

Nhằm khẳng định phương pháp giảm thuộc tính đề xuất là hiệu quả hơn về khả năng phân lớp so với một số phương pháp giảm thuộc tính khác trên nền tập mờ trực cảm, chương này tiến hành thực nghiệm các thuật toán đề xuất theo các kịch bản như sau:

1) So sánh reduct của thuật toán đề xuất F\_IFT với các thuật toán filter theo tiếp cận khoảng cách mờ trực cảm F\_IFD [113], thuật toán filter theo tiếp cận miền dương

**Hình 3.1:** Tập rút gọn thu được từ thuật toán  $F\_IFT$ 

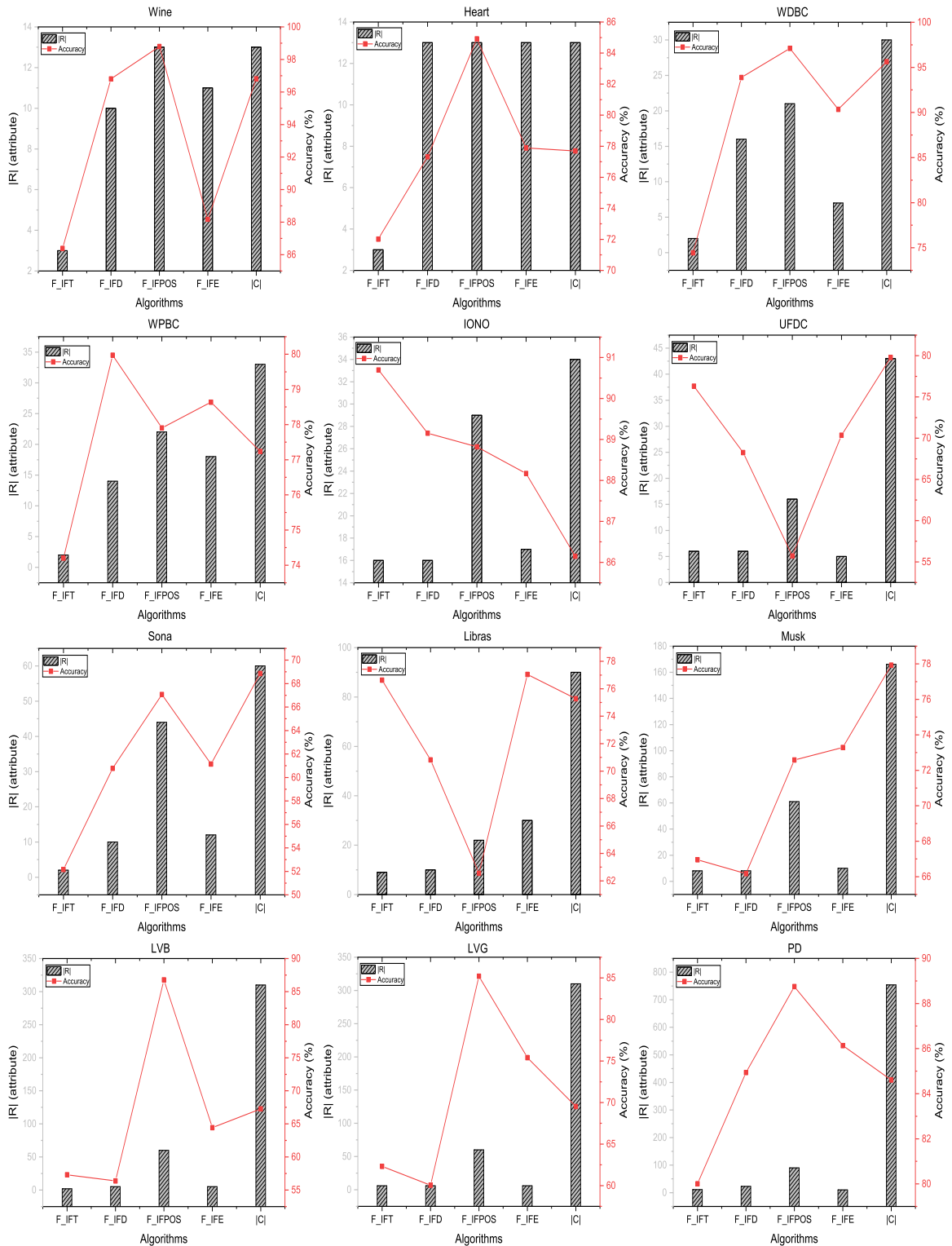
ID	Datasets	Tập rút gọn
1	Wine	[7, 11, 9]
2	Heart	[0, 7, 9]
3	Wdbc	[21, 4]
4	Wpbc	[0, 1]
5	Iono	[18, 9, 3, 25, 30, 6, 7, 5, 2, 19, 23, 32, 12, 4, 27, 8]
6	UFDC	[3, 26, 10, 42, 6, 29]
7	Sona	[19, 30]
8	Libras	[66, 1, 71, 0, 88, 39, 89, 51, 58]
9	Musk	[31, 62, 162, 164, 89, 25, 115, 106]
10	LVB	[58, 51]
11	LVG	[79, 82, 56, 84, 47, 49]
12	PD	[0, 420, 421, 422, 423, 424, 426, 427, 430, 437, 502, 613]

mờ trực cảm  $F\_IFPOS$  [36] và thuật toán filter theo tiếp cận Entropy thông tin mờ trực cảm  $F\_IFE$  [15]. Trong đó các tiêu chí được sử dụng để so sánh và đánh giá bao gồm khả năng phân lớp (accuracy), số lượng phần tử của reduct ( $|R|$ ) và chi phí tính toán của thuật toán (second).

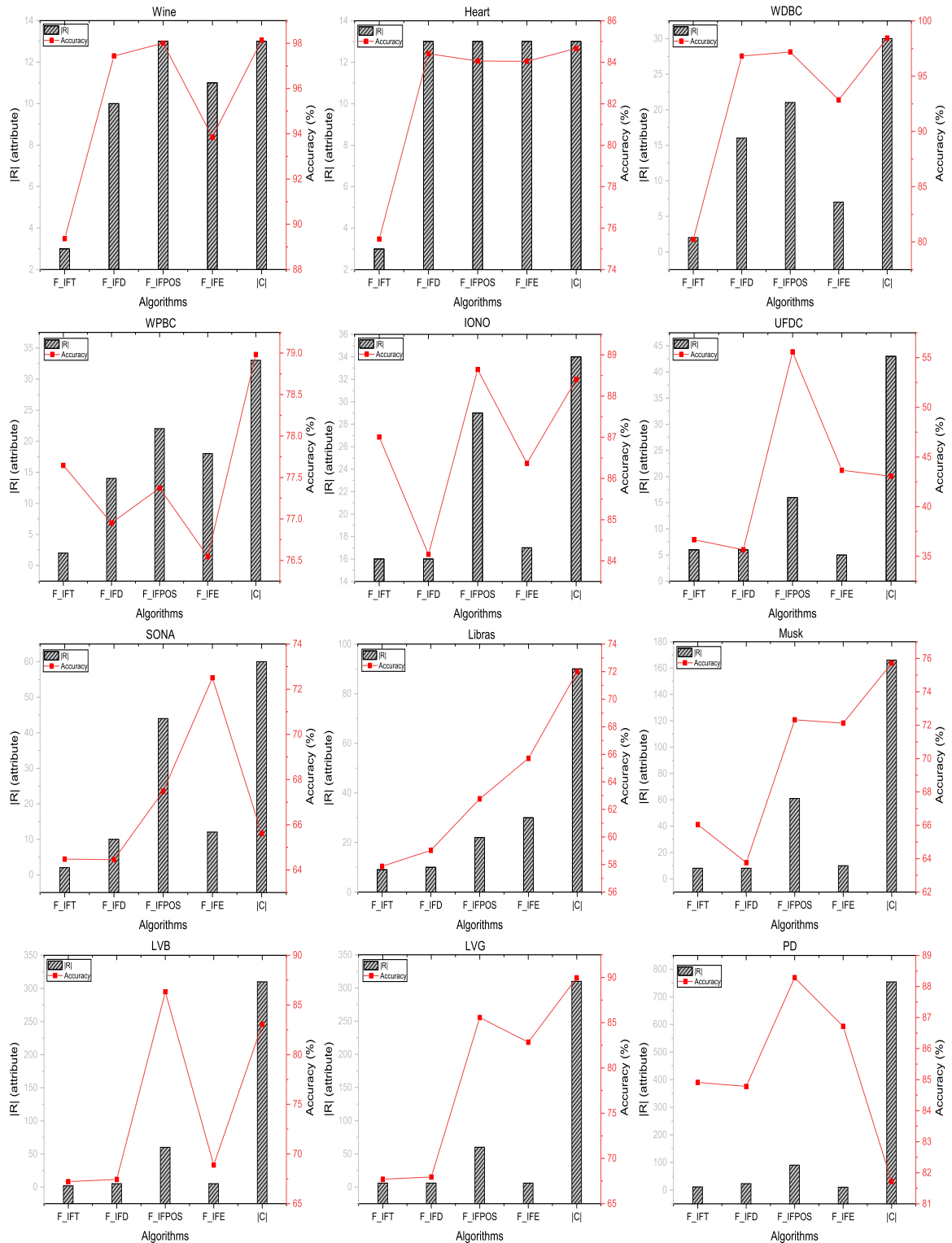
2) So sánh reduct của thuật toán đề xuất  $F\_IFT$  với thuật toán filter - wrapper theo tiếp cận khoảng cách mờ trực cảm  $FW\_IFD$  [113]. Trong đó các tiêu chí được sử dụng để so sánh và đánh giá bao gồm khả năng phân lớp (accuracy), số lượng phần tử của reduct ( $|R|$ ) và chi phí tính toán của thuật toán (second).

#### 3.4.3.3. Đánh giá thuật toán $F\_IFT$

Bảng 3.1 mô tả các reduct thu được từ thuật toán  $F\_IFT$  trên từng tập dữ liệu. Trong đó tên của các thuộc tính được đánh số lần lượt từ 0 đến  $|C - 1|$ . Bảng 3.2 so sánh số lượng phần tử reduct thu được từ các thuật toán. Kết quả thực nghiệm được trình bày trong Bảng 3.2 cho thấy số lượng phần tử trung bình của các reduct thu được từ thuật toán đề xuất  $F\_IFT$  thấp hơn đáng kể so với các thuật toán khác. Quan sát biểu đồ phân tích sự tương quan về số lượng phần tử reduct (trái) trong Hình 3.4, ta có thể



**Hình 3.2:** Sự tương quan về số lượng phân tử và khả năng phân lớp chính xác của reduct thu được từ các thuật toán so với  $F\_IFT$  trên mô hình phân lớp KNN.



**Hình 3.3:** Sự tương quan về số lượng phân tử và khả năng phân lớp chính xác của reduct thu được từ các thuật toán so với  $F\_IFT$  trên mô hình phân lớp SVM.

**Bảng 3.2:** So sánh số lượng phần tử của các reduct thu được từ các thuật toán theo tiếp cận filter

ID	F_IFT	F_IFD	F_IFPOS	F_IFE	C
1	3	10	13	11	13
2	3	13	13	13	13
3	2	16	21	7	30
4	2	14	22	18	33
5	16	16	29	17	34
6	6	6	16	5	43
7	2	10	44	12	60
8	9	10	22	30	90
9	8	8	61	10	166
10	2	5	60	5	310
11	6	6	60	6	310
12	5	23	90	10	754

**Bảng 3.3:** So sánh khả năng phân lớp chính xác của reduct thu được từ các thuật toán theo tiếp cận filter trên mô hình phân lớp KNN

ID	F_IFT	F_IFD	F_IFPOS	F_IFE	C
1	86.62	96.93	98.96	88.12	96.24
2	72.84	77.57	84.85	77.71	77.7
3	74.91	93.66	97.72	90.48	95.13
4	74.82	79.21	77.61	79	77.81
5	90.89	89.72	88.28	88.63	86.26
6	76.71	68.9	55.65	70.61	79.2
7	52.83	60.67	67.45	61.93	68.95
8	76.62	70.79	62.46	77.88	75.21
9	66.14	66.17	72.25	73.16	77.61
10	57.55	56.31	86.76	64.92	67.85
11	62.07	60.91	85.83	75.76	69.54
12	80.11	84.91	88.28	86.6	84.05

thấy số thuộc tính tăng lên nhưng số lượng phần tử reduct không tăng. Do đó, chúng ta có thể khẳng định phương pháp giảm thuộc tính theo tiếp cận tập ô giảm chiều tốt hơn so với các tiếp cận đo đo trên nền tập mờ trực cảm.

Hơn nữa, chi phí tính toán của thuật toán giảm thuộc tính đề xuất F\_IFT cũng được cải thiện đáng kể so với các thuật toán khác. Quan sát biểu đồ phân tích sự tương quan

**Bảng 3.4:** So sánh khả năng phân lớp chính xác của reduct thu được từ các thuật toán theo tiếp cận filter trên mô hình phân lớp SVM

ID	F_IFT	F_IFD	F_IFPOS	F_IFE	C
1	89.37	97.45	98.14	93.55	98.64
2	75.33	84.35	84.81	84.36	84.91
3	80.79	96.96	97.62	92.34	98.99
4	77.71	76.3	77.51	76.13	78.9
5	87.09	84.46	88.17	86.22	88.07
6	36.39	35.13	55.7	43.94	43.92
7	64.23	64.34	67.18	72.71	65.3
8	57.46	59.31	62.53	65.92	71.68
9	66.48	63.76	72.22	72.71	75.49
10	67.03	67.55	86.45	68.86	83.06
11	67.83	67.98	85.3	82.37	89.92
12	84.83	84.75	88.38	86.39	81.37

về chi phí tính toán (phải) trong Hình 3.4, ta có thể thấy chi phí tính toán của thuật toán hầu như không biến động trên các bộ dữ liệu có số lượng thuộc tính nhỏ hơn 500, và biến động thấp nhất trên các bộ dữ liệu có số lượng thuộc tính lớn hơn 500.

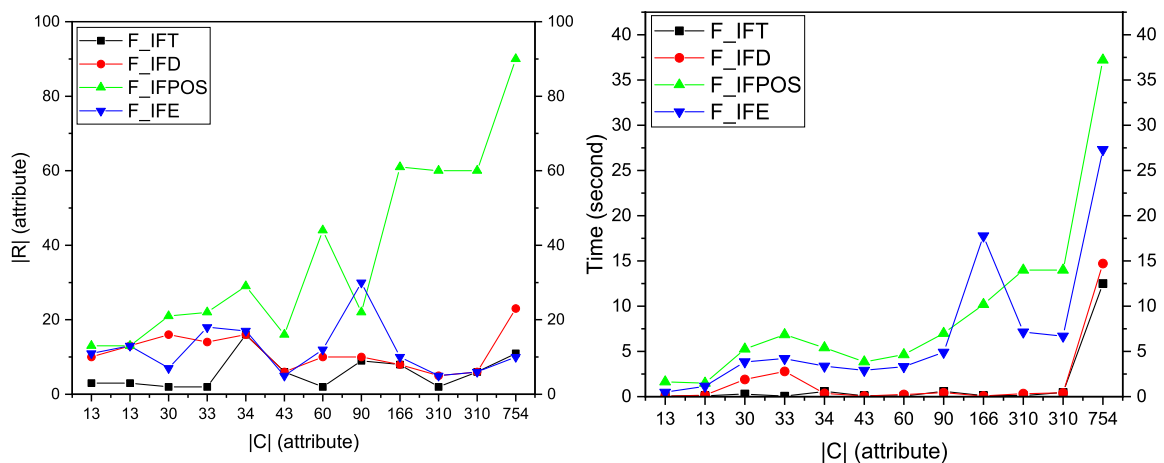
Tuy nhiên, khả năng phân lớp chính xác của reduct thu được từ thuật toán đề xuất vẫn còn hạn chế. Quan sát kết quả thống kê về khả năng phân lớp các reduct thu được từ thuật toán đề xuất trên mô hình phân lớp KNN trong Bảng 3.3, trên mô hình phân lớp SVM trong Bảng 3.4. Ta có thể thấy độ chính xác trung bình trên toàn bộ các tập dữ liệu của hai mô hình đều thấp hơn so với các thuật toán khác. Sau đây là các phân tích về nguyên nhân ảnh hưởng tới chi phí tính toán của thuật toán F\_IFT, số lượng phân tử và khả năng phân lớp chính xác của reduct thu được bởi thuật toán F\_IFT.

- *Kích thước của reduct*: Các kết quả phân tích và thống kê trong phần thực nghiệm cho thấy thuật toán đề xuất F\_IFT hiệu quả về giảm thuộc tính. Nguyên nhân chính ảnh hưởng tốt tới khả năng giảm thuộc tính đó là phương pháp định nghĩa reduct theo tiếp cận tôpô như đã được đề xuất trong phần nghiên cứu lý thuyết của Chương 3. Kết quả này là hoàn toàn phù hợp với phương pháp đánh giá độ tương đồng về mặt cấu trúc trong tôpô như đã được khẳng định trong các kết quả nghiên cứu của Yu và các

cộng sự trong [38].

- *Chi phí tính toán của thuật toán:* Cũng dựa trên phương pháp định nghĩa reduct theo tiếp cận tôpô. Thay vì phải đối sánh độ tương đồng giữa tôpô của tập thuộc rút gọn thông qua  $B_R$  với tôpô của tập thuộc tính ban đầu thông qua  $B_C$ , chương này sử dụng khái niệm tôpô đơn vị thông qua  $B_I$  để làm điều kiện dừng cho thuật toán. Khi đó, chúng ta không phải tính toán  $B_C$  nên giảm đáng kể chi phí tính toán của thuật toán. Hơn nữa khi reduct có số lượng phần tử càng nhỏ thì thời gian hội tụ của thuật toán càng nhanh. Đây là những nguyên nhân quan trọng cải thiện tốt chi phí tính toán của thuật toán F\_IFT.

- *Khả năng phân lớp:* Trái lại với số lượng phần tử của reduct được cải thiện đáng kể thì khả năng phân lớp còn gặp nhiều hạn chế. Nguyên nhân chính đó là phương pháp đánh giá reduct đề xuất trong nghiên cứu của Chương 3 vẫn dựa trên tiếp cận độ đo độ tương tự. Đây cũng chính là nhược điểm của hầu hết các phương pháp giảm thuộc tính hiện nay như đã được trình bày trong phần Mở đầu của luận án.



**Hình 3.4:** Biểu đồ đánh giá sự tương quan về số lượng phần tử reduct (trái) và chi phí tính toán (phải) với số lượng thuộc tính ban đầu của thuật toán F\_IFT so với các thuật toán khác

#### 3.4.3.4. Đánh giá thuật toán FW\_IFT

Các Bảng 3.5 và Bảng 3.6 trình bày reduct thu được từ thuật toán đề xuất FW\_IFT trên từng tập dữ liệu tương ứng trên hai mô hình phân lớp SVM và KNN. Trong các bảng này, tên của các thuộc tính cũng được đánh số thứ tự từ 0 đến  $|C| - 1$  để tiện cho quá trình thống kê và quan sát các thuộc tính thu được sau rút gọn.

Bảng 3.7 so sánh số lượng phần tử reduct thu được từ các thuật toán. Cả hai thuật toán so sánh đều sử dụng phương pháp lai ghép filter - wrapper trên hai mô hình phân lớp dữ liệu SVM và KNN.

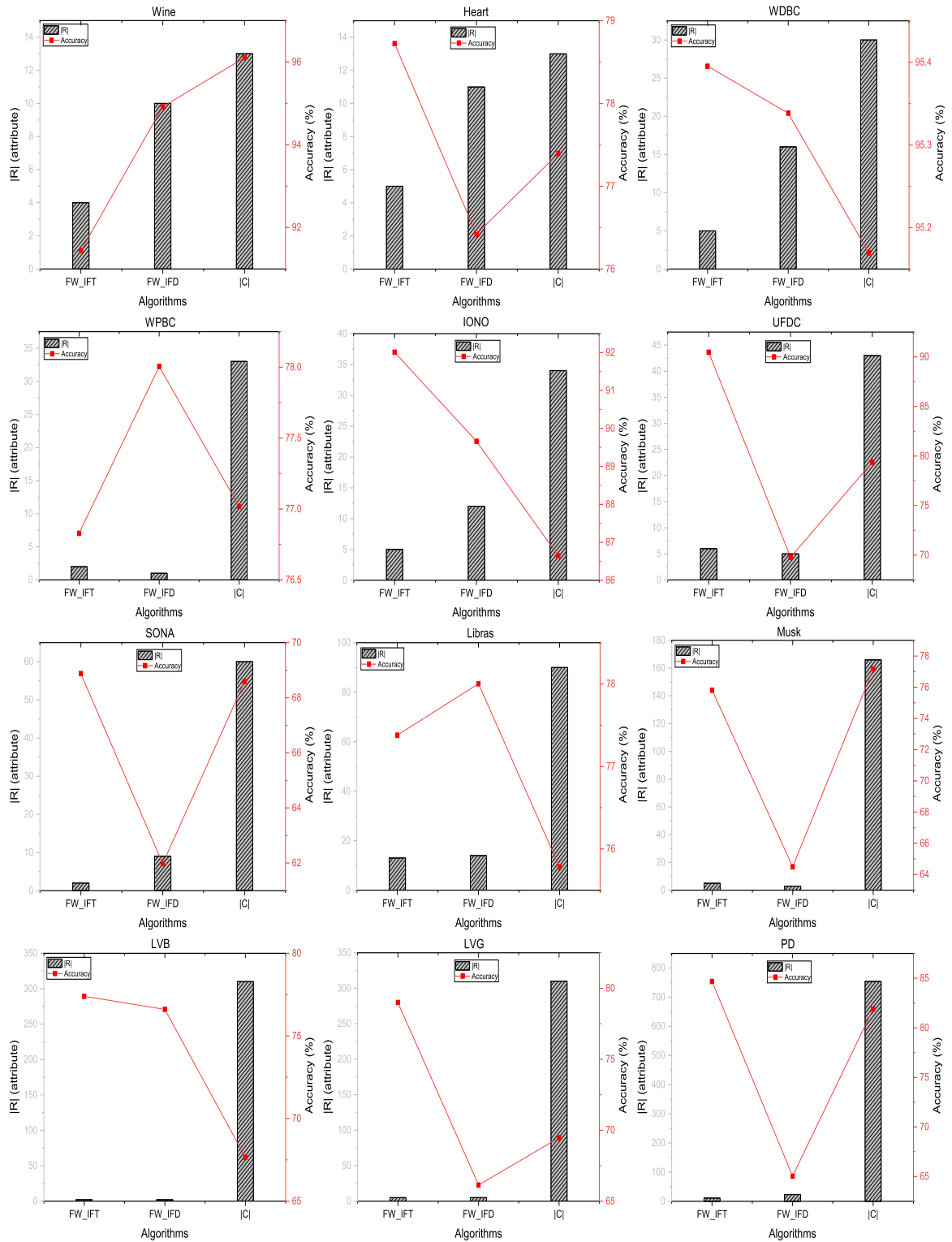
**Bảng 3.5:** Tập rút gọn thu được từ thuật toán FW\_IFT trên mô hình phân lớp SVM

ID	Data	Reduct sets
1	Wine	[7, 9, 10, 11, 12]
2	Heart	[0, 8, 9, 11, 12]
3	Wdbc	[21, 22, 24]
4	Wpbc	[0, 1]
5	Iono	[18, 0, 2, 3, 4, 5, 10]
6	UFDC	[3, 17, 19, 20, 21, 22, 23, 25]
7	Sona	[19, 10, 11]
8	Libras	[66, 10, 11, 14, 15, 16, 19, 20, 23, 26, 29, 34, 35, 41, 44, 51, 61, 72]
9	Musk	[31, 161, 162, 163, 164]
10	LVB	[58, 81, 82, 83, 84, 88]
11	LVG	[79, 88, 89, 90, 91, 92, 93]
12	PD	[0, 420, 421, 422, 427, 430, 437, 502, 613]

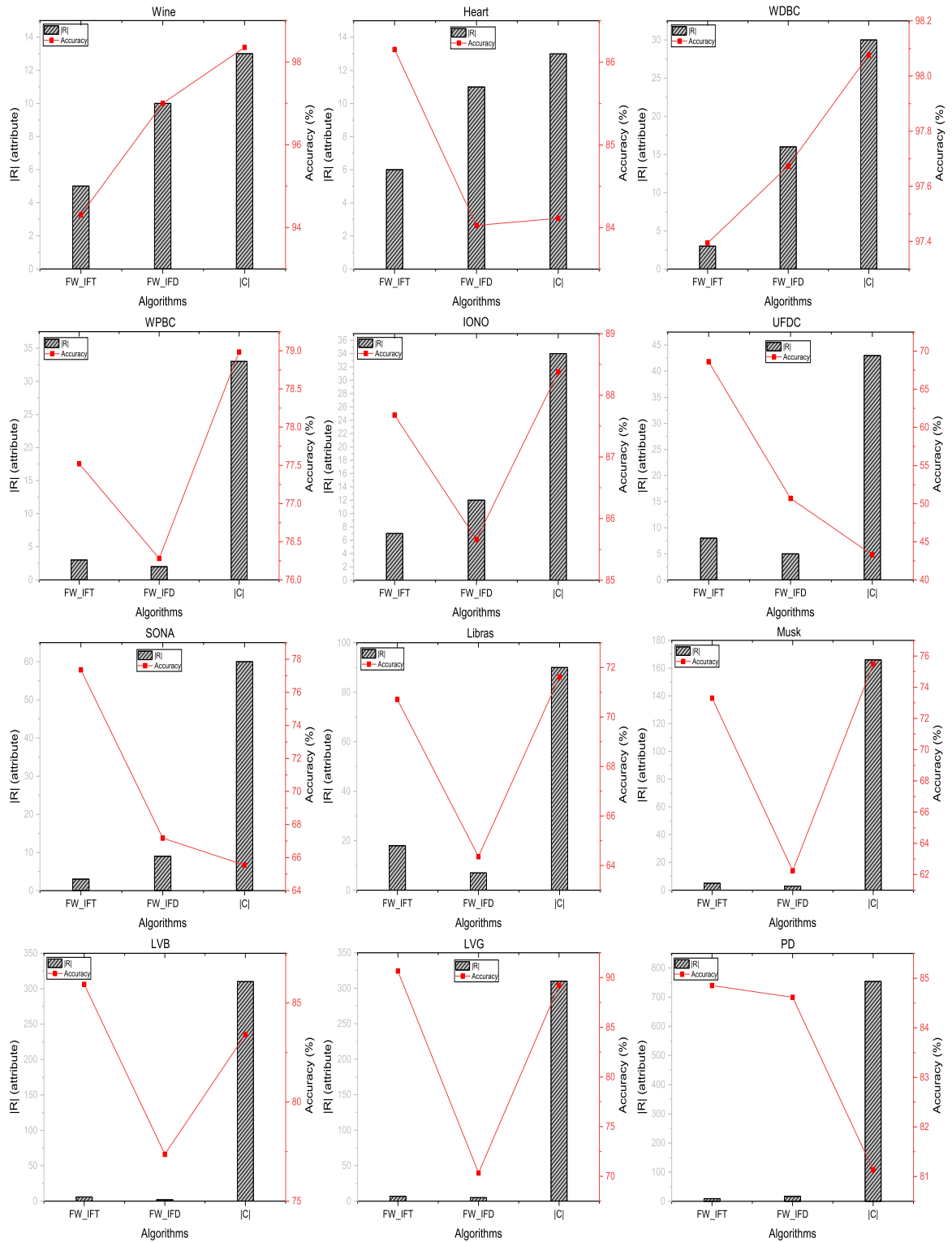
Kết quả thực nghiệm được trình bày trong Bảng 3.7 cho thấy số lượng phần tử trung bình của các reduct thu được từ thuật toán đề xuất FW\_IFT cải thiện tốt hơn so với thuật toán theo tiếp cận độ đo khoảng cách mờ. Quan sát biểu đồ phân tích sự tương quan về số lượng phần tử reduct (trái) trong Hình 3.7 và Hình 3.8, ta có thể thấy số thuộc tính tăng lên nhưng số lượng phần tử reduct không tăng đáng kể, hơn nữa tính ổn định về số lượng phần tử đạt được trên hầu hết các bộ dữ liệu.

Hơn nữa, khả năng phân lớp của các reduct thu được từ thuật toán đề xuất được cải thiện đáng kể, tốt hơn so với thuật toán theo tiếp cận độ đo khoảng cách mờ trực cảm.





**Hình 3.5:** Sự tương quan về số lượng phần tử và khả năng phân lớp chính xác của reduct thu được từ các thuật toán so với FW\_IFT trên mô hình phân lớp KNN.



**Hình 3.6:** Sự tương quan về số lượng phần tử và khả năng phân lớp chính xác của reduct thu được từ các thuật toán so với FW\_IFT trên mô hình phân lớp SVM.

**Bảng 3.6:** Tập rút gọn thu được từ thuật toán FW\_IFT trên mô hình phân lớp KNN

ID	Data	Reduct sets
1	Wine	[7, 10, 11, 12]
2	Heart	[0, 8, 9, 11, 12]
3	Wdbc	[19, 20, 21, 24, 28]
4	Wpbc	[0, 1]
5	Iono	[18, 3, 4, 5, 9]
6	UFDC	[3, 29, 31, 33, 35, 36]
7	Sona	[19, 11]
8	Libras	[66, 4, 12, 13, 16, 19, 22, 25, 33, 36, 43, 54, 59]
9	Musk	[31, 109, 110, 111, 114]
10	LVB	[58, 86]
11	LVG	[79, 90, 91, 92, 93]
12	PD	[0, 420, 421, 422, 427, 430, 437, 502, 503, 504, 613]

Các kết quả phân tích và thống kê về khả năng phân lớp chính xác của reduct thu từ thuật toán FW\_IFT trên cả hai mô hình phân lớp KNN và SVM trong Bảng 3.8 cho thấy khả năng phân lớp trung bình trên cả hai mô hình phân lớp đều tốt hơn đáng kể so với phương pháp dựa trên khoảng cách mờ trực cảm. Đặc biệt, một số bộ dữ liệu có khả năng phân lớp ban đầu thấp như Sona và UFDC đã được cải thiện đáng kể so với thuật toán được so sánh.

Tuy nhiên chi phí tính toán của thuật toán đề xuất FW\_IFT còn gặp nhiều hạn chế. Đây là sự đánh đổi chi phí tính toán để cải thiện khả năng phân lớp cho thuật toán F\_IFT. Quan sát biểu đồ phân tích sự tương quan về chi phí tính toán của các thuật toán (phải) trong Hình 3.7 và Hình 3.8, ta có thể thấy số thuộc tính tăng lên thì chi phí tính toán của thuật toán đề xuất cũng tăng lên đáng kể trên hầu hết các bộ dữ liệu. Sau đây là các phân tích về nguyên nhân ảnh hưởng tới chi phí tính toán của thuật toán FW\_IFT, số lượng phần tử và khả năng phân lớp chính xác của reduct thu được bởi thuật toán FW\_IFT.

- *Kích thước của reduct*: Các kết quả phân tích được trình bày trong phần thực nghiệm của Chương 3 trong luận án cho thấy số lượng phần tử các reduct thu được từ thuật toán FW\_IFT gần như tương đương với thuật toán F\_IFT. Theo cùng tiếp cận

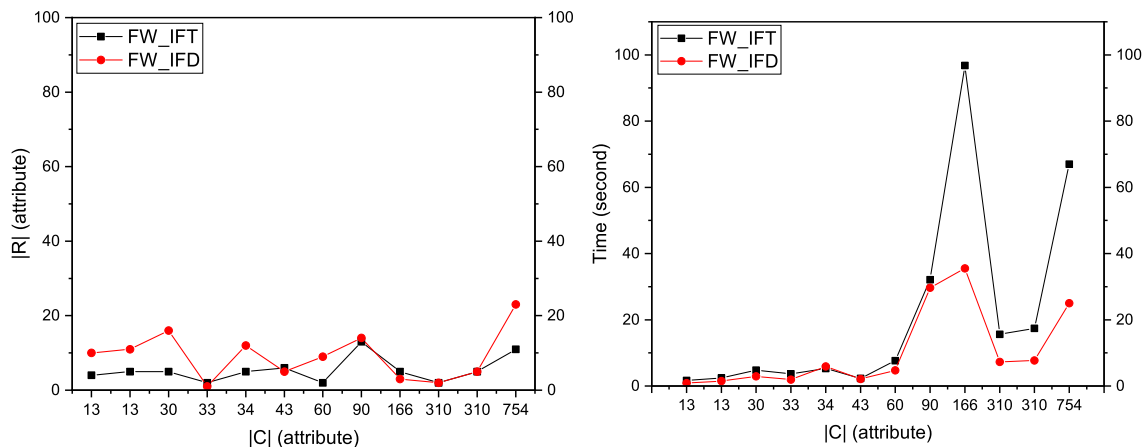
**Bảng 3.7:** So sánh số lượng phần tử của các reduct thu được từ các thuật toán theo tiếp cận filter - wrapper trên mô hình phân lớp SVM và KNN

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

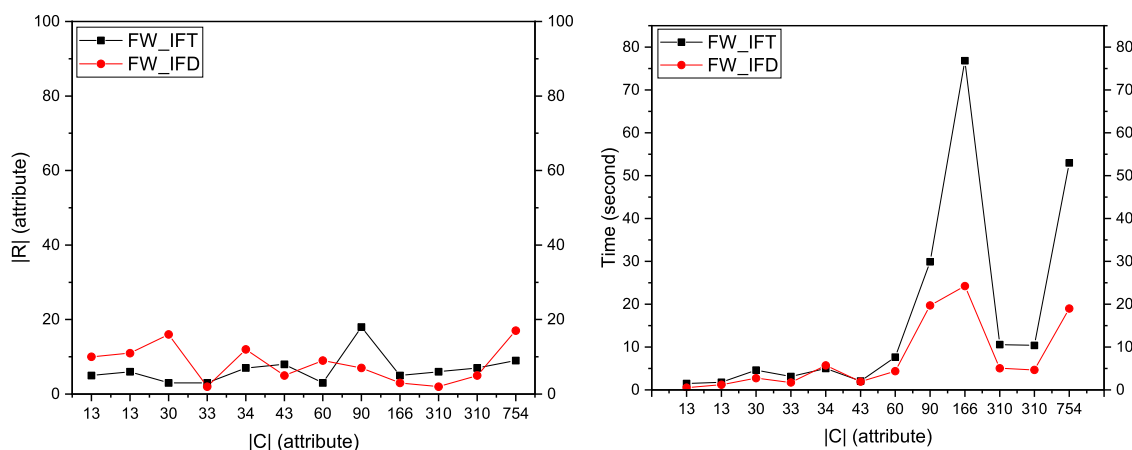
**Bảng 3.8:** So sánh khả năng phân lớp của các reduct thu được từ các thuật toán theo tiếp cận filter - wrapper trên mô hình phân lớp SVM và KNN

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

lai ghép filter - wrapper, thuật toán đề xuất theo tiếp cận tập mờ trực cảm cho reduct có số lượng phần tử tốt hơn thuật toán theo tiếp cận khoảng cách mờ trực cảm. Yếu tố ảnh hưởng chính vẫn là phương pháp định nghĩa reduct theo tiếp cận tập mờ như đã được trình bày.



**Hình 3.7:** Biểu đồ đánh giá sự tương quan về số lượng phần tử reduct (trái) và chi phí tính toán (phải) với số lượng thuộc tính ban đầu của thuật toán FW\_IFT so với các thuật toán khác trên mô hình phân lớp KNN



**Hình 3.8:** Biểu đồ đánh giá sự tương quan về số lượng phần tử reduct (trái) và chi phí tính toán (phải) với số lượng thuộc tính ban đầu của thuật toán FW\_IFT so với các thuật toán khác trên mô hình phân lớp SVM

- *Khả năng phân lớp chính xác của reduct:* Yếu tố ảnh hưởng đến khả năng cải thiện khả năng phân lớp của thuật toán đó chính là phương pháp filter - wrapper thông qua cấu trúc dữ liệu Stack. Nhược điểm của tiếp cận độ đo là bỏ sót nhiều thuộc tính ứng viên có cùng độ quan trọng, do đó cấu trúc Stack sẽ lưu vết lại các reduct ứng viên này để xây dựng các reduct ứng viên. phục vụ cho bước wrapper của thuật toán.

- *Chi phí tính toán của thuật toán:* Trái lại với khả năng nâng cao độ chính xác của reduct thu được từ thuật toán thì chi phí tính toán còn nhiều hạn chế. Nguyên nhân

chính cũng là phương pháp xây dựng các reduct ứng viên thông qua cấu trúc dữ liệu Stack. Theo tiếp cận này ta sẽ thu được nhiều reduct ứng viên để xét nhưng phải trả giá về mặt thời gian xác định reduct cuối cùng của thuật toán.

### 3.5. Kết luận Chương 3

Chương 3, luận án trình bày về phương pháp giảm thuộc tính theo tiếp cận tô pô mờ trực cảm. Các đóng góp chính của Chương này gồm có:

- Đề xuất cấu trúc tô pô mờ trực cảm dựa trên quan hệ ưu tiên mờ trực cảm. Nghiên cứu các tính chất của IF-base và IF-subbase và các phép toán cơ bản nhằm xây dựng độ đo đánh giá sự tương đồng giữa hai tô pô mờ trực cảm.

- Đề xuất hai thuật toán theo phương pháp filter và filter - wrapper tìm reduct trong bảng quyết định số với định nghĩa mới về reduct theo tiếp cận tô pô đơn vị.

Các kết quả thực nghiệm cho thấy thuật toán đề xuất theo phương pháp filter cho chi phí tính toán hiệu quả. Trong khi đó thuật toán đề xuất theo phương pháp lai ghép filter - wrapper cho các reduct hiệu quả về số lượng phần tử và khả năng phân lớp trên hầu hết các tập dữ liệu, đặc biệt là mục tiêu nâng cao chất lượng phân lớp và giảm số lượng phần tử của reduct cho các bộ dữ liệu có khả năng phân lớp ban đầu thấp.

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

### 4.1. Mở đầu

Năm 2005, Lashin và các cộng sự lần đầu tiên giới thiệu khái niệm tôpô rút gọn theo tiếp cận rough set [37]. Từ đó, phương pháp xây dựng cấu trúc tôpô theo tiếp cận RS được nhiều nhà nghiên cứu quan tâm và đề xuất. Hiện nay có hai phương pháp xây dựng tôpô theo tiếp cận RS gồm có, các phương pháp xây dựng tôpô từ không gian xấp xỉ của RS [38], [40]–[42], các phương pháp xây dựng tôpô từ các phép toán xấp xỉ của RS [43].

Chương 3 của luận án đã trình bày phương pháp giảm thuộc tính theo tiếp cận tôpô mờ trực cảm, trong đó cấu trúc tôpô mờ trực cảm được xác định qua quan hệ ưu tiên mờ trực cảm. Tuy nhiên, phương pháp đánh giá độ quan trọng của thuộc tính trong Chương 3 vẫn sử dụng tiếp cận độ đo, chưa thể hiện được tinh thần chọn lọc thuộc tính theo cấu trúc tôpô. Hơn nữa, sự đánh đổi về mặt chi phí tính toán của thuật toán với số lượng phần tử và khả năng phân lớp chính xác của reduct là quá lớn.

Dựa trên tính chất khả li của tôpô Hausdorff, Chương 4 luận án đề xuất phương pháp giảm thuộc tính theo tiếp cận tôpô Hausdorff nhằm thay đổi hướng tiếp cận chọn lọc thuộc tính theo tiếp cận độ đo sang hướng tiếp cận chọn lọc thuộc tính theo cấu trúc tôpô. Một số đóng góp mới của chương nghiên cứu này gồm có:

- Đề xuất phương pháp giảm thuộc tính theo tiếp cận tôpô Hausdorff. Trong đó, phương pháp xây dựng cấu trúc tôpô theo tiếp cận RS trên nền không gian xấp xỉ mờ ngưỡng  $\beta$ .
- Đề xuất phương pháp xác định cấu trúc tôpô Hausdorff dựa trên định nghĩa về

tính phân biệt được của ma trận quan hệ mờ trực cảm ngưỡng  $\beta$ .

- Đề xuất phương pháp gom cụm các thuộc tính dựa trên định nghĩa về sự tương đồng cấu trúc phụ thuộc của tôpô Hausdorff. Trên cơ sở đó, luận án đề xuất thuật toán giảm thuộc tính mới hiệu quả về thời gian trong khi nâng cao chất lượng về số lượng phần tử và khả năng phân lớp cho reduct thu được.

Các kết quả thực nghiệm trên các bộ dữ liệu tải về từ UCI cho thấy, thuật toán đề xuất là hoàn toàn vượt trội so với các thuật toán khác. Các kết quả nghiên cứu trong Chương này được công bố trên các công trình nghiên cứu [CT1]. [CT5] đang chờ phản biện vòng 1.

## 4.2. Đề xuất cấu trúc tôpô từ không gian xấp xỉ mờ ngưỡng $\beta$

**Định nghĩa 4.1** (Không gian xấp xỉ mờ ngưỡng  $\beta$ ). **Không gian xấp xỉ mờ ngưỡng  $\beta$  kí hiệu bởi  $(U, R^\beta)$ . Trong đó  $R^\beta = \{R(p, q) \mid \beta \leq R(p, q), \beta \in [0, 1], \forall p, q \in U\}$ .**

**Định nghĩa 4.2** (Công thức quan hệ mờ ngưỡng  $\beta$ ). **Quan hệ tương đương mờ ngưỡng  $\beta$  của  $p, q \in U$  được định nghĩa như sau:**

$$R^\beta(p, q) = \begin{cases} 1 - |p - q| : \text{if } 1 - |p - q| \geq \beta \\ 0 : \text{if } 1 - |p - q| < \beta. \end{cases} \quad (4.1)$$

**Định nghĩa 4.3** (Thứ tự bộ phận của hai quan hệ mờ ngưỡng  $\beta$ ). Cho hai quan hệ tương đương mờ  $R_1^\beta$  and  $R_2^\beta$  xác định trên  $U$ . Khi đó  $R_1^\beta$  được gọi là nhỏ hơn ( $\prec$ )  $R_2^\beta$  nếu mọi  $p, q \in U$  ta có  $R_1^\beta(p, q) \leq R_2^\beta(p, q)$ .

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

*Chứng minh.* Ta cần chứng minh 3 điều kiện ràng buộc của cấu trúc tôpô như sau:

(1): Dựa trên tính chất 2) của mệnh đề 1.1, ta có  $\underline{R^\beta}(U) = U$  và  $\overline{R^\beta}(\emptyset) = \emptyset$ . Khi đó  $\emptyset \in \mathcal{T}$  và  $U \in \mathcal{T}$ ;



(2): Giả sử  $X, Y \in \mathcal{T}$ , khi đó  $\underline{R}^\beta(X) = \overline{R}^\beta(X)$ , và  $\underline{R}^\beta(Y) = \overline{R}^\beta(Y)$ . Hơn nữa  $\underline{R}^\beta(X \cap Y) \subseteq \overline{R}^\beta(X \cap Y)$  và  $\overline{R}^\beta(X \cap Y) \subseteq \overline{R}^\beta(X) \cap \overline{R}^\beta(Y) = \underline{R}^\beta(X) \cap \underline{R}^\beta(Y) = \underline{R}^\beta(X \cap Y)$ . Khi đó  $\overline{R}^\beta(X \cap Y) = \underline{R}^\beta(X \cap Y)$ . Do đó  $X \cap Y \in \mathcal{T}$ ;

(3): Giả sử  $X_k \in \mathcal{T} | k \in K$ . Khi đó  $\underline{R}^\beta(X_k) = \overline{R}^\beta(X_k)$ , do đó với mọi  $k \in K$  ta có  $\overline{R}^\beta(\bigcup_{k \in K} X_k) = \bigcup_{k \in K} \overline{R}^\beta(X_k) = \bigcup_{k \in K} \underline{R}^\beta(X_k) \subseteq \underline{R}^\beta(\bigcup_{k \in K} X_k)$ . Hơn nữa  $\underline{R}^\beta(\bigcup_{k \in K} X_k) \subseteq \overline{R}^\beta(\bigcup_{k \in K} X_k)$  do đó  $\underline{R}^\beta(\bigcup_{k \in K} X_k) = \overline{R}^\beta(\bigcup_{k \in K} X_k)$ .

Từ (1), (2) và (3) ta có thể kết luận  $\mathcal{T}$  là một tôpô trên  $U$ .  $\square$

**Ví dụ 4.1.** Xét bảng quyết định như trong Bảng 1.3, sử dụng công thức quan hệ 4.1 với  $\beta = 0.5$  ta có:

$$R_a^\beta = \begin{bmatrix} 1 & 1 & 0.8 & 0 & 0 & 0 \\ 1 & 1 & 0.8 & 0 & 0 & 0 \\ 0.8 & 0.8 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 1 & 1 & 1 \end{bmatrix}, R_c^\beta = \begin{bmatrix} 1 & 0 & 0.8 & 1 & 1 & 0 \\ 0 & 1 & 0 & 0 & 0 & 1 \\ 0.8 & 0 & 1 & 0.8 & 0.8 & 0 \\ 1 & 0 & 0.8 & 1 & 1 & 0 \\ 1 & 0 & 0.8 & 1 & 1 & 0 \\ 0 & 1 & 0 & 0 & 0 & 1 \end{bmatrix}$$

Khi đó:

$$\mathcal{T}_a = \{\emptyset, \{u_1, u_2, u_3\}, \{u_4, u_5, u_6\}, U\}; \mathcal{T}_c = \{\emptyset, \{u_1, u_3, u_4, u_5\}, \{u_2, u_6\}, U\}$$

**Định nghĩa 4.4** (Quan hệ bao thuộc của hai tôpô). Cho  $\mathcal{T}_p = \{X \subseteq U | \underline{R}_p^\beta(X) = \overline{R}_p^\beta(X)\}$

và  $\mathcal{T}_q = \{X \subseteq U | \underline{R}_q^\beta(X) = \overline{R}_q^\beta(X)\}$  là hai tôpô xác định trên  $U$  tương ứng với  $p, q \subseteq C$ . Khi đó  $\mathcal{T}_p \subseteq \mathcal{T}_q$  nếu với mọi  $e \in \mathcal{T}_p$  thì  $e \in \mathcal{T}_q$ .

**Ví dụ 4.2.** Cho hai quan hệ tương đương mờ như sau:

$$R_p^\beta = \begin{bmatrix} 1 & 1 & 0.8 & 0 & 0 & 0 \\ 1 & 1 & 0.8 & 0 & 0 & 0 \\ 0.8 & 0.8 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 1 & 1 & 1 \end{bmatrix}, R_q^\beta = \begin{bmatrix} 1 & 0 & 0.8 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0.8 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}.$$

$$\mathcal{T}_p = \{\emptyset, \{u_1, u_2, u_3\}, \{u_4, u_5, u_6\}, U\};$$

Khi đó:  $\mathcal{T}_q = \{\emptyset, \{u_1, u_2, u_3\}, \{u_4, u_5, u_6\}, \{u_1, u_3, u_4, u_5\}, \{u_2, u_6\}, \{u_2\},$

$$\{u_1, u_3, u_4, u_5, u_6\}, \{u_6\}, \{u_1, u_2, u_3, u_4, u_5\}, \{u_4, u_5\}, \{u_1, u_2, u_3, u_6\}, U\}$$

Do đó:  $\mathcal{T}_p \subseteq \mathcal{T}_q$

**Mệnh đề 4.2** (Thứ tự bộ phận của hai tôpô). Cho  $\mathcal{T}_p = \{X \subseteq U | \underline{R}_p^\beta(X) = \overline{R}_p^\beta(X)\}$  và

$\mathcal{T}_q = \{X \subseteq U | \underline{R}_q^\beta(X) = \overline{R}_q^\beta(X)\}$  là hai tôpô xác định trên  $U$  tương ứng với  $p, q \subseteq C$ .

Khi đó  $\mathcal{T}_p \subseteq \mathcal{T}_q$  nếu  $R_q^\beta \prec R_p^\beta$ .

*Chứng minh.* Giả sử rằng  $R_p^\beta \prec R_q^\beta$ , khi đó với mọi  $u \in U$ , ta có  $[u]_p^\beta \subseteq [u]_q^\beta$ . Do đó  $u \subseteq U$ , nếu  $[u]_q^\beta \subseteq X$  thì  $[x]_p^\beta \subseteq X$ . Ta có đpcm.  $\square$

**Ví dụ 4.3.** Quan sát ví dụ 4.2, chúng ta có thể thấy rõ vì  $R_p^\beta \prec R_q^\beta$  nên  $\mathcal{T}_p \leq \mathcal{T}_q$ .

**Định nghĩa 4.5** (Quan hệ mịn nhất). Cho  $R_1^\beta$  là một quan hệ tương đương mờ trên  $U$ , khi đó  $R_1^\beta$  được gọi là mịn nhất khi và chỉ khi với mọi  $p, q \in U$ ,  $R_1^\beta(p, q) = 1$  nếu  $p = q$  và  $R_1^\beta(p, q) = 0$  nếu  $p \neq q$ .

**Mệnh đề 4.3** (Tôpô lớn nhất). Cho  $\mathcal{T}_1 = \{X \subseteq U | \underline{R}^\beta(X) = \overline{R}^\beta(X)\}$ . Khi đó  $\mathcal{T}_1$  được gọi là lớn nhất nếu  $R^\beta = R_1^\beta$ .

*Chứng minh.* Chứng minh tương tự như mệnh đề 4.2 ta có đpcm.  $\square$

**Ví dụ 4.4.** Cho ma trận quan hệ mờ:  $R_1^\beta =$

$$\begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

Khi đó:  $\mathcal{T}_1 = \{X \subseteq U | \underline{R}^\beta(X) = \overline{R}^\beta(X)\} = P(U) = 2^U$

**Định nghĩa 4.6** (Quan hệ thô nhất). Cho  $R_0^\beta$  là một quan hệ tương đương mờ trên  $U$ , khi đó  $R_0^\beta$  được gọi là thô nhất khi và chỉ khi với mọi  $p, q \in U$ ,  $R_0^\beta(p, q) = 1$ .

**Mệnh đề 4.4** (Tôpô nhỏ nhất). Cho  $\mathcal{T}_0 = \{X \subseteq U \mid \underline{R}^\beta(X) = \overline{R}^\beta(X)\}$ . Khi đó  $\mathcal{T}_0$  được gọi là nhỏ nhất nếu  $R^\beta = R_0^\beta$ .

*Chứng minh.* Chứng minh tương tự như mệnh đề 4.2 ta có đpcm.  $\square$

**Ví dụ 4.5.** Cho ma trận quan hệ mờ:  $R_0^\beta = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 \end{bmatrix}$ .

Khi đó:  $\mathcal{T}_0 = \{X \subseteq U \mid \underline{R}^\beta(X) = \overline{R}^\beta(X)\} = \{\emptyset, U\}$

**Định nghĩa 4.7** (Quan hệ bù). Cho quan hệ tương đương  $R^\beta$  xác định trên  $U$ . Khi đó quan hệ bù của quan hệ  $R^\beta$  được định nghĩa như sau: với mọi  $p, q \in U$

$$(R^\beta)^c(p, q) = \begin{cases} 1 - R^\beta(p, q) & | 1 - R^\beta(p, q) \geq \beta \\ 0 & | 1 - R^\beta(p, q) < \beta \end{cases} \quad (4.2)$$

**Mệnh đề 4.5** (Tôpô bù). Cho  $\mathcal{T} = \{X \subseteq U \mid \underline{R}^\beta(X) = \overline{R}^\beta(X)\}$ .

Khi đó  $\sim \mathcal{T} = \{X \subseteq U \mid \sim \underline{R}^\beta(X) = \sim \overline{R}^\beta(X)\}$  là một tôpô bù của  $\mathcal{T}$ .

**Ví dụ 4.6.** Quan sát các ví dụ 4.4 và 4.5, chúng ta có thể thấy  $\mathcal{T}_0$  và  $\mathcal{T}_1$  là các tôpô bù của nhau.

**Định nghĩa 4.8** (Phép toán hợp hai tôpô). Cho  $\mathcal{T}_p = \{X \subseteq U \mid \underline{R}_p^\beta(X) = \overline{R}_p^\beta(X)\}$  và

$\mathcal{T}_q = \{X \subseteq U \mid \underline{R}_q^\beta(X) = \overline{R}_q^\beta(X)\}$  là hai tôpô xác định trên  $U$  tương ứng với  $p, q \subseteq C$ .

Khi đó hợp của hai tôpô được định nghĩa như sau:

$$\mathcal{T}_p \cup \mathcal{T}_q = \{X \subseteq U \mid X \in \mathcal{T}_p \vee X \in \mathcal{T}_q\} \quad (4.3)$$

**Mệnh đề 4.6** (Hợp hai tôpô theo tiếp cận RS). Cho  $\mathcal{T}_p = \{X \subseteq U \mid \underline{R}_p^\beta(X) = \overline{R}_p^\beta(X)\}$

và  $\mathcal{T}_q = \{X \subseteq U \mid \underline{R}_q^\beta(X) = \overline{R}_q^\beta(X)\}$  là hai tôpô xác định trên  $U$  tương ứng với  $p, q \subseteq C$

C. Khi đó, nếu  $R_q \prec R_p$  hoặc  $R_p \prec R_q$  thì  $\mathcal{T}_p \cup \mathcal{T}_q = \{X \subseteq U \mid \underline{R}_{pq}^\beta(X) = \overline{R}_{pq}^\beta(X)\}$ , với  $R_{pq}^\beta = R_p^\beta \cap R_q^\beta$ .

*Chứng minh.* Giả sử rằng  $R_q^\beta \prec R_p^\beta$  hoặc  $R_p^\beta \prec R_q^\beta$ , do đó với mọi  $u \in U$ , ta có  $[u]_p^\beta \subseteq [u]_q^\beta$  hoặc  $[u]_q^\beta \subseteq [u]_p^\beta$ . Khi đó, với mọi  $u \subseteq U$ , nếu  $[u]_q^\beta \subseteq X$  thì  $[u]_p^\beta \subseteq X$  hay nếu  $[u]_p^\beta \subseteq X$  thì  $[u]_q^\beta \subseteq X$ . Khi đó  $[u]_p^\beta \cap [u]_q^\beta = [u]_p^\beta$  hoặc  $[u]_p^\beta \cap [u]_q^\beta = [u]_q^\beta$ . Do đó, ta có đpcm.  $\square$

**Ví dụ 4.7.** Xem ví dụ 4.2 để thêm thông tin chi tiết.

**Định nghĩa 4.9** (Phép toán giao hai tôpô). Cho  $\mathcal{T}_p = \{X \subseteq U \mid \underline{R}_p^\beta(X) = \overline{R}_p^\beta(X)\}$  và  $\mathcal{T}_q = \{X \subseteq U \mid \underline{R}_q^\beta(X) = \overline{R}_q^\beta(X)\}$  là hai tôpô xác định trên  $U$  tương ứng với  $p, q \subseteq C$ . Khi đó giao của hai tôpô được định nghĩa như sau:

$$\mathcal{T}_p \cap \mathcal{T}_q = \{X \subseteq U \mid X \in \mathcal{T}_p \wedge X \in \mathcal{T}_q\} \quad (4.4)$$

**Mệnh đề 4.7** (Giao hai tôpô theo tiếp cận RS). Cho  $DT = (U, C, D, f)$  với  $\mathcal{T}_p = \{X \subseteq U \mid \underline{R}_p^\beta(X) = \overline{R}_p^\beta(X)\}$  và  $\mathcal{T}_q = \{X \subseteq U \mid \underline{R}_q^\beta(X) = \overline{R}_q^\beta(X)\}$  là hai tôpô xác định trên  $U$  tương ứng với  $p, q \subseteq C$ . Khi đó  $\mathcal{T}_p \cap \mathcal{T}_q = \{X \subseteq U \mid \underline{R}_{pq}^\beta(X) = \overline{R}_{pq}^\beta(X)\}$  với  $R_{pq}^\beta = R_p^\beta \cup R_q^\beta$ .

*Chứng minh.* Giả sử rằng  $X \in R_p^\beta \cup R_q^\beta$ , khi đó  $X \subseteq \underline{R}_p^\beta \cup \underline{R}_q^\beta(X)$ , do đó  $X \subseteq \underline{R}_{pq}^\beta \cap \underline{R}_{pq}^\beta \leftrightarrow X \subseteq \underline{R}_{pq}^\beta \leftrightarrow X \in \mathcal{T}_p \cap \mathcal{T}_q$ . Ta có đpcm.  $\square$

**Định nghĩa 4.10** (Nhóm tôpô). Cho không gian xấp xỉ mờ  $(U, R)$  và tôpô  $\mathcal{T} = \{X \subseteq U \mid \underline{R}(X) = \overline{R}(X)\}$  xác định trên  $U$ . Khi đó họ các tôpô cùng với các phép toán kí hiệu bởi  $G \langle (U, R), \cup, \cap, \sim, \mathcal{T}_0, \mathcal{T}_1 \rangle$  được gọi là một nhóm nếu với mọi  $\mathcal{T} \in G$  thỏa mãn:

- (1):  $(\mathcal{T}_a \cup \mathcal{T}_b) \cup \mathcal{T}_c = \mathcal{T}_a \cup (\mathcal{T}_b \cup \mathcal{T}_c)$ ; (2):  $(\mathcal{T}_a \cap \mathcal{T}_b) \cap \mathcal{T}_c = \mathcal{T}_a \cap (\mathcal{T}_b \cap \mathcal{T}_c)$ ;
- (3):  $(\mathcal{T}_a \cup \mathcal{T}_b) \cap \mathcal{T}_c = \mathcal{T}_a \cap \mathcal{T}_c \cup \mathcal{T}_b \cap \mathcal{T}_c$ ; (4):  $\mathcal{T}_a \cup \mathcal{T}_b = \mathcal{T}_b \cup \mathcal{T}_a$ ;
- (5):  $\mathcal{T}_a \cap \mathcal{T}_b = \mathcal{T}_b \cap \mathcal{T}_a$ ; (6):  $\mathcal{T}_0 \cup \mathcal{T}_a = \mathcal{T}_a$ ; (7):  $\mathcal{T}_1 \cap \mathcal{T}_a = \mathcal{T}_a$ .

**Mệnh đề 4.8** (Nhóm tôpô abel). Cho không gian xấp xỉ mờ  $(U, R^\beta)$  với tôpô  $\mathcal{T} =$

$\{X \subseteq U | \underline{R}^\beta(X) = \overline{R}^\beta(X)\}$  xác định trên  $U$ . Khi đó:  $G \langle (U, R^\beta), \cup, \cap, \sim, \mathcal{T}_0, \mathcal{T}_1 \rangle$  được gọi là một nhóm abel.

**Ví dụ 4.8.** Cho các tôpô  $\mathcal{T}_0, \mathcal{T}_1, \mathcal{T}_a, \mathcal{T}_b, \mathcal{T}_c \in G$  tương ứng với các quan hệ:

$$R_0^\beta = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 \end{bmatrix}, \quad R_1^\beta = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

$$R_a^\beta = \begin{bmatrix} 1 & 1 & 0.8 & 0 & 0 & 0 \\ 1 & 1 & 0.8 & 0 & 0 & 0 \\ 0.8 & 0.8 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 1 & 1 & 1 \end{bmatrix}, \quad R_b^\beta = \begin{bmatrix} 1 & 0 & 0.8 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0.8 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

$$R_c^\beta = \begin{bmatrix} 1 & 0 & 0.6 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0.6 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}, \quad U = \{u_1, u_2, u_3, u_4, u_5, u_6\}. \text{ Khi đó ta có:}$$

$$\mathcal{T}_0 = \left\{ X \subseteq U | \underline{R}_0^\beta(X) = \overline{R}_0^\beta(X) \right\} = \{\emptyset, U\}$$

$$\mathcal{T}_1 = \left\{ X \subseteq U | \underline{R}_1^\beta(X) = \overline{R}_1^\beta(X) \right\} = P(U) = 2^U$$

$$\mathcal{T}_a = \left\{ X \subseteq U | \underline{R}_a^\beta(X) = \overline{R}_a^\beta(X) \right\} = \{\emptyset, \{u_1, u_2, u_3\}, \{u_4, u_5, u_6\}, U\}$$

$$\mathcal{T}_b = \left\{ X \subseteq U | \underline{R}_b^\beta(X) = \overline{R}_b^\beta(X) \right\}$$

$$= \left\{ \begin{array}{l} \emptyset, \{u_1, u_2, u_3\}, \{u_4, u_5, u_6\}, \{u_1, u_3, u_4, u_5\}, \{u_2, u_6\}, \{u_2\}, \\ \{u_1, u_3, u_4, u_5, u_6\}, \{u_6\}, \{u_1, u_2, u_3, u_4, u_5\}, \{u_2, u_6\}, \{u_1, u_3, u_4, u_5\}, U \end{array} \right\}$$

(1):  $\forall i \left( R_a^\beta \cap R_b^\beta \right) \cap R_c^\beta = R_a^\beta \cap \left( R_b^\beta \cap R_c^\beta \right) = R_c^\beta$ , do đó  $(\mathcal{T}_a \cup \mathcal{T}_b) \cup \mathcal{T}_c = \mathcal{T}_a \cup (\mathcal{T}_b \cup \mathcal{T}_c) = \mathcal{T}_c$ ;

(2):  $\forall i \left( R_a^\beta \cup R_b^\beta \right) \cup R_c^\beta = R_a^\beta \cup \left( R_b^\beta \cup R_c^\beta \right) = R_a^\beta$ , do đó  $(\mathcal{T}_a \cup \mathcal{T}_b) \cup \mathcal{T}_c = \mathcal{T}_a \cup (\mathcal{T}_b \cup \mathcal{T}_c) = \mathcal{T}_a$ ;

(3):  $\forall i \left( R_a^\beta \cup R_b^\beta \right) \cap R_c^\beta = R_a^\beta \cap R_c^\beta \cup R_b^\beta \cap R_c^\beta = R_c^\beta$ , do đó  $(\mathcal{T}_a \cup \mathcal{T}_b) \cup \mathcal{T}_c = \mathcal{T}_a \cup (\mathcal{T}_b \cup \mathcal{T}_c) = \mathcal{T}_c$ ;

(4):  $\forall i R_a^\beta \cap R_b^\beta = R_b^\beta \cap R_a^\beta = R_b^\beta$ , do đó  $\mathcal{T}_a \cup \mathcal{T}_b = \mathcal{T}_b \cup \mathcal{T}_a = \mathcal{T}_b$ ;

(5):  $\forall i R_a^\beta \cup R_b^\beta = R_b^\beta \cup R_a^\beta = R_a^\beta$ , do đó  $\mathcal{T}_a \cap \mathcal{T}_b = \mathcal{T}_b \cap \mathcal{T}_a = \mathcal{T}_a$ ;

(6):  $\forall i R_0^\beta \cap R_a^\beta = R_a^\beta$ , do đó  $\mathcal{T}_0 \cup \mathcal{T}_a = \mathcal{T}_a$ ; (7):  $\forall i R_1^\beta \cup R_a^\beta = R_a^\beta$ , do đó  $\mathcal{T}_1 \cap \mathcal{T}_a = \mathcal{T}_a$ ;

### 4.3. Đề xuất cấu trúc tôpô Hausdorff

Bảng quyết định số  $DT = (U, C, D, f)$  như được trình bày trong Bảng 1.3 có thuộc tính quyết định  $D$  là các giá trị rời rạc, tức là các đối tượng giống nhau là không phân biệt được. Khi đó lớp tương đương của các đối tượng này không giao nhau, rõ ràng tôpô trên thuộc tính  $D$  là một cấu trúc tôpô Hausdorff. Do đó, để xác định một thuộc tính điều kiện  $c \in C$  có ảnh hưởng tới thuộc tính  $D$  thì thuộc tính đó cũng phải có cấu trúc tôpô Hausdorff. Do đó, phần này đề xuất xây dựng cấu trúc tôpô Hausdorff cho bài toán giảm thuộc tính.

**Định nghĩa 4.11** (Tính khả li của quan hệ mờ ngưỡng  $\beta$ ). Cho không gian xấp xỉ  $(U, R^\beta)$  trong đó  $R^\beta$  là quan hệ tương đương mờ  $\beta$ . Khi đó  $R^\beta$  được gọi là phân biệt được nếu với mọi  $p \in U$  tồn tại  $q \neq p \in U$  sao cho  $[p]_{R^\beta} \cap [q]_{R^\beta} = \emptyset$ . Kí hiệu quan hệ này là  $R_H^\beta$ .

**Mệnh đề 4.9** (Tôpô Hausdorff từ quan hệ  $R_H^\beta$ ). Cho tôpô  $\mathcal{T}_H = \{X \subseteq U \mid \underline{R}^\beta(X) = \overline{R}^\beta(X)\}$  xác định trên  $U$ . Khi đó,  $\mathcal{T}_H$  được gọi là tôpô Hausdorff nếu  $R^\beta$  là một  $R_H^\beta$ .

*Chứng minh.* Ta cần chứng minh hai điều kiện sau đây thỏa mãn:

(1): Chứng minh tương tự như mệnh đề 4.1, ta có  $\mathcal{T}_H$  là một tôpô xác định trên  $U$ ;

(2): Giả sử  $X, Y \in \mathcal{T}$ , khi đó:

$$\underline{R}_H^\beta(X) = \overline{R}_H^\beta(X) \Leftrightarrow \cup_{u \in U} \{ [u]_{R_H^\beta} \mid [u]_{R_H^\beta} \subseteq X \} = \cup_{u \in U} \{ [u]_{R_H^\beta} \mid [u]_{R_H^\beta} \cap X \neq \emptyset \}, \text{ và}$$

$$\underline{R}_H^\beta(Y) = \overline{R}_H^\beta(Y) \Leftrightarrow \cup_{y \in U} \{ [y]_{R_H^\beta} \mid [y]_{R_H^\beta} \subseteq Y \} = \cup_{y \in U} \{ [y]_{R_H^\beta} \mid [y]_{R_H^\beta} \cap Y \neq \emptyset \}.$$

$$\begin{aligned} X \cap Y &= \cup_{p, q \in U} \{ [u]_{R_H^\beta} \cap [y]_{R_H^\beta} \mid [u]_{R_H^\beta} \subseteq X, [y]_{R_H^\beta} \subseteq Y \} \\ \text{Hơn nữa:} &= \cup_{p, q \in U} \{ [u]_{R_H^\beta} \cap [y]_{R_H^\beta} \mid [u]_{R_H^\beta} \cap X \neq \emptyset, [y]_{R_H^\beta} \cap Y \neq \emptyset \}. \end{aligned}$$

Khi đó, nếu  $[u]_{R_H^\beta} \cap [y]_{R_H^\beta} = \emptyset$  thì  $X \cap Y = \emptyset$ .

Từ (1) và (2), ta có  $\mathcal{T}_H$  là một tôpô Hausdorff.  $\square$

**Mệnh đề 4.10** (Xác định thuộc tính có quan hệ  $R_H^\beta$ ). *Cho bảng quyết định  $DT = (U, C, D, f)$  và  $c \in C$ . Khi đó  $c$  được gọi là thuộc tính có quan hệ  $R_H^\beta$  nếu  $\max_1(V_c) - \max_2(V_c) > \beta$ . Trong đó  $V_c$  là tập giá trị của thuộc tính  $c$ .*

*Chứng minh.* Đặt,  $m_1 = \max_1(V_c)$  và  $m_2 = \max_2(V_c)$ . Rõ ràng nếu  $m_1 - m_2 > \beta$  thì với mọi  $m < m_2$  ta luôn có  $m_1 - m > \beta$ . Hơn nữa, theo công thức 4.1 nếu  $m_2 < m_1$  thì  $[m_1]_{R_H^\beta} \cap [m_2]_{R_H^\beta} = \emptyset$ , do đó  $[m_1]_{R_H^\beta} \cap [m]_{R_H^\beta} = \emptyset$ . ta có đpcm.  $\square$

## 4.4. Giảm thuộc tính trong bảng quyết định theo tiếp cận tôpô Hausdorff

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

**Định nghĩa 4.12** (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  $\mathcal{T}_c$  là một tôpô Hausdorff.

**Định nghĩa 4.13** (Đồng cấu trúc phụ thuộc). Cho bảng quyết định  $DT = (U, C, D, f)$  và hai tôpô  $\mathcal{T}_p, \mathcal{T}_q$  xác định trên  $U$  tương ứng với  $p, q \in C$ . Khi đó  $\mathcal{T}_p$  được gọi là

đồng cấu trúc phụ thuộc với  $\mathcal{T}_q$  nếu  $\mathcal{T}_p \cup \mathcal{T}_D = \mathcal{T}_q \cup \mathcal{T}_D$ .

Sau đây là thuật toán giảm thuộc tính theo phương pháp phân cụm và đánh giá các nhóm thuộc tính Hausdorff.

---

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

---

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

**Output** Tập rút gọn  $R$

```

1:  $R = \emptyset$ ;
2: for all  $\beta \in \Delta$  do
3:    $H^\beta \leftarrow \emptyset$ ;
4:    $CH^\beta \leftarrow \emptyset$ ;
5:    $R^\beta \leftarrow \emptyset$ ;
6:   for all  $c \in C$  do
7:     if  $\max_1(V_c) - \max_2(V_c) > \beta$  then
8:        $H^\beta = H^\beta \cup \{c\}$ ;                                {Filter các thuộc tính Hausdorff}
9:     end if
10:  end for
11:  for all  $p \in \{H^\beta - CH^\beta\}$  do
12:     $U_p = \emptyset$ ;
13:    for all  $q \in \{H^\beta - CH^\beta - p\}$  do
14:      if  $\mathcal{T}_p \cup \mathcal{T}_D = \mathcal{T}_q \cup \mathcal{T}_D$  then
15:         $U_p = U_p \cup \{q\}$ ;                                {Phân cụm thuộc tính Hausdorff}
16:      end if
17:    end for
18:     $CH^\beta = CH^\beta \cup U_p$ ;
19:    if  $ACC_{U_p}^{Model} > ACC_{R^\beta}^{Model}$  then
20:       $R^\beta = U_p$ ;                                        {Wrapper các cụm thuộc tính Hausdorff}
21:    end if
22:  end for
23:  if  $ACC_{R^\beta}^{Model} > ACC_R^{Model}$  then
24:     $R = R^\beta$ ;                                          {Wrapper các reduct ứng viên  $\beta$ }
25:  end if
26: end for
27: return  $R$ ;

```

---

Trong thuật toán này, kí hiệu  $H^\beta$  là tập các thuộc tính Hausdorff thu được từ tập thuộc tính  $C$  ban đầu trong bảng quyết định  $DT$ .  $CH^\beta$  là các cụm thuộc tính được



phân loại từ  $H^\beta$ , trong đó mỗi cụm thuộc tính  $U_p$  là các thuộc tính  $q \in H^\beta$  có cùng cấu trúc phụ thuộc với thuộc tính  $p$  theo định nghĩa 4.13.

Tiếp theo sẽ là phân 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  $\mathcal{O}(2|U||C|)$ , độ phức tạp từ 11-22 là  $\mathcal{O}(|U|^2|H^\beta|^2)$ . Giả sử  $\mathbb{T}$  là chi phí tính toán của mô hình phân lớp *Model*. Với số lượng  $\Delta$  rất nhỏ, khi đó độ phức tạp của thuật toán là  $\mathcal{O}(2|U||C|) + \mathcal{O}(|U|^2|H^\beta| + |CH^\beta|\mathbb{T})$ .

**Ví dụ 4.9.** Xét bảng quyết định  $DT = (U, C, D, f)$  như được trình bày trong Bảng 1.3. Xét  $\beta = 0.7$  ta có:

Bước 1:  $H^{0.7} \leftarrow \emptyset; CH^\beta \leftarrow \emptyset; R^\beta \leftarrow \emptyset;$

Bước 2: Xác định các thuộc tính Hausdorff dựa trên mệnh đề 4.10 ta có  $H^{0.7} = \{a, c, d, e, f\};$

Bước 3: Tính các ma trận quan hệ của các thuộc tính trong  $H^{0.7}$  theo công thức quan hệ 4.1 với  $\beta = 0.7$  ta có:

$$R_a^{0.7} = \begin{bmatrix} 1 & 1 & 0.8 & 0 & 0 & 0 \\ 1 & 1 & 0.8 & 0 & 0 & 0 \\ 0.8 & 0.8 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 1 & 1 & 1 \end{bmatrix} \quad R_c^{0.7} = \begin{bmatrix} 1 & 0 & 0.8 & 1 & 1 & 0 \\ 0 & 1 & 0 & 0 & 0 & 1 \\ 0.8 & 0 & 1 & 0.8 & 0.8 & 0 \\ 1 & 0 & 0.8 & 1 & 1 & 0 \\ 1 & 0 & 0.8 & 1 & 1 & 0 \\ 0 & 1 & 0 & 0 & 0 & 1 \end{bmatrix}$$

$$R_d^{0.7} = \begin{bmatrix} 1 & 0.8 & 0.8 & 1 & 1 & 0 \\ 0.8 & 1 & 0 & 0.8 & 0.8 & 0 \\ 0.8 & 0 & 1 & 0.8 & 0.8 & 0 \\ 1 & 0.8 & 0.8 & 1 & 1 & 0 \\ 1 & 0.8 & 0.8 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} \quad R_D = \begin{bmatrix} 1 & 0 & 1 & 0 & 0 & 1 \\ 0 & 1 & 0 & 1 & 1 & 0 \\ 1 & 0 & 1 & 0 & 0 & 1 \\ 0 & 1 & 0 & 1 & 1 & 0 \\ 0 & 1 & 0 & 1 & 1 & 0 \\ 1 & 0 & 1 & 0 & 0 & 1 \end{bmatrix}$$

$$R_f^{0.7} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0.8 & 0.8 & 0.8 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0.8 & 0 & 1 & 1 & 1 \\ 0 & 0.8 & 0 & 1 & 1 & 1 \\ 0 & 0.8 & 0 & 1 & 1 & 1 \end{bmatrix} \quad R_e^{0.7} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0.8 & 0.8 & 0.8 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0.8 & 0 & 1 & 1 & 1 \\ 0 & 0.8 & 0 & 1 & 1 & 1 \\ 0 & 0.8 & 0 & 1 & 1 & 1 \end{bmatrix}$$

Vì  $R_a^{0.7} \cap R_D^{0.7} = R_c^{0.7} \cap R_D^{0.7} \Leftrightarrow \mathcal{T}_a \cup \mathcal{T}_D = \mathcal{T}_c \cup \mathcal{T}_D$ , do đó  $CH^{0.7} = \emptyset \cup \{a, c\} = \{\{a, c\}\}$ ;  $R_e^{0.7} \cap R_D^{0.7} = R_f^{0.7} \cap R_D^{0.7} \Leftrightarrow \mathcal{T}_e \cup \mathcal{T}_D = \mathcal{T}_f \cup \mathcal{T}_D$ .

Do đó  $CH^{0.7} = \{a, c\} \cup \{e, f\} = \{\{a, c\}, \{e, f\}\}$ ;

Cuối cùng thuộc tính  $d$  được bổ xung vào  $CH^{0.7}$ , do đó  $CH^{0.7} = \{\{a, c\}, \{e, f\}, \{d\}\}$ .

Bước 4: wrapper từng nhóm thuộc tính trên mô hình phân lớp *Model*, nhóm thuộc tính nào có khả năng phân lớp cao nhất thì nhóm đó được gán cho  $R^{0.7}$

Bước 5: Giả sử  $R^{0.7}$  có khả năng phân lớp cao nhất trong các ngưỡng  $\delta$ . Khi đó reduct thu được của thuật toán  $R = R^{0.7}$ .

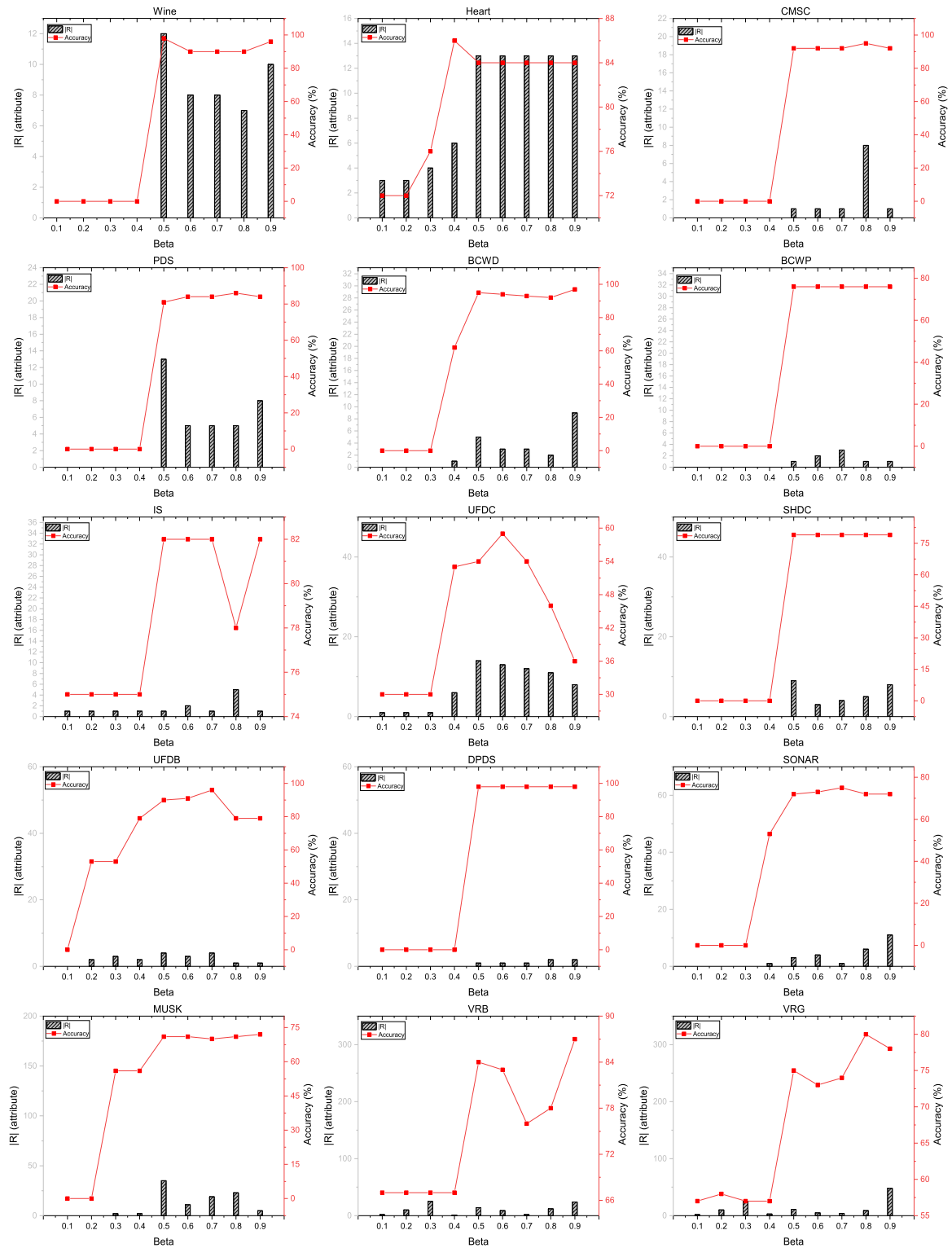
#### 4.4.2. Thực nghiệm và đánh giá thuật toán

Mục tiêu của phần thực nghiệm nhằm đánh giá tính hiệu quả của thuật toán đề xuất khi áp dụng với các bộ dữ liệu trong thực tiễn. Trên cơ sở đó có thể khẳng định tính đúng đắn của khung nền tảng lý thuyết bài toán giảm thuộc tính theo tiếp cận tôpô Hausdorff. Sau đây là kế hoạch thực nghiệm thuật toán đề xuất.

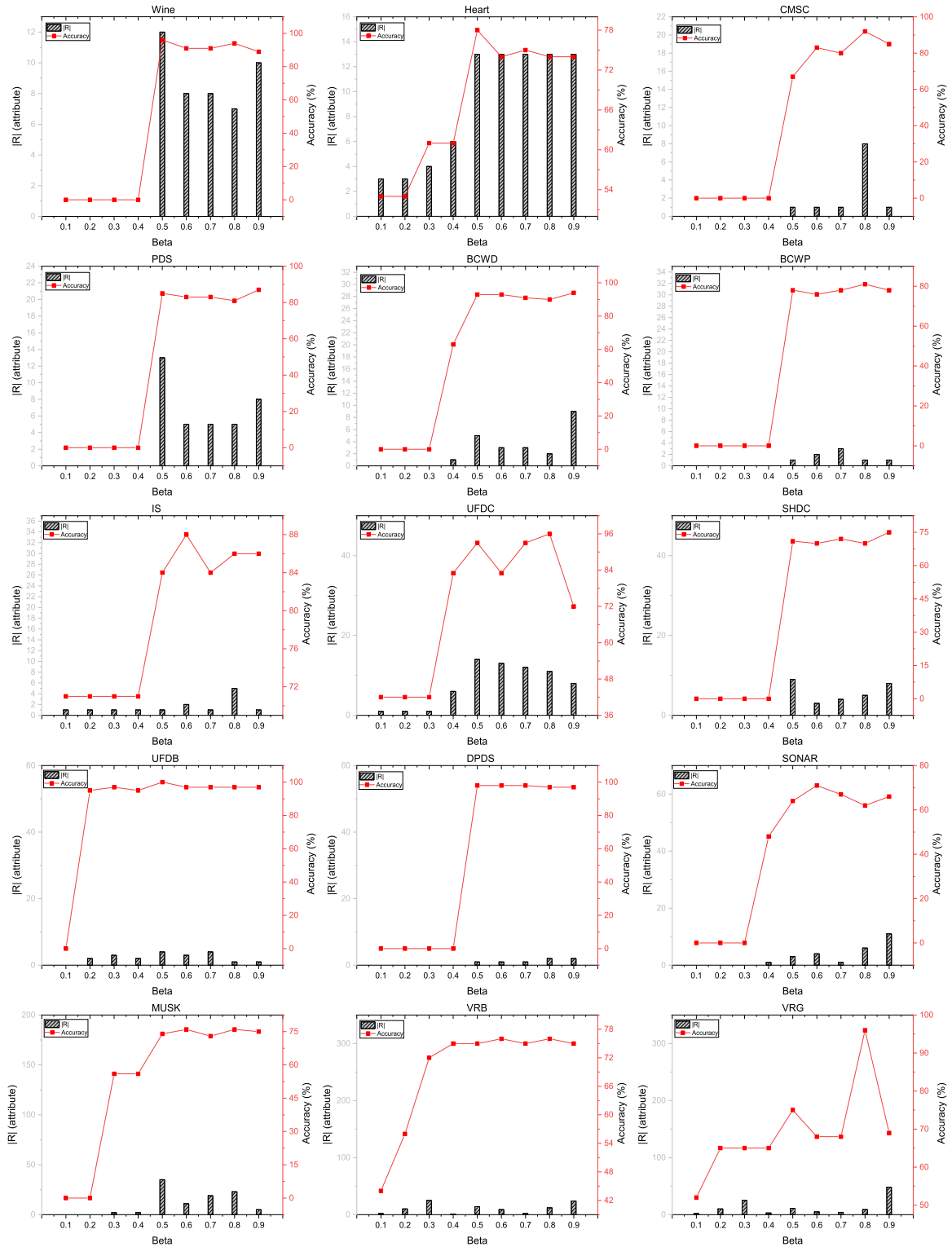
##### 4.4.2.1. Kích bản và môi trường thực nghiệm

1) Chọn lựa giá trị  $\beta$ . Mục tiêu của quá trình thực nghiệm này nhằm tìm kiếm giá trị  $\beta$  phù hợp nhất cho từng bộ dữ liệu của thuật toán đề xuất, trong đó giá trị  $\beta$  được chọn trong khoảng  $[0.1, 0.9]$  với mỗi bước nhảy là 0.1.

2) Đánh giá thuật toán đề xuất. Sau khi chọn lựa được các giá trị  $\beta$  phù hợp, thực hiện so sánh và đánh giá thuật toán đề xuất với các thuật toán giảm thuộc tính điển



**Hình 4.1:** Biểu đồ phân tích sự tương quan giữa số lượng phân tử và khả năng phân lớp chính xác của reduct tại mỗi giá trị  $\beta$  trên mô hình phân lớp SVM.



**Hình 4.2:** Biểu đồ phân tích sự tương quan giữa số lượng phân tử và khả năng phân lớp chính xác của reduct tại mỗi giá trị  $\beta$  trên mô hình phân lớp KNN.

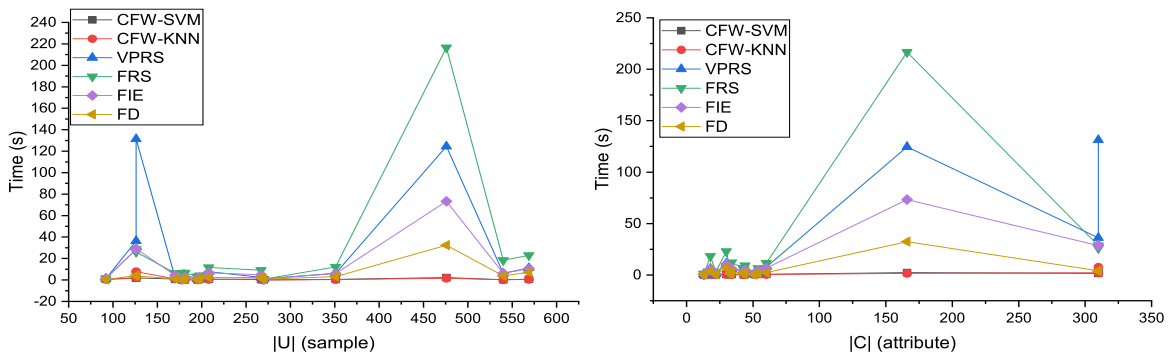
**Bảng 4.1:** Mô tả các tập dữ liệu thực nghiệm

ID	Data	Describe	$ U $	$ C $	$ D $
1	Wine	Wine	178	13	3
2	Heart	Statlog (Heart)	270	13	2
3	CMSC	Climate Model Simulation Crashes Data Set	540	18	2
4	PDS	Parkinsons Data Set	196	22	2
5	BCWD	Breast Cancer Wisconsin (Diagnostic)	569	30	2
6	BCWP	Breast Cancer Wisconsin (Prognostic)	198	33	2
7	IS	Ionosphere	351	34	2
8	UFDC	Ultrasonic flowmeter diagnostics (C)	181	43	4
9	UFDD	Ultrasonic flowmeter diagnostics (D)	181	43	4
10	SHDC	SPECTF Heart Data Set	267	44	2
11	UFDB	Ultrasonic flowmeter diagnostics (B)	92	51	3
12	DPDS	Divorce Predictors data set	170	54	2
13	Sona	Connectionist Bench	208	60	2
14	Musk	Musk	476	166	2
15	VRB	Voice Rehabilitation(Binary)	126	310	2
16	VRG	Voice Rehabilitation(Gender)	126	310	2

hình trên tiếp cận độ đo gồm có: (1) thuật toán giảm thuộc tính theo tiếp cận RS với độ chính xác điều chỉnh (VPRS) [132]; (2) thuật toán giảm thuộc tính theo tiếp cận RS mờ (FRS) [68]; (3) thuật toán giảm thuộc tính theo tiếp cận Entropy thông tin mờ (IFE) [95]; (4) thuật toán giảm thuộc tính theo tiếp cận khoảng cách mờ (FD) [33].

Tất cả các thuật toán đều được đánh giá trên 16 bộ dữ liệu được tải về từ kho dữ liệu học máy UCI. Các tập dữ liệu được chọn là các tập dữ liệu có thuộc tính quyết định miền giá trị số và thuộc tính điều kiện có miền giá trị rời rạc. Các thuật toán được cài đặt bằng ngôn ngữ lập trình Python trên nền hệ điều hành Window 10 với cấu hình phần cứng là bộ xử lý Core-i5, bộ nhớ RAM 8G.

Các tập dữ liệu được sắp xếp theo trật tự tăng dần về số lượng thuộc tính điều kiện. Chi tiết các tập dữ liệu được mô tả trong Bảng 4.1 trong đó kí hiệu  $|U|$  là số lượng mẫu, kí hiệu  $|C|$  là số lượng các thuộc tính điều kiện và  $|D|$  là số phân lớp trong thuộc tính quyết định. Trong các bộ dữ liệu thực nghiệm, bộ dữ liệu UFDC và Sonar là các bộ dữ liệu nhiễu, có khả năng phân lớp ban đầu trên mô hình huấn luyện *Model* thấp



**Hình 4.3:** Biểu đồ phân tích sự tương quan giữa chi phí tính toán của thuật toán và  $|U|$  (left), giữa chi phí tính toán của thuật toán và  $|C|$  (right).

< 70%. Các tiêu chí đánh giá gồm: chi phí tính toán của thuật toán (seconds), số lượng phần tử của reduct ( $|R|$ ), và khả năng phân lớp của tập dữ liệu trên mô hình phân lớp dữ liệu *Model* (percentage). Quá trình thực nghiệm cũng coi trọng khả năng loại bỏ nhiễu của các thuật toán trên các tập dữ liệu xấu. Trước khi thực hiện thuật toán giảm thuộc tính, các tập dữ liệu được chuẩn hóa giá trị về đoạn  $[0, 1]$  để nâng cao hiệu năng cho thuật toán và các mô hình phân lớp.

Mỗi thuật toán được thực hiện 10 lần trên từng bộ dữ liệu với 90% dữ liệu được lấy ngẫu nhiên từ tập dữ liệu gốc. Hai mô hình phân lớp được sử dụng để đánh giá gồm có mô hình phân lớp máy vector hỗ trợ<sup>1</sup> (Support Vector Machine - SVM) và mô hình phân lớp k-láng giềng<sup>2</sup> (k-Nearest Neighbor - kNN,  $k=|D|$ ). Chỉ số đánh giá độ chính xác (accuracy) và phương pháp đánh giá chéo 10-fold được kết hợp để đánh giá chất lượng phân lớp của reduct.

#### 4.4.2.2. Chọn lọc giá trị $\beta$ cho mỗi tập dữ liệu

Trước khi thực hiện so sánh thuật toán đề xuất với các thuật toán giảm thuộc tính khác, ta cần lựa chọn giá trị  $\beta$  phù hợp nhất trên từng bộ dữ liệu khác nhau cho thuật toán. Với mỗi tập dữ liệu thực nghiệm, thực nghiệm thuật toán với từng giá trị  $\beta$  khác nhau trong khoảng  $[0.1, 0.9]$  với mỗi bước nhảy là 0.1. Như vậy, mỗi bộ dữ liệu được

<sup>1</sup><https://brilliant.org/wiki/support-vector-machines/>

<sup>2</sup><https://brilliant.org/wiki/k-nearest-neighbors/>

**Bảng 4.2:** So sánh số lượng phần tử của reduct thu được từ các thuật toán

ID	Dataset	IRI						
		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

thực nghiệm với 09 lần khác nhau. Với mỗi giá trị  $\beta$  khác nhau, số lượng phần tử reduct sẽ khác nhau và khả năng phân lớp với mỗi reduct cũng có thể khác nhau. Tuy nhiên, với các giá trị  $\beta$  quá nhỏ, có thể sẽ không tồn tại reduct.

Quan sát biểu đồ của các tập dữ liệu (Wine, CSMC, PDSB, BCWP, ..) trong Hình 4.1 và Hình 4.2 để biết thêm thông tin chi tiết. Hơn nữa, biểu đồ trong các hình này cũng phân tích sự tương quan giữa số lượng phần tử của reduct và khả năng phân lớp trên từng tập dữ liệu. Chúng ta có thể thấy, nhiều giá trị  $\beta$  cho reduct số lượng phần tử nhỏ nhưng độ chính xác lại cao hơn so với các giá trị  $\delta$  khác. Tuy nhiên môi quan hệ nay không tuyến tính nên chúng ta phải cân nhắc giá trị  $\beta$  cho phù hợp với mục tiêu về độ chính xác hay số lượng phần tử của reduct. Ở đây, lựa chọn giá trị  $\beta$  sao cho đảm bảo tính cân bằng giữa khả năng phân lớp và số lượng phần tử của reduct thu được từ thuật toán đề xuất.

**Bảng 4.3:** So sánh khả năng phân lớp chính xác của reduct 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	<b>59±0.7</b>	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	<b>73±0.2</b>	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

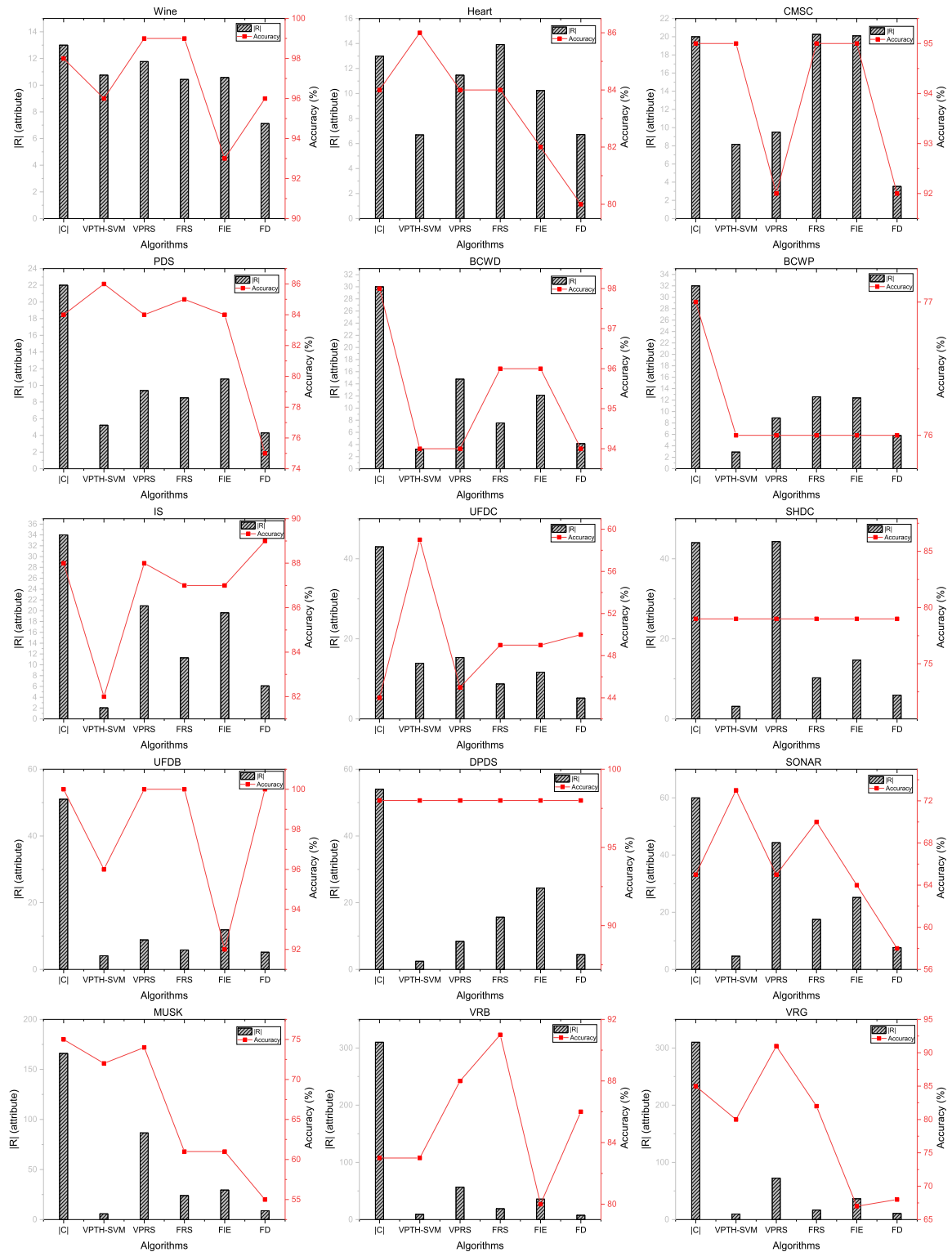
#### 4.4.2.3. Đánh giá thuật toán đề xuất

Giai đoạn wrapper của thuật toán đề xuất sử dụng hai mô hình phân lớp là SVM và k-NN( $k=|D|$ ). Trước tiên, chương này đánh giá hiệu năng của thuật toán đề xuất trên mô hình phân lớp SVM.

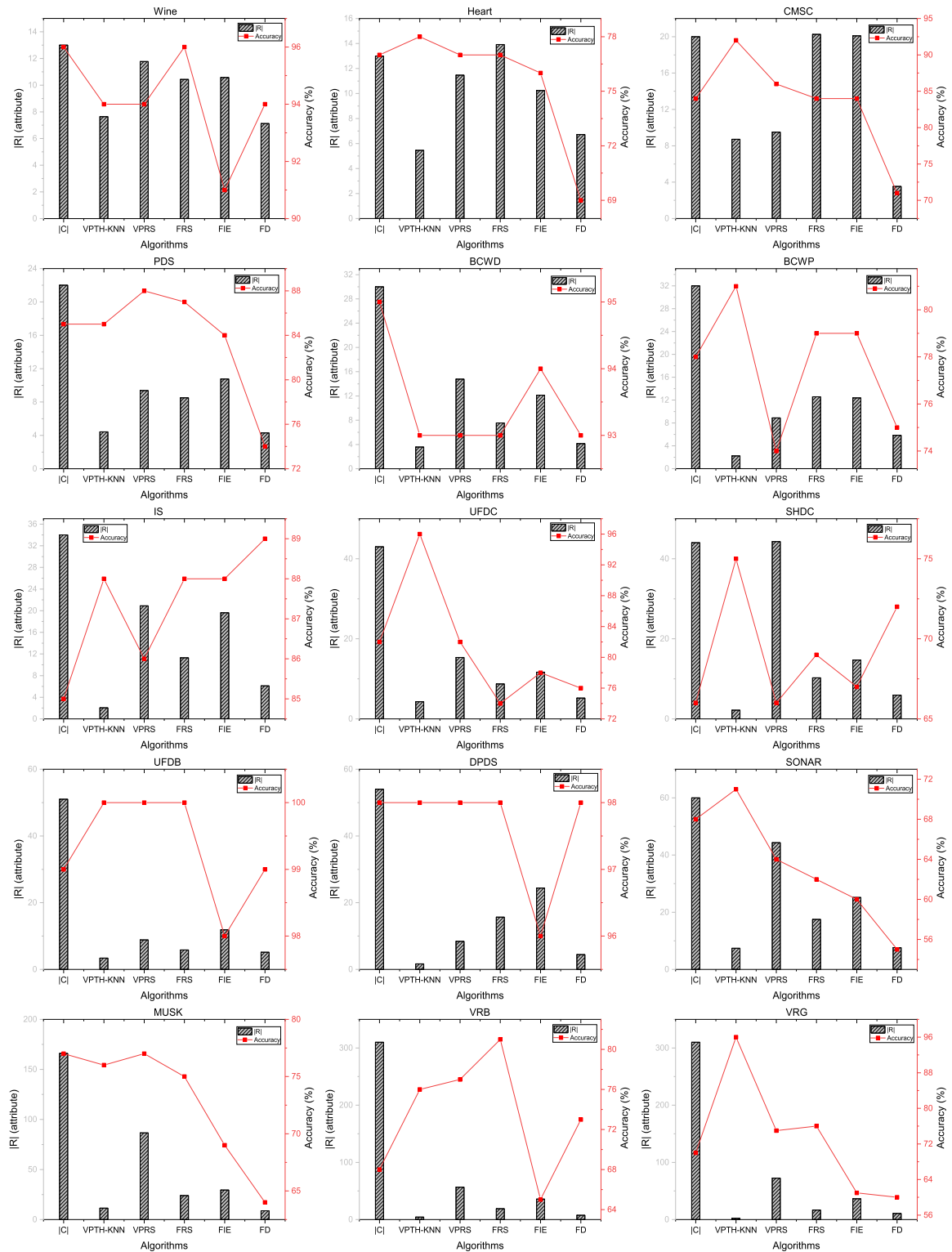
##### *Đánh giá thuật toán trên mô hình phân lớp SVM*

Các kết quả thực nghiệm của thuật toán trên mô hình phân lớp SVM được phân tích và thống kê về số lượng phần tử reduct thu được trình bày trong Bảng 4.2, về khả năng phân lớp được trình bày trong Bảng 4.3, và chi phí tính toán được trình bày trong Bảng 4.5. Quan sát dữ liệu trong Bảng 4.2 ta có thể thấy thuật toán đề xuất CFW và thuật toán giảm thuộc tính theo tiếp cận khoảng cách mờ FD cho reduct có số lượng phần tử tốt nhất, tuy nhiên khả năng phân lớp trên reduct của thuật toán đề xuất hoàn toàn tốt hơn thuật toán theo tiếp cận FD. Bảng 4.3 cho thấy thuật toán đề xuất CFW





**Hình 4.4:** Biểu đồ phân tích sự tương quan giữa số lượng phần tử và khả năng phân lớp chính xác của reduct của mỗi thuật toán trên mô hình phân lớp SVM.



**Hình 4.5:** Biểu đồ phân tích sự tương quan giữa số lượng phần tử và khả năng phân lớp chính xác của reduct của mỗi thuật toán trên mô hình phân lớp KNN.

**Bảng 4.4:** So sánh khả năng phân lớp chính xác của reduct 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

và hai thuật toán dựa trên độ đo miền dương là VPRS và FRS cho reduct có độ chính xác tốt nhất, hầu như không chênh lệch so với tập dữ liệu gốc. Tuy nhiên số lượng phần tử reduct thu được từ thuật toán đề xuất hoàn toàn vượt trội so với hai thuật toán theo tiếp cận VPRS và FRS. Đặc biệt là các bộ dữ liệu nhiễu (**UFDS, Sonar**), khả năng phân lớp được cải thiện từ 44% to 59%. Bảng 4.5 cho thấy thuật toán đề xuất có thời gian hoàn toàn vượt trội so với các thuật toán khác.

Quan sát biểu đồ trong Hình 4.4 ta có thể thấy sự tương quan về số lượng phần tử và khả năng phân lớp của các reduct thu được từ các thuật toán, hầu hết các reduct của thuật toán đề xuất đều có số lượng phần tử nhỏ hơn nhưng khả năng phân lớp không chênh lệch so với các thuật toán tốt nhất. Quan sát biểu đồ của các tập dữ liệu (**Heart, CMCS, PDS, BCWP, UFDC, SHDC, DPDS, Sonar**) ta có thể thấy tính hiệu quả về số lượng phần tử và khả năng phân lớp của thuật toán đề xuất là hoàn toàn vượt trội so với các thuật toán khác. Qua đó ta có thể thấy nhiều reduct có số lượng phần tử lớn

**Bảng 4.5:** So sánh chi phí tính toán của các thuật toán

ID	Dataset	Computation time (s)					
		CFW-SVM	CFW-kNN	VPRS	FRS	FIE	FD
1	wine	0.05	0.07	0.7	1.27	0.58	0.3
2	heart	0.08	0.1	1.08	0.79	1.03	0.63
3	CMSC	0.28	0.26	6.19	18.28	6.14	3.79
4	PDS	0.17	0.09	1.24	3.09	1.23	0.6
5	BCWD	0.67	0.49	11.19	23.01	10.69	7.3
6	BCWP	0.71	0.57	1.81	4.13	2.14	1.04
7	IS	0.6	0.43	6.13	12.17	6.08	3.17
8	UFDC	0.11	0.49	4.23	6.73	2.49	1.17
9	UFDD	0.18	0.62	4.84	5.43	2.11	1.12
10	SHDC	0.45	0.85	2.41	9.16	4.8	2.29
11	UFDB	0.56	0.77	1.22	1.39	1.22	0.38
12	DPDS	0.82	1.06	2.41	6.42	4.34	1.31
13	sonar	0.63	0.52	7.8	11.63	6.65	2.25
14	musk	2.26	1.55	124.48	216.6	73.32	32.31
15	VRB	1.7	2.12	36.25	28.67	28.59	4.1
16	VRG	1.93	7.67	131.24	26.09	29.11	3.63

nhưng chưa chắc đã có độ chính xác cao hơn.

Tóm lại, thuật toán đề xuất thực hiện trên mô hình phân lớp SVM cho reduct có số lượng phần tử và khả năng phân lớp là không chênh lệch đáng kể so với các thuật toán tốt nhất. Tuy nhiên chi phí tính toán của thuật toán đề xuất là hoàn toàn vượt trội so với các thuật toán khác.

#### *Đánh giá thuật toán trên mô hình phân lớp KNN*

Kết quả thực nghiệm trên mô hình phân loại kNN của thuật toán đề xuất được mô tả chi tiết trong Bảng 4.2, 4.4 và 4.5. Kết quả trung bình của các tiêu chí đánh giá về số lượng phần tử, khả năng phân lớp và chi phí tính toán của thuật toán đề xuất đều có kết quả vượt trội so với các thuật toán khác. Quan sát Bảng 4.2 ta có thể thấy số lượng phần tử trung bình của reduct thu được từ thuật toán đề xuất có kết quả tốt nhất và tốt hơn khi thực hiện trên mô hình phân lớp SVM. Quan sát Bảng 4.4 và Bảng 4.5 ta thấy khả năng phân lớp chính xác của reduct và chi phí tính toán của thuật toán đề xuất là hoàn toàn vượt trội so với các thuật toán tốt nhất.

Quan sát biểu đồ trong Hình 4.4 ta có thể thấy sự tương quan về số lượng phần tử và khả năng phân lớp của các reduct thu được từ các thuật toán, hầu hết các reduct của thuật toán đề xuất đều có số lượng phần tử nhỏ hơn nhưng khả năng phân lớp lại cao hơn so với các thuật toán tốt nhất. Quan sát biểu đồ của các tập dữ liệu (**Heart**, **CMCS**, **BCWP**, **IS**, **UFDC**, **UFDD**, **SHDC**, **UFDB**, **DPDS**, **Sonar**, **Musk**, **VRG**). Đặc biệt bộ dữ liệu nhiễu (**VRG**) cải thiện khả năng phân lớp từ 70% lên 96%.

**Bảng 4.6:** Mô tả reduct thu được từ các thuật toán

ID	Dataset	Reducts					
		CFW-SVM	CFWK-NN	VPRS	FRS	FIE	FD
1	wine	[0, 2, 3, 4, 5, 7, 8, 9, 10, 11]	[0, 4, 5, 6, 7, 8, 10]	[12, 11, 0, 9, 6, 1, 3, 10, 7, 5, 2]	[12, 9, 5, 0, 1, 11, 3, 4, 2, 6]	[7, 1, 5, 3, 4, 8, 0, 11, 9, 10]	[11, 12, 7, 0, 1, 5, 9]
2	heart	[1, 2, 5, 8, 11, 12]	[2, 3, 4, 9, 11]	[12, 2, 11, 6, 8, 10, 1, 5, 7, 3, 9]	[0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12]	[6, 1, 8, 12, 5, 10, 2, 11, 0, 3]	[12, 6, 8, 1, 2, 10]
3	CMSC	[2, 3, 4, 5, 8, 9, 14, 15]	[2, 3, 4, 5, 8, 9, 14, 15]	[2, 3, 19, 9, 11, 7, 12, 18, 14]	[2, 3, 15, 0, 5, 1, 6, 4, 7, 9, 13, ...]	[0, 1, 4, 6, 5, 7, 10, 8, 9, 3, ...]	[0, 1, 5]
							Tiếp theo trang sau

**Bảng 4.6 – Tiếp theo trang trước**

ID	Dataset	Reducts					
		CFW-SVM	CFWK-NN	VPRS	FRS	FIE	FD
4	PDS	[16, 17, 18, 19, 21]	[15, 18, 19, 20]	[0, 2, 18, 17, 19, 20, 21, 9, 11]	[18, 0, 10, 16, 17, 2, 19, 20]	[16, 2, 17, 1, 0, 10, 3, 19, 6, 20]	[0, 16, 2, 1]
5	BCWD	[0, 20, 27]	[0, 20, 27]	[27, 7, 20, 25, 6, 28, 16, 0, 17, ...]	[20, 27, 1, 11, 4, 8, 21]	[9, 21, 27, 6, 8, 11, 3, 15, 1, 18, ..]	[27, 7, 6, 20]
6	BCWP	[0, 26]	[11, 31]	[23, 16, 15, 0, 19, 20, 24, 22]	[0, 4, 2, 5, 6, 19, 11, 9, 12, .., 20]	[0, 31, 6, 19, 1, 11, 22, 5, 20, 8, 12, 18]	[0, 1, 31, 23, 19]
7	IS	[4, 26]	[4, 26]	[0, 4, 2, 7, 9, 5, 27, 23, 29, 13, ...]	[0, 4, 2, 5, 27, 30, 7, 3, 9, 16, 17]	[14, 0, 27, 28, 7, 31, 18, 23, 26, 4, ...]	[14, 0, 4, 24, 22, 27]
8	UFDC	[0, 8, 9, 10, 25, 27, 28, 29, 30, 31, 32, 33, 34]	[29, 30, 31, 32]	[23, 27, 5, 25, 36, 32, 12, 11, 35, 21, ...]	[7, 9, 25, 5, 27, 0, 39, 11]	[3, 23, 8, 25, 0, 2, 27, 15, 29, 33, 31]	[3, 23, 27, 7, 25]
Tiếp theo trang sau							

**Bảng 4.6 – Tiếp theo trang trước**

ID	Dataset	Reducts					
		CFW-SVM	CFWK-NN	VPRS	FRS	FIE	FD
9	UFDD	[10, 27, 28, 41, 42]	[25, 29, 30]	[33, 27, 6, 40, 5, 30, 12, ...]	[25, 27, 17, 3, 11, 42]	[5, 27, 21, 42, 39, 31, 0, 41]	[33, 5, 27]
10	SHDC	[0, 21, 22]	[25, 41]	[0, 1, 2, 3, 4, 5, 6, 7, 8, 9, ...]	[40, 29, 1, 2, 13, 25, 3, 4, 18, 9]	[43, 3, 18, 1, 12, 14, 21, 9, ...]	[43, 24, 41, 29, 36]
11	UFDB	[17, 22, 41, 42]	[19, 39, 40]	[41, 14, 4, 32, 15, 29, 34, 23]	[41, 14, 13, 16, 12]	[35, 6, 12, 39, 31, 0, 19, 22, ..]	[41, 14, 16, 20, 7]
12	DPDS	[10, 19]	[18]	[16, 39, 25, 43, 8, 24, 40, 27]	[10, 32, 30, 0, 48, 15, 6, 39, 35, ...]	[44, 45, 3, 51, 34, 5, 30, 46, 41, ...]	[39, 34, 10, 43]
13	sonar	[11, 12, 29, 32]	[9, 16, 17, 43, 47, 52, 53]	[35, 20, 11, 19, 44, 7, 36, 16, ...]	[0, 11, 15, 36, 26, 19, 21, 9, ...]	[19, 25, 16, 22, 34, 27, ...]	[19, 35, 16, 22, 25, 34, 28]
Tiếp theo trang sau							

**Bảng 4.6 – Tiếp theo trang trước**

ID	Dataset	Reducts					
		CFW-SVM	CFWK-NN	VPRS	FRS	FIE	FD
14	musk	[62, 101, 161, 163, 164]	[12, 20, 50, 62, 93, 104, 109, 110, 128, 140, 161]	[91, 36, 76, 57, 15, 31, 156, 162, 135, 83, ...]	[49, 16, 147, 3, 96, 1, 38, 13, 31, 87, ...]	[16, 124, 1, 36, 23, 46, 60, 68, 131, 15, ...]	[31, 16, 40, 131, 124, 63, 64, 36]
15	VRB	[58, 83, 92, 107, 117, 121, 152, 197, 279]	[79, 84, 270, 271]	[72, 79, 26, 70, 52, 77, 83, ...]	[54, 84, 79, 83, 91, 41, 3, ...]	[58, 70, 59, 60, 62, 92, ...]	[79, 58, 70, 82, 59, 84, 69]
16	VRG	[75, 92, 117, 124, 196, 197, 198, 279, 309]	[124, 309]	[60, 69, 62, 30, 85, 86, 77, 25, 70, ...]	[90, 95, 31, 102, 59, 73, 16, 82, 89, ...]	[59, 16, 70, 62, 57, 92, 69, 127, ...]	[79, 58, 70, 86, 84, 121, 56, 69, 16, 62]

Nhìn chung, các kết quả thực nghiệm của thuật toán đề xuất trên mô hình phân lớp kNN đều có kết quả tốt hơn so với các thuật toán còn lại. Trong đó khả năng cải thiện nhiều và chi phí tính toán của thuật toán là hoàn toàn vượt trội. Dưới góc nhìn thực nghiệm, ta có thể khẳng định tiếp cận giảm thuộc tính theo tiếp cận tô pô là hoàn toàn phù hợp. Tiếp cận này có khả năng tạo ra những thay đổi lớn trong cải thiện hiệu năng cho các thuật toán giảm thuộc tính. Sau đây là phần phân tích các nguyên nhân ảnh hưởng tới chi phí tính toán của thuật toán, khả năng phân lớp và số lượng phần tử



reduct thu được từ thuật toán.

1) Thời gian tính toán của thuật toán đề xuất.

Như đã trình bày trong phần đánh giá độ phức tạp, thuật toán đề xuất có thời gian tính toán lý thuyết tốt hơn đáng kể so với các thuật toán sử dụng tiếp cận độ đo hiện nay. Hầu hết các thuật toán giảm thuộc tính truyền thống đều có độ phức tạp tính toán là  $\mathcal{O}(|U|^2|C|^2)$ . Khi  $|U|$  lớn dẫn đến không gian xấp xỉ sẽ rất lớn, chiếm dụng nhiều tài nguyên lưu trữ và tài nguyên tính toán của hệ thống. Khi  $|C|$  lớn, sẽ có nhiều thuộc tính cần phải đánh giá. Ngược lại, độ phức tạp của thuật toán được đề xuất là  $\mathcal{O}(2|U||C|) + \mathcal{O}(|U|^2|H^\beta| + |CH^\beta|\mathbb{T})$  có chi phí tính toán nhỏ hơn đáng kể. Trong đó giai đoạn filter có thời gian chỉ  $\mathcal{O}(2|U||C|)$ , giai đoạn này là yếu tố chính làm giảm mạnh chi phí tính toán của thuật toán. Khi  $|H^\beta|$  nhỏ thời gian phân cụm các thuộc tính sẽ nhanh. Trong đó thời gian phân cụm của  $|H^\beta|$  thuộc tính là  $\mathcal{O}(|U|^2|H^\beta|^2)$ . Hơn nữa, nếu số cụm phân cụm được ít thì thời gian xác định cụm thuộc tính có khả năng phân lớp sẽ nhanh hơn với số lượng cụm lớn. Do đó, khi  $|H^\beta|$  càng nhỏ thì chi phí tính toán của thuật toán càng nhanh và ngược lại.

2) Khả năng phân lớp của thuật toán đề xuất.

Hầu hết các phương pháp giảm thuộc tính truyền thống đều sử dụng độ đo để đánh giá độ quan trọng của thuộc tính cũng như đo lường lượng thông tin bảo toàn của tập thuộc tính rút gọn so với tập thuộc tính gốc. Tuy nhiên, cách tiếp cận độ đo chủ yếu đánh giá độ tương tự giữa các tập dựa trên tổng thành phần mà không xem xét đến nội dung bên trong thành phần đó. Trong khi đó, cách tiếp cận dựa trên cấu trúc tôpô cho phép đánh giá sự tương tự giữa hai tập hợp dựa trên sự tương đồng giữa hai cấu trúc. Tiếp cận đánh giá dựa trên sự tương đồng cấu trúc chặt chẽ hơn so với tiếp cận dựa trên lực lượng của các tập hợp. Hơn nữa, cấu trúc tôpô được dùng là cấu trúc tôpô Hausdorff. Tại sao sử dụng cấu trúc tôpô này vì phương pháp chung để đánh giá độ quan trọng của thuộc tính đó là đánh giá sự phụ thuộc của thuộc tính quyết định. Trong đó, cấu trúc tôpô của thuộc tính quyết định là cấu trúc tôpô Hausdorff nên chương này đề xuất chọn lọc các thuộc tính điều kiện có cấu trúc tôpô Hausdorff.

3) Kích thước rút gọn từ thuật toán đề xuất.

Như đã đề cập bên trên, tiếp cận đánh giá thuộc tính theo cấu trúc tôpô chặt hơn so với các tiếp cận độ đo truyền thống. Theo các kết quả nghiên cứu của Yu và các cộng [38] đã chỉ ra hai phân hoạch khác nhau có thể có cùng cấu trúc tôpô. Do đó, reduct theo tiếp cận tôpô sẽ có số lượng phần tử nhỏ hơn so với reduct theo tiếp cận độ đo truyền thống. Hơn nữa, giai đoạn phân cụm các thuộc tính lại tiếp tục chia nhỏ tập thuộc tính ứng viên thành các nhóm thuộc tính con có cùng cấu trúc phụ thuộc. Đây là nguyên nhân chủ đạo ảnh hưởng tới việc cải thiện số lượng phần tử reduct với các thuật toán tốt nhất hiện nay.

#### 4.5. Kết luận Chương 4

Chương 4, luận án trình bày về phương pháp giảm thuộc tính theo tiếp cận tôpô Hausdorff. Các đóng góp chính của Chương này gồm có:

- Đề xuất cấu trúc tôpô dựa trên quan hệ của các phép toán xấp xỉ trên không gian xấp xỉ mờ ngưỡng  $\beta$ ;
- Đề xuất cấu trúc tôpô Hausdorff dựa trên định nghĩa tính phân biệt được của ma trận quan hệ mờ ngưỡng  $\beta$ ;
- Đề xuất thuật toán tìm reduct dựa trên cấu trúc tôpô Hausdorff và định nghĩa khái niệm đồng cấu trúc phụ thuộc trong không gian tôpô Hausdorff.

Các kết quả thực nghiệm cho thấy thuật toán đề xuất là hoàn toàn vượt trội so với các phương pháp khác cả về chi phí tính toán của thuật toán, khả năng phân lớp và số lượng phần tử của reduct thu được.

## **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 giảm thuộc tính theo tiếp cận lai ghép filter - wrapper, sử dụng độ đo khoảng cách mờ trực cảm (IFD).
2. Xây dựng thuật toán giảm thuộc tính theo tiếp cận filter (F\_IFT) và thuật toán lai ghép filter - wrapper (FW\_IFT), sử dụng cấu trúc tập mờ trực cảm
3. Xây dựng thuật toán giảm thuộc tính theo tiếp cận lai ghép filter - wrapper cụm (CFW), sử dụng cấu trúc tập Hausdorff.

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

- Thuật toán IFD có khả năng cải thiện nhiều khá tốt, tuy nhiên số lượng phần tử và khả năng phân lớp của reduct chưa hiệu quả hơn so với các thuật toán được so sánh.
- Thuật toán F\_IFT có chi phí tính toán hiệu quả và số lượng phần tử reduct thu được tốt nhưng khả năng phân lớp còn hạn chế so với các thuật toán được so sánh.
- Thuật toán FW\_IFT cho reduct có số lượng phần tử và khả năng phân lớp hiệu quả, tuy nhiên chi phí tính toán của thuật toán còn hạn chế so với các thuật toán được so sánh.
- Thuật toán CFW là hoàn toàn vượt trội về chi phí tính toán, số lượng phần tử và khả năng phân lớp của reduct 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 giảm thuộc tính gồm có:

- Phương pháp giảm thuộc tính theo tiếp cận RS 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 giảm 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  $\beta$ .

### **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 giảm thuộc tính trong bảng quyết định không đầy đủ theo tiếp cận mô hình RS mở rộng, tuy nhiên các kết quả nghiên cứu vẫn còn hạn chế về số lượng phần tử và khả năng phân lớp trên các reduct thu được. Do đó, hướng nghiên cứu tương lai của luận án sẽ nhắm tới giảm thuộc tính cho bảng quyết định không đầy đủ thông qua một số các hướng mở rộng cấu trúc tôpô theo tiếp cận RS như sau:

1) Mở rộng cấu trúc tôpô dựa trên không gian xấp xỉ của mô hình RS 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 sự tương quan của các phép toán xấp xỉ của mô hình RS 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 reduct.

## DANH MỤC CÁC CÔNG TRÌNH NGHIÊN CỨU

### A. Các công trình đã công bố

[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. Các công trình đang chờ phản biện

[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.

## TÀI LIỆU THAM KHẢO

### Tiếng Việt:

- [1] H. T. Phương, *Phương pháp gia tăng rút gọn thuộc tính trong bảng quyết định thay đổi theo tiếp cận tập thô mờ*. Luận án Tiến sĩ Khoa học máy tính, Học viện Khoa học và Công nghệ-Viện Hàn lâm Khoa học và Công nghệ Việt Nam, 2021.
- [2] N. V. Thiện, *Rút gọn thuộc tính và trích lọc luật theo tiếp cận tập thô mờ*. Luận án Tiến sĩ Khoa học máy tính, Học viện Khoa học và Công nghệ-Viện Hàn lâm Khoa học và Công nghệ Việt Nam, 2018.
- [3] C. C. Nghĩa, *Rút gọn thuộc tính trong bảng quyết định theo tiếp cận tập thô mờ*. Luận án Tiến sĩ Khoa học máy tính, Học viện Khoa học và Công nghệ-Viện Hàn lâm Khoa học và Công nghệ Việt Nam, 2014.

### Tiếng Anh:

- [4] S. Bashir, I. U. Khattak, A. Khan, F. H. Khan, A. Gani, and M. Shiraz, “A Novel Feature Selection Method for Classification of Medical Data Using Filters, Wrappers, and Embedded Approaches,” *Complexity*, vol. 2022, pp. 1–12, 2022.
- [5] L. Meenachi and S. Ramakrishnan, “Differential evolution and ACO based global optimal feature selection with fuzzy rough set for cancer data classification,” *Soft Computing*, vol. 24, no. 24, pp. 18 463–18 475, 2020.
- [6] S. Ahlawat and R. Rishi, “A Genetic Algorithm Based Feature Selection for Handwritten Digit Recognition,” *Recent Patents on Computer Science*, vol. 12, no. 4, pp. 304–316, 2018.

- [7] H. huang Zhao, H. Liu, H. huang Zhao, and H. Liu, “Multiple classifiers fusion and CNN feature extraction for handwritten digits recognition,” *Granular Computing*, vol. 5, no. 3, pp. 411–418, 2020.
- [8] L. Sun, S. Fu, and F. Wang, “Decision tree SVM model with Fisher feature selection for speech emotion recognition,” *Eurasip Journal on Audio, Speech, and Music Processing*, vol. 2019, no. 1, 2019.
- [9] S. Yildirim, Y. Kaya, and F. Kılıç, “A modified feature selection method based on metaheuristic algorithms for speech emotion recognition,” *Applied Acoustics*, vol. 173, 2021.
- [10] G. Ansari, T. Ahmad, and M. N. Doja, “Spam review classification using ensemble of global and local feature selectors,” *Cybernetics and Information Technologies*, vol. 18, no. 4, pp. 29–42, 2018.
- [11] H. Mohammadzadeh and F. S. Gharehchopogh, “A novel hybrid whale optimization algorithm with flower pollination algorithm for feature selection: Case study Email spam detection,” *Computational Intelligence*, vol. 37, no. 1, pp. 176–209, 2021.
- [12] A. J. Fernández-García, L. Iribarne, A. Corral, J. Criado, and J. Z. Wang, “A recommender system for component-based applications using machine learning techniques,” *Knowledge-Based Systems*, vol. 164, pp. 68–84, 2019.
- [13] B. Saravanan, V. Mohanraj, and J. Senthilkumar, “A fuzzy entropy technique for dimensionality reduction in recommender systems using deep learning,” *Soft Computing*, vol. 23, no. 8, pp. 2575–2583, 2019.
- [14] Z. Pawlak, “Rough sets,” *International Journal of Computer & Information Sciences*, vol. 11, no. 5, pp. 341–356, 1982.
- [15] A. Tan, S. Shi, W. Z. Wu, J. Li, and W. Pedrycz, “Granularity and Entropy of Intuitionistic Fuzzy Information and Their Applications,” *IEEE Transactions on Cybernetics*, vol. 52, no. 1, pp. 192–204, 2022.

- [16] B. Liang, L. Wang, and Y. Liu, "Attribute reduction based on improved information entropy," *Journal of Intelligent and Fuzzy Systems*, vol. 36, no. 1, pp. 709–718, 2019.
- [17] J. He, L. Qu, Z. Wang, Y. Chen, D. Luo, and C. F. Wen, "Attribute reduction in an incomplete categorical decision information system based on fuzzy rough sets," *Artificial Intelligence Review*, vol. 55, no. 7, pp. 5313–5348, 2022.
- [18] Z. Pawlak, S. K. Wong, and W. Ziarko, "Rough sets: probabilistic versus deterministic approach," *International Journal of Man-Machine Studies*, vol. 29, no. 1, pp. 81–95, 1988.
- [19] Z. Pawlak, "Granularity of knowledge, indiscernibility and rough sets," in *1998 IEEE International Conference on Fuzzy Systems Proceedings - IEEE World Congress on Computational Intelligence*, vol. 1, 1998.
- [20] Z. Bashir, M. G. Abbas Malik, S. Asif, and T. Rashid, "The topological properties of intuitionistic fuzzy rough sets," *Journal of Intelligent and Fuzzy Systems*, vol. 38, no. 1, pp. 795–807, 2020.
- [21] D. Dubois and H. Prade, "Rough fuzzy sets and fuzzy rough sets," *International Journal of General Systems*, vol. 17, no. 2-3, pp. 191–209, 1990.
- [22] E. C. Tsang, D. Chen, D. S. Yeung, X. Z. Wang, and J. W. Lee, "Attributes reduction using fuzzy rough sets," *IEEE Transactions on Fuzzy Systems*, vol. 16, no. 5, pp. 1130–1141, 2008.
- [23] X. Yang, H. Chen, T. Li, and C. Luo, "A noise-aware fuzzy rough set approach for feature selection," *Knowledge-Based Systems*, vol. 250, no. 109092, p. 109 092, 2022.
- [24] Z. Qiu and H. Zhao, "A fuzzy rough set approach to hierarchical feature selection based on Hausdorff distance," *Applied Intelligence*, vol. 52, no. 10, pp. 11 089–11 102, 2022.



- [25] R. K. Huda and H. Banka, “Efficient feature selection methods using PSO with fuzzy rough set as fitness function,” *Soft Computing*, vol. 26, no. 5, pp. 2501–2521, 2022.
- [26] P. Liang, D. Lei, K. S. Chin, and J. Hu, “Feature selection based on robust fuzzy rough sets using kernel-based similarity and relative classification uncertainty measures,” *Knowledge-Based Systems*, vol. 255, no. 109795, p. 109 795, 2022.
- [27] J. Ye, J. Zhan, W. Ding, and H. Fujita, “A novel fuzzy rough set model with fuzzy neighborhood operators,” *Information Sciences*, vol. 544, pp. 266–297, 2021.
- [28] A. Kumar and P. S. Sai Prasad, “Incremental fuzzy rough sets based feature subset selection using fuzzy min-max neural network preprocessing,” *International Journal of Approximate Reasoning*, vol. 139, pp. 69–87, 2021.
- [29] S. An, Q. Hu, and C. Wang, “Probability granular distance-based fuzzy rough set model,” *Applied Soft Computing*, vol. 102, 2021.
- [30] Z. Li, Y. Chen, G. Zhang, L. Qu, and N. Xie, “Entropy measurement for a hybrid information system with images: an application in attribute reduction,” *Soft Computing*, vol. 26, no. 21, pp. 11 243–11 263, 2022.
- [31] J. Xu, Y. Wang, H. Mu, and F. Huang, “Feature genes selection based on fuzzy neighborhood conditional entropy,” *Journal of Intelligent and Fuzzy Systems*, vol. 36, no. 1, pp. 117–126, 2019.
- [32] P. Zhang, T. Li, G. Wang, *et al.*, *Multi-source information fusion based on rough set theory: A review*, 2021.
- [33] N. L. Giang, L. H. Son, T. T. Ngan, *et al.*, “Novel Incremental Algorithms for Attribute Reduction from Dynamic Decision Tables Using Hybrid Filter-Wrapper with Fuzzy Partition Distance,” *IEEE Transactions on Fuzzy Systems*, vol. 28, no. 5, pp. 858–873, 2020.

- [34] X. Zhang, B. Zhou, and P. Li, “A general frame for intuitionistic fuzzy rough sets,” *Information Sciences*, vol. 216, pp. 34–49, 2012.
- [35] C. Cornelis, M. De Cock, and E. E. Kerre, “Intuitionistic fuzzy rough sets: At the crossroads of imperfect knowledge,” *Expert Systems*, vol. 20, no. 5, pp. 260–270, 2003.
- [36] A. Tan, W. Z. Wu, Y. Qian, J. Liang, J. Chen, and J. Li, “Intuitionistic Fuzzy Rough Set-Based Granular Structures and Attribute Subset Selection,” *IEEE Transactions on Fuzzy Systems*, vol. 27, no. 3, pp. 527–539, 2019.
- [37] E. F. Lashin and T. Medhat, “Topological reduction of information systems,” *Chaos, Solitons and Fractals*, vol. 25, no. 2, pp. 277–286, 2005.
- [38] H. Yu and W. R. Zhan, “On the topological properties of generalized rough sets,” *Information Sciences*, vol. 263, pp. 141–152, 2014.
- [39] K. Qin and Z. Pei, “On the topological properties of fuzzy rough sets,” *Fuzzy Sets and Systems*, vol. 151, no. 3, pp. 601–613, 2005.
- [40] P. K. Singh and S. Tiwari, “Topological structures in rough set theory: A survey,” *Hacettepe Journal of Mathematics and Statistics*, vol. 49, no. 4, pp. 1270–1294, 2020.
- [41] Q. E. Wu, T. Wang, Y. Huang, and J. S. Li, “Topology theory on rough sets,” *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, vol. 38, no. 1, pp. 68–77, 2008.
- [42] S. E. Han, “Topological properties of locally finite covering rough sets and K-topological rough set structures,” *Soft Computing*, vol. 25, no. 10, 2021.
- [43] M. K. El-Bably, K. K. Fleifel, and O. A. Embaby, “Topological approaches to rough approximations based on closure operators,” *Granular Computing*, vol. 7, no. 1, 2022.

- [44] T. M. Al-shami and I. Alshammari, “Rough sets models inspired by supra-topology structures,” *Artificial Intelligence Review*, 2022.
- [45] C. Y. Wang, “Topological characterizations of generalized fuzzy rough sets,” *Fuzzy Sets and Systems*, vol. 312, 2017.
- [46] H. Wu and G. Liu, “The relationships between topologies and generalized rough sets,” *International Journal of Approximate Reasoning*, vol. 119, pp. 313–324, 2020.
- [47] L. Zhou, W. Z. Wu, and W. X. Zhang, “On intuitionistic fuzzy rough sets and their topological structures,” *International Journal of General Systems*, vol. 38, no. 6, pp. 589–616, 2009.
- [48] Z. Pei, D. Pei, and L. Zheng, “Topology vs generalized rough sets,” *International Journal of Approximate Reasoning*, vol. 52, no. 2, pp. 231–239, 2011.
- [49] Z. Zhao, “On some types of covering rough sets from topological points of view,” *International Journal of Approximate Reasoning*, vol. 68, pp. 1–14, 2016.
- [50] W. Yao and S. E. Han, “A topological approach to rough sets from a granular computing perspective,” *Information Sciences*, vol. 627, 2023.
- [51] T. M. Al-shami, “Topological approach to generate new rough set models,” *Complex and Intelligent Systems*, vol. 8, no. 5, pp. 4101–4113, 2022.
- [52] Y. Xie and X. Gao, “Topological reduction algorithm for relation systems,” *Soft Computing*, vol. 26, no. 22, pp. 11 961–11 971, 2022.
- [53] J. Xu, J. Yang, Y. Ma, K. Qu, and Y. Kang, “Feature selection method for color image steganalysis based on fuzzy neighborhood conditional entropy,” *Applied Intelligence*, vol. 52, no. 8, pp. 9388–9405, 2022.
- [54] Z. Pawlak, “Rough set approach to knowledge-based decision support,” *European Journal of Operational Research*, vol. 99, no. 1, pp. 48–57, 1997.

- [55] L. A. Zadeh, “Fuzzy sets,” *Information and Control*, vol. 8, no. 3, pp. 338–353, 1965.
- [56] Z. Yuan, H. Chen, P. Xie, P. Zhang, J. Liu, and T. Li, “Attribute reduction methods in fuzzy rough set theory: An overview, comparative experiments, and new directions,” *Applied Soft Computing*, vol. 107, no. 107353, p. 107 353, 2021.
- [57] K. T. Atanassov, “Intuitionistic fuzzy sets,” *Fuzzy Sets and Systems*, vol. 20, no. 1, pp. 87–96, 1986.
- [58] S. Mishra and R. Srivastava, “Fuzzy topologies generated by fuzzy relations,” *Soft Computing*, vol. 22, no. 2, pp. 373–385, 2018.
- [59] C. Wang, Y. Huang, M. Shao, and X. Fan, “Fuzzy rough set-based attribute reduction using distance measures,” *Knowledge-Based Systems*, vol. 164, pp. 205–212, 2019.
- [60] D. Yu, Q. Hu, and C. Wu, “Uncertainty measures for fuzzy relations and their applications,” *Applied Soft Computing Journal*, vol. 7, no. 3, 2007.
- [61] “Fuzzy probabilistic approximation spaces and their information measures,” *IEEE Transactions on Fuzzy Systems*, vol. 14, no. 2, 2006.
- [62] Q. Hu, L. Zhang, D. Chen, W. Pedrycz, and D. Yu, “Gaussian kernel based fuzzy rough sets: Model, uncertainty measures and applications,” *International Journal of Approximate Reasoning*, vol. 51, no. 4, 2010.
- [63] J. W. Grzymala-Busse, “On the unknown attribute values in learning from examples,” in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 542 LNAI Part F2, 1991.
- [64] C. Cornelis, R. Jensen, and Hurtado, “Attribute selection with fuzzy decision reducts,” *Information Sciences*, vol. 180, no. 2, 2010.

- [65] Q. Hu, D. Yu, W. Pedrycz, and D. Chen, “Kernelized fuzzy rough sets and their applications,” *IEEE Transactions on Knowledge and Data Engineering*, vol. 23, no. 11, 2011.
- [66] C. Wang, Y. Wang, M. Shao, Y. Qian, and D. Chen, “Fuzzy Rough Attribute Reduction for Categorical Data,” *IEEE Transactions on Fuzzy Systems*, vol. 28, no. 5, 2020.
- [67] A. Skowron and C. Rauszer, “The Discernibility Matrices and Functions in Information Systems,” in *Intelligent Decision Support*, 1992.
- [68] R. Jensen and Q. Shen, “New approaches to fuzzy-rough feature selection,” *IEEE Transactions on Fuzzy Systems*, vol. 17, no. 4, 2009.
- [69] C. Degang and Z. Suyun, “Local reduction of decision system with fuzzy rough sets,” *Fuzzy Sets and Systems*, vol. 161, no. 13, 2010.
- [70] D. Chen and Y. Yang, “Attribute reduction for heterogeneous data based on the combination of classical and fuzzy rough set models,” *IEEE Transactions on Fuzzy Systems*, vol. 22, no. 5, 2014.
- [71] Y. Yang, D. Chen, H. Wang, E. C. Tsang, and D. Zhang, “Fuzzy rough set based incremental attribute reduction from dynamic data with sample arriving,” *Fuzzy Sets and Systems*, vol. 312, 2017.
- [72] J. Dai, H. Hu, W. Z. Wu, Y. Qian, and D. Huang, “Maximal-discernibility-pair-based approach to attribute reduction in fuzzy rough sets,” *IEEE Transactions on Fuzzy Systems*, vol. 26, no. 4, pp. 2174–2187, 2018.
- [73] Y. Yang, D. Chen, H. Wang, and X. Wang, “Incremental Perspective for Feature Selection Based on Fuzzy Rough Sets,” *IEEE Transactions on Fuzzy Systems*, vol. 26, no. 3, 2018.
- [74] Y. Liu, L. Zheng, Y. Xiu, *et al.*, “Discernibility matrix based incremental feature selection on fused decision tables,” *International Journal of Approximate Reasoning*, vol. 118, pp. 1–26, 2020.

- [75] Y. Li, M. Cai, J. Zhou, and Q. Li, “Accelerated multi-granularity reduction based on neighborhood rough sets,” *Applied Intelligence*, vol. 52, no. 15, pp. 17 636–17 651, 2022.
- [76] C. Wang, Q. He, M. Shao, and Q. Hu, “Feature selection based on maximal neighborhood discernibility,” *International Journal of Machine Learning and Cybernetics*, vol. 9, no. 11, 2018.
- [77] Q. Hu, D. Yu, and Z. Xie, “Neighborhood classifiers,” *Expert Systems with Applications*, vol. 34, no. 2, pp. 640–649, 2008.
- [78] X. Yang, H. Chen, T. Li, J. Wan, and B. Sang, “Neighborhood rough sets with distance metric learning for feature selection[Formula presented],” *Knowledge-Based Systems*, vol. 224, no. 107076, p. 107 076, 2021.
- [79] J. Liu, Y. Lin, J. Du, H. Zhang, Z. Chen, and J. Zhang, “ASFS: A novel streaming feature selection for multi-label data based on neighborhood rough set,” *Applied Intelligence*, 2022.
- [80] W. Shu, W. Qian, and Y. Xie, “Incremental feature selection for dynamic hybrid data using neighborhood rough set,” *Knowledge-Based Systems*, vol. 194, no. 105516, p. 105 516, 2020.
- [81] J. Wan, H. Chen, Z. Yuan, T. Li, X. Yang, and B. B. Sang, “A novel hybrid feature selection method considering feature interaction in neighborhood rough set[Formula presented],” *Knowledge-Based Systems*, vol. 227, no. 107167, p. 107 167, 2021.
- [82] D. Liu and J. Li, “Safety monitoring data classification method based on wireless rough network of neighborhood rough sets,” *Safety Science*, vol. 118, pp. 282–296, 2019.
- [83] R. Benouini, I. Batioua, S. Ezghari, K. Zenkouar, and A. Zahi, “Fast feature selection algorithm for neighborhood rough set model based on Bucket and Trie structures,” *Granular Computing*, vol. 5, no. 3, pp. 329–347, 2020.

- [84] M. Hu, E. C. Tsang, Y. Guo, D. Chen, and W. Xu, “A novel approach to attribute reduction based on weighted neighborhood rough sets,” *Knowledge-Based Systems*, vol. 220, no. 106908, 2021.
- [85] S. Li, K. Zhang, Y. Li, S. Wang, and S. Zhang, “Online streaming feature selection based on neighborhood rough set,” *Applied Soft Computing*, vol. 113, no. 108025, p. 108 025, 2021.
- [86] Q. Hu, D. Yu, J. Liu, and C. Wu, “Neighborhood rough set based heterogeneous feature subset selection,” *Information Sciences*, vol. 178, no. 18, 2008.
- [87] C. Wang, M. Shao, Q. He, Y. Qian, and Y. Qi, “Feature subset selection based on fuzzy neighborhood rough sets,” *Knowledge-Based Systems*, vol. 111, pp. 173–179, 2016.
- [88] B. Sang, H. Chen, L. Yang, T. Li, and W. Xu, “Incremental Feature Selection Using a Conditional Entropy Based on Fuzzy Dominance Neighborhood Rough Sets,” *IEEE Transactions on Fuzzy Systems*, vol. 30, no. 6, pp. 1683–1697, 2022.
- [89] P. Chen, M. Lin, and J. Liu, “Multi-Label Attribute Reduction Based on Variable Precision Fuzzy Neighborhood Rough Set,” *IEEE Access*, vol. 8, pp. 133 565–133 576, 2020.
- [90] K. Zhang, J. Zhan, and W. Z. Wu, “On Multicriteria Decision-Making Method Based on a Fuzzy Rough Set Model with Fuzzy  $\alpha$ -Neighborhoods,” *IEEE Transactions on Fuzzy Systems*, vol. 29, no. 9, pp. 2491–2505, 2021.
- [91] J. Xu, K. Shen, and L. Sun, “Multi-label feature selection based on fuzzy neighborhood rough sets,” *Complex and Intelligent Systems*, vol. 8, no. 3, pp. 2105–2129, 2022.
- [92] B. Sang, H. Chen, L. Yang, T. Li, W. Xu, and C. Luo, “Feature selection for dynamic interval-valued ordered data based on fuzzy dominance

- neighborhood rough set,” *Knowledge-Based Systems*, vol. 227, no. 107223, p. 107 223, 2021.
- [93] S. Shreevastava, A. K. Tiwari, and T. Som, “Intuitionistic fuzzy neighborhood rough set model for feature selection,” *International Journal of Fuzzy System Applications*, vol. 7, no. 2, pp. 75–84, 2018.
- [94] J. Xie, B. Q. Hu, and H. Jiang, “A novel method to attribute reduction based on weighted neighborhood probabilistic rough sets,” *International Journal of Approximate Reasoning*, vol. 144, pp. 1–17, 2022.
- [95] Q. Hu, D. Yu, and Z. Xie, “Information-preserving hybrid data reduction based on fuzzy-rough techniques,” *Pattern Recognition Letters*, vol. 27, no. 5, 2006.
- [96] N. N. Thuy and S. Wongthanavas, “Hybrid filter–wrapper attribute selection with alpha-level fuzzy rough sets,” *Expert Systems with Applications*, vol. 193, no. 116428, p. 116 428, 2022.
- [97] Y. Lin, Y. Li, C. Wang, and J. Chen, “Attribute reduction for multi-label learning with fuzzy rough set,” *Knowledge-Based Systems*, vol. 152, pp. 51–61, 2018.
- [98] X. Zhang, J. Yang, and L. Tang, “Three-way class-specific attribute reducts from the information viewpoint,” *Information Sciences*, vol. 507, pp. 92–126, 2020.
- [99] Y. Fang and F. Min, “Cost-sensitive approximate attribute reduction with three-way decisions,” *International Journal of Approximate Reasoning*, vol. 104, pp. 112–139, 2019.
- [100] A. M. Ghroutkhar and H. M. Nehi, “Fuzzy–rough set models and fuzzy-rough data reduction,” *Croatian Operational Research Review*, vol. 19, no. 1, pp. 67–80, 2020.



- [101] X. Zhang, C. Mei, D. Chen, and Y. Yang, “A fuzzy rough set-based feature selection method using representative instances,” *Knowledge-Based Systems*, vol. 151, pp. 216–229, 2018.
- [102] P. Jain, A. K. Tiwari, and T. Som, “A fitting model based intuitionistic fuzzy rough feature selection,” *Engineering Applications of Artificial Intelligence*, vol. 89, no. 103421, p. 103 421, 2020.
- [103] A. K. Tiwari, S. Shreevastava, K. Subbiah, and T. Som, “An intuitionistic fuzzy-rough set model and its application to feature selection,” *Journal of Intelligent and Fuzzy Systems*, vol. 36, no. 5, pp. 4969–4979, 2019.
- [104] B. B. Sang, X. Y. Zhang, and W. H. Xu, “Attribute reduction of relative knowledge granularity in intuitionistic fuzzy ordered decision table,” *Filomat*, vol. 32, no. 5, pp. 1727–1736, 2018.
- [105] Z. Zhang, “Attributes reduction based on intuitionistic fuzzy rough sets,” *Journal of Intelligent and Fuzzy Systems*, vol. 30, no. 2, pp. 1127–1137, 2016.
- [106] Z. Chuanchao, “Generalized dynamic attribute reduction based on similarity relation of intuitionistic fuzzy rough set,” *Journal of Intelligent and Fuzzy Systems*, vol. 39, no. 5, pp. 7107–7122, 2020.
- [107] Z. Zhang and J. Tian, “On attribute reduction with intuitionistic fuzzy rough sets,” *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, vol. 20, no. 1, pp. 59–76, 2012.
- [108] Z. ao Xue, M. meng Jing, Y. xiang Li, *et al.*, “Variable precision multi-granulation covering rough intuitionistic fuzzy sets,” *Granular Computing*, 2022.
- [109] M. Rahimi, P. Kumar, B. Moomivand, and G. Yari, “An intuitionistic fuzzy entropy approach for supplier selection,” *Complex and Intelligent Systems*, vol. 7, no. 4, pp. 1869–1876, 2021.

- [110] M. B. Revanasiddappa and B. S. Harish, “A New Feature Selection Method based on Intuitionistic Fuzzy Entropy to Categorize Text Documents,” *International Journal of Interactive Multimedia and Artificial Intelligence*, vol. 5, no. 3, p. 106, 2018.
- [111] P. Zhang, T. Li, Z. Yuan, C. Luo, K. Liu, and X. Yang, “Heterogeneous Feature Selection Based on Neighborhood Combination Entropy,” *IEEE Transactions on Neural Networks and Learning Systems*, vol. PP, pp. 1–14, 2022.
- [112] Z. Yuan, H. Chen, X. Yang, T. Li, and K. Liu, “Fuzzy complementary entropy using hybrid-kernel function and its unsupervised attribute reduction,” *Knowledge-Based Systems*, vol. 231, no. 107398, p. 107 398, 2021.
- [113] T. T. Nguyen, N. L. Giang, D. T. Tran, *et al.*, “A novel filter-wrapper algorithm on intuitionistic fuzzy set for attribute reduction from decision tables,” *International Journal of Data Warehousing and Mining*, vol. 17, no. 4, pp. 67–100, 2021.
- [114] H. I. Mustafa and O. A. Tantawy, “A new approach of attribute reduction of rough sets based on soft metric,” *Journal of Intelligent and Fuzzy Systems*, vol. 39, no. 3, pp. 4473–4489, 2020.
- [115] B. Huang, H. X. Li, G. F. Feng, and Y. L. Zhuang, “Distance-based Information Granularity and Hierarchical Structure for an Intuitionistic Fuzzy Granular Space,” *Fuzzy Information and Engineering*, vol. 8, no. 2, pp. 147–168, 2016.
- [116] W. Yang, K. Gong, S. Tu, and C. Zeng, “Distance measurement on intuitionistic fuzzy granular structure sets,” *Microsyst. Technol.*, vol. 27, no. 4, pp. 1633–1639, 2019.
- [117] Z. Li, T. Xie, and Q. Li, “Topological structure of generalized rough sets,” *Computers and Mathematics with Applications*, vol. 63, no. 6, pp. 1066–1071, 2012.

- [118] Z. M. Ma and B. Q. Hu, “Topological and lattice structures of L-fuzzy rough sets determined by lower and upper sets,” *Information Sciences*, vol. 218, pp. 194–204, 2013.
- [119] Z. Zhang, “Generalized intuitionistic fuzzy rough sets based on intuitionistic fuzzy coverings,” *Information Sciences*, vol. 198, pp. 186–206, 2012.
- [120] L. Yang and L. Xu, “Topological properties of generalized approximation spaces,” *Information Sciences*, vol. 181, no. 17, 2011.
- [121] S. M. Yun and S. J. Lee, “New approach to intuitionistic fuzzy rough sets,” *International Journal of Fuzzy Logic and Intelligent Systems*, vol. 20, no. 2, pp. 129–137, 2020.
- [122] S. M. Yun, Y. S. Eom, and S. J. Lee, “Topology of the Redefined Intuitionistic Fuzzy Rough Sets,” *International Journal of Fuzzy Logic and Intelligent Systems*, vol. 21, no. 4, pp. 369–377, 2021.
- [123] W. Zhu, “Topological approaches to covering rough sets,” *Information Sciences*, vol. 177, no. 6, pp. 1499–1508, 2007.
- [124] L. Su and W. Zhu, “Dependence space of topology and its application to attribute reduction,” *International Journal of Machine Learning and Cybernetics*, vol. 9, no. 4, pp. 691–698, 2018.
- [125] R. A. Ibrahim, M. Abd Elaziz, D. Oliva, and S. Lu, “An improved runner-root algorithm for solving feature selection problems based on rough sets and neighborhood rough sets,” *Applied Soft Computing*, vol. 97, no. 105517, p. 105 517, 2020.
- [126] L. Kong, W. Qu, J. Yu, *et al.*, “Distributed Feature Selection for Big Data Using Fuzzy Rough Sets,” *IEEE Transactions on Fuzzy Systems*, vol. 28, no. 5, pp. 846–857, 2020.

- [127] M. Prasad, S. Tripathi, and K. Dahal, “An efficient feature selection based Bayesian and Rough set approach for intrusion detection,” *Applied Soft Computing Journal*, vol. 87, no. 105980, p. 105 980, 2020.
- [128] Q. Zhang and C. Hu, “Reduction Algorithm of Interval-Valued Intuitionistic Fuzzy Probability Rough Set Under Dominant Relation,” *International Journal of Pattern Recognition and Artificial Intelligence*, 2022.
- [129] J. Yang, Q. Zhang, and Q. Xie, “Attribute reduction based on misclassification cost in variable precision rough set model,” *Journal of Intelligent and Fuzzy Systems*, vol. 37, no. 4, pp. 5129–5142, 2019.
- [130] A. J. Reddy and B. K. Tripathy, “Topological properties of multigranular rough sets on intuitionistic fuzzy approximation spaces,” *International Journal of Intelligent Enterprise*, vol. 8, no. 1, pp. 1–17, 2021.
- [131] A. Tan, W. Z. Wu, J. Li, and T. Li, “Reduction foundation with multigranulation rough sets using discernibility,” *Artificial Intelligence Review*, vol. 53, no. 4, pp. 2425–2452, 2020.
- [132] Q. Hu, Z. Xie, and D. Yu, “Hybrid attribute reduction based on a novel fuzzy-rough model and information granulation,” *Pattern Recognition*, vol. 40, no. 12, 2007.