Về một thuật toán lai ghép lọc-đóng gói lựa chọn thuộc tính trên hệ thông tin quyết định theo tiếp cận tập thô lân cận mờ

Phạm Việt Anh[1,2], Đặng Trọng Hợp[3], Ngô Quang Huy[4], Lê Xuân Hùng[3], Trần Phi Lực[5], Đỗ Đình Lực[6]
[1] Học viện Khoa học và công nghệ, Viện Hàn lâm Khoa học và công nghệ Việt Nam, Hà Nội, Việt Nam
[2] Viện Công nghệ HaUI, Trường Đại học Công nghiệp Hà Nội, Hà Nội, Việt Nam
[3] Khoa Công nghệ Thông tin, Trường Đại học Công nghiệp Hà Nội, Hà Nội, Việt Nam
[4] Trung tâm Truyền thông và Quan hệ công chúng, Trường Đại học Công nghiệp Hà Nội, Hà Nội, Việt Nam
[5] Phòng Khoa học Công nghệ, Trường Đại học Công nghiệp Hà Nội, Hà Nội, Việt Nam
[6] Đại học Công nghệ thông tin và Truyền thông, Đại học Thái Nguyên, Thái Nguyên, Việt Nam

Tác giả liên hệ: Phạm Việt Anh, anhpv@haui.edu.vn, vietanh.hn.4078@gmail.com
Ngày nhận bài: 01/08/2024, ngày sửa chữa: 05/09/2024, ngày duyệt đăng: 14/10/2024
Định danh DOI: 10.32913/mic-ict-research-vn.v2024.n2.1296

Tóm tắt: Lý thuyết tập thô lân cận là công cụ hữu hiệu giúp giải quyết rất tốt bài toán lựa chọn thuộc tính trên các hệ thông tin có miền giá trị số liên tục. Tuy nhiên, mô hình này còn gặp nhiều hạn chế trong việc mô tả mối quan hệ từ các đối tượng trong cùng một lớp lân cận. Điều này dẫn tới sự kém hiệu quả về một số độ đo khi chỉ xét tới các lớp lân cận nằm hoàn toàn trong một phân lớp. Rõ ràng, việc bỏ qua những lớp lân cận còn lại sẽ có những ảnh hưởng rất lớn tới hiệu quả, mặc dù chúng vẫn tồn tại những ý nghĩa về mặt thông tin cũng như đóp góp giá trị vào các độ đo. Để giải quyết những khó khăn này, nghiên cứu ban đầu sẽ trình bày về một mở rộng mới được gọi là tập thô lân cận mờ. Mô hình này hiệu quả trong việc giảm thiểu nhiễu và rút gọn không gian tính toán. Bên cạnh đó, mô hình này cũng mô tả đầy đủ đặc trưng mức độ thuộc của các đối tượng trong một lớp lân cận. Dựa trên những ưu điểm của mô hình, nghiên cứu đề xuất một độ đo mới nhằm đánh giá khả năng phân lớp của các hạt thông tin lân cận mờ. Qua đó, những phân tích và chứng minh tính hiệu quả từ độ đo trong việc đánh giá ý nghĩa của các thuộc tính trong hệ thông tin nhất quán và không nhất quán sẽ được làm rõ. Tiếp theo, chúng tôi sẽ định nghĩa lại một rút gọn mới để thiết kế một thuật toán theo hướng tiếp cận lọc-đóng gói nhằm trích chọn một tập con thuộc tính tối ưu trên hệ thông tin quyết định. Một số kết quả thực nghiệm đã chứng minh hiệu quả của thuật toán đề xuất so với một số thuật toán theo tiếp cận tập thô mờ.

Từ khóa: *lựa chọn thuộc tính, tập thô lân cận mờ, phủ lân cận mờ, hệ thông tin quyết định.*

Title: **A Novel Hybrid Filter-Wrapper Algorithm for Feature Selection in Decision Information Systems Based on Fuzzy Neighborhood Rough Sets**

Abstract: Neighborhood rough set theory is an effective tool for solving the problem of attribute selection in information systems with continuous numerical value domains. However, this model still has many limitations in describing the relationships between objects within the same neighborhood class. This leads to inefficiency in measures when only considering neighborhood classes entirely within a decision class. Clearly, ignoring the remaining neighborhood classes will significantly affect efficiency, even though they still hold informational value and contribute to the measures. To address these challenges, the initial research will present a new extension called the fuzzy neighborhood rough set. This model is effective in reducing noise and simplifying the computational space. Additionally, this model fully describes the characteristics of the membership degree of objects in a neighborhood class. Based on these advantages, the research also proposes a new measure to evaluate the classification capability of fuzzy neighborhood information granules. Accordingly, the analysis and proof of the effectiveness of the measures in evaluating the significance of attributes in both consistent and inconsistent information systems will be clarified. Next, we will redefine a new reduct to design an algorithm following the filter-wrapper approach for selecting an optimal subset of attributes in decision information systems. Several experimental results demonstrated the effectiveness of the proposed algorithm compared to other algorithms based on the fuzzy rough set approach.

Keywords: *attribute selection, fuzzy neighborhood rough sets, fuzzy neighborhood cover, decision information system.*

I. GIỚI THIỆU

Trong giai đoạn tiền xử lý dữ liệu, lựa chọn thuộc tính được coi là vấn đề quan trọng nhằm cải thiện hiệu quả trên các mô hình học máy. Với mục tiêu loại bỏ các thuộc tính dư thừa trong dữ liệu, lựa chọn thuộc tính là một hướng nghiên cứu không thể thiếu trong xu thế của dữ liệu lớn. Lý thuyết tập thô [1] được xem là công cụ đầu tiên đánh dấu sự khởi nguồn cho các phương pháp về lựa chọn thuộc tính. Dựa trên mô hình này, một số thuật toán về lựa chọn thuộc tính sau đó đã được phát triển. Cụ thể, Giang trong [2] đã xây dựng một độ đo entropy cải tiến để định nghĩa một rút gọn mới tương đương với rút gọn của Pawlak làm cơ sở cho việc thiết kế một thuật toán lựa chọn thuộc tính trên hệ thông tin không nhất quán. Tiếp đó, trên hệ thông tin không đầy đủ, Giang cùng các cộng sự [3] mở rộng độ đo entropy Liang và đề xuất thuật toán heuristic để tìm kiếm một rút gọn của bảng. Bên cạnh đó, Anh cùng các cộng sự [4] đã chứng minh các rút gọn trên một hệ thông tin không đầy đủ, không nhất quán là tương đương với một hệ Spener và đề xuất một thuật toán tìm toàn bộ các rút gọn. Có thể thấy, mô hình tập thô là công cụ hữu hiệu cho bài toán lựa chọn thuộc tính với rất nhiều phương pháp đã được phát triển. Tuy nhiên, mô hình này gặp nhiều hạn chế khi sử dụng một quan hệ tương đương chặt chẽ. Do đó, các mở rộng của nó sau đó đã trở thành một hướng nghiên cứu rất sôi động.

Dựa trên sự kết hợp giữa lý thuyết tập thô [1] và tập mờ [5], Dubois cùng các cộng sự đã đề xuất lý thuyết tập thô mờ để giải quyết bài toán lựa chọn thuộc tính trên các hệ thông tin có miền giá trị số liên tục [6]. Một số nghiên cứu sau đó đã được công bố trên không gian xấp xỉ mờ và mang tới những kết quả ấn tượng. Điển hình là các phương pháp sử dùng hàm thuộc mờ [7], miền dương mờ [8, 9], entropy thông tin mờ [10, 11] và khoảng cách mờ [12–14]. Trong [15], các tác giả đã đánh giá cấu trúc hạt của các xấp xỉ thô mờ dưới và trên để xây dựng ma trận phân biệt làm cơ sở cho việc đề xuất thuật toán lựa chọn thuộc tính. Với ưu điểm mô tả rõ ràng mối quan hệ giữa các đối tượng trong các lớp tương đương mờ, các độ đo đánh giá thuộc tính trên không gian tập thô mờ đã chứng minh được khả năng vượt trội so với các phương pháp trên không gian tập thô. Tuy nhiên, các tác giả trong [16, 17] đã chỉ ra rằng với trường hợp dữ liệu nhiễu thì một số phương pháp theo cách tiếp cận này vẫn còn gặp nhiều khó khăn. Một trong số đó đến từ khả năng hạn chế tầm ảnh hưởng từ các đối tượng nhiễu. Cụ thể, trong mô hình tập thô mờ, các lớp tương đương mờ được tạo ra bởi quan hệ tương đương mờ vẫn luôn tồn tại nhiều đối tượng mang giá trị hàm thuộc nhỏ. Về mặt trực cảm, những đối tượng này có thể được tạo ra bởi nhiễu [18] và can thiệp vào quá trình tính toán từ các độ đo đánh giá thuộc tính. Nói cách khác, các độ đo phải chịu sự ảnh hưởng từ các thông tin được tạo ra bởi các đối tượng nhiễu. Do đó, các rút gọn tìm được đạt hiệu quả không cao trong việc cải thiện độ chính xác trên các mô hình phân lớp.

Mô hình tập thô lân cận từ lâu đã trở thành công cụ nhằm giải quyết bài toán lựa chọn thuộc tính trên các hệ thông tin chứa miền giá trị số liên tục. Ưu điểm của mô hình này là sử dụng các quan hệ lân cận được mở rộng từ quan hệ không phân biệt được trong lý thuyết tập thô. Theo đó, mỗi đối tượng trong tập vũ trụ được biểu diễn bởi một lớp lân cận chứa các đối tượng nằm trong một bán kính λ. Do đó, mô hình này có thể loại bỏ các đối tượng có sự phân bố khác biệt so với phần lớn các đối tượng còn lại trong không gian vũ trụ, nghĩa là hạn chế tầm ảnh hưởng của một số đối tượng nhiễu. Tuy nhiên, hạn chế của mô hình này là khả năng mô tả mối quan hệ của các đối tượng trong lớp lân cận còn đơn giản. Ngoài ra, trong mô hình này, một số độ đo chỉ tính toán trên các lớp lân cận chứa toàn bộ các đối tượng thuộc về một phân lớp, nghĩa là nếu một lân cận chỉ chứa một vài đối tượng không thuộc lớp quyết định thì lân cận đó sẽ không được xét. Điều này dẫn tới các độ đo đánh giá thuộc tính đạt hiệu quả chưa tốt khi tìm kiếm một rút gọn tối ưu.

Để giải quyết vấn đề nêu trên, trong nghiên cứu này, chúng tôi ban đầu trình bày về một mô hình dựa trên sự kết hợp giữa mô hình tập thô lân cận và mô hình tập thô mờ, được gọi là mô hình tập thô lân cận mờ [19] nhằm giải quyết tốt bài toán lựa chọn thuộc tính trên các hệ thông tin nhiễu. Tiếp theo, chúng tôi đề xuất một độ đo mới được gọi là độ chắc chắn trong việc phân lớp của một tập đối tượng. Chúng tôi sử dụng độ đo này để đánh giá hiệu quả phân lớp của các hạt thông tin lân cận mờ từ một tập con thuộc tính điều kiện. Theo đó, một số tính chất quan trọng từ độ đo cũng được chúng tôi làm rõ để chứng minh được tính hiệu quả trong việc áp dụng trên cả hệ thông tin quyết định nhất quán và không nhất quán. Dựa trên độ đo này, chúng tôi định nghĩa một rút gọn mới là một tập con thuộc tính điều kiện tối ưu bảo toàn được các thông tin của bảng dữ liệu. Cuối cùng, chúng tôi thiết kế một thuật toán lựa chọn thuộc tính theo hướng tiếp cận lai ghép lọc-đóng gói để trích xuất ra một rút gọn tối ưu. Các đánh giá về những ưu điểm và hạn chế của thuật toán sau đó cũng được chúng tôi trình bày chi tiết. Cấu trúc của bài báo này gồm sáu phần chính, trong đó, Phần II sẽ trình bày một số khái niệm cơ bản liên quan tới hệ thông tin quyết định và mô hình tập thô lân cận. Phần III trình bày về mô hình tập thô lân cận mờ và định nghĩa một rút gọn mới làm cơ sở cho việc đề xuất một thuật toán lai ghép lựa chọn thuộc tính trong Phần IV. Phần V sẽ trình bày một số thực nghiệm để chứng minh tính hiệu quả của thuật toán đề xuất và Phần VI sẽ là một số kết luận từ nghiên cứu.

II. CÁC KHÁI NIỆM LIÊN QUAN

1. Hệ thông tin quyết định

Hệ thông tin quyết định được biểu diễn bởi một cặp $\mathcal{IS} = (X, F \cup D)$, trong đó X là một tập hữu hạn và khác rỗng các đối tượng, $F \cup D$ là tập hữu hạn khác rỗng các thuộc tính thỏa mãn $F \cap D = \emptyset$. Khi đó, cho $x \in X$ và $f \in F \cup D$, giá trị của thuộc tính f với đối tượng x được viết là $f(x)$. Chúng tôi gọi F là tập các thuộc tính điều kiện và D là tập các thuộc tính quyết định.

Ví dụ 1: Cho một hệ thông tin quyết định $\mathcal{IS} = (X, F \cup D)$ với $X = \{x_1, x_2, \ldots, x_5\}$ và $F = \{f_1, f_2, \ldots, f_5\}$ được biểu diễn trong Bảng I.

Bảng I
HỆ THÔNG TIN QUYẾT ĐỊNH

X	f_1	f_2	f_3	f_4	f_5	D
x_1	0.35	0.40	0.70	0.90	0.85	Type A
x_2	0.20	0.25	0.60	0.80	0.75	Type C
x_3	0.95	0.50	0.40	0.25	0.70	Type B
x_4	0.90	0.00	0.45	1.00	0.90	Type A
x_5	0.15	1.00	0.50	0.15	0.60	Type B

2. Mô hình tập thô lân cận

Cho hệ thông tin $\mathcal{IS} = (X, F \cup D)$, một tập con thuộc tính $B \subseteq F$ và hai đối tượng $x, y \in X$, khoảng cách giữa hai đối tượng x và y theo tập thuộc tính B ký hiệu là $\delta_B(x, y)$, được xác định như sau:

$$\delta_B(x, y) = \sqrt[p]{\sum_{f \in B} |f(x) - f(y)|^p} \tag{1}$$

trong đó, $\delta_B(x, y)$ được gọi là khoảng cách Manhattan nếu $p = 1$, khoảng cách Euclid nếu $p = 2$ và khoảng cách Chebyshev nếu $p = \infty$. Trong bài báo này, chúng tôi sử dụng khoảng cách Chebyshev để xác định khoảng cách giữa hai đối tượng x và y.

Giả sử rằng λ là một bán kính lân cận có giá trị nằm trong đoạn $[0, 1]$ và δ là một hàm khoảng cách cho trước, khi đó một quan hệ nhị phân R_B^λ sẽ trở thành một quan hệ lân cận trên X từ tập thuộc tính B và có dạng như sau [20]:

$$R_B^\lambda = \left\{(x, y) \in X \times X \,\middle|\, \forall f \in B, \delta_{\{f\}}(x, y) \leq \lambda\right\} \tag{2}$$

Dựa trên quan hệ lân cận, lớp lân cận của đối tượng $x \in X$ trên tập con thuộc tính B được biểu diễn như công thức dưới đây [20]:

$$[x]_B^\lambda = \left\{y \in X \,\middle|\, (x, y) \in R_B^\lambda\right\} \tag{3}$$

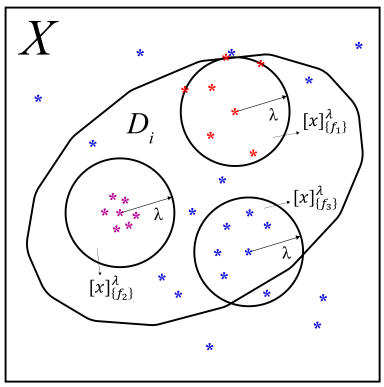
Rõ ràng, lớp lân cận của một đối tượng x bất kỳ trong X là một tập rõ thỏa mãn $[x]_B^\lambda \subseteq X$. Nếu xét toàn bộ các đối tượng trong không gian X, chúng ta sẽ thu được một họ chứa các lớp lân cận, ký hiệu là $X/R_B^\lambda = \left\{[x]_B^\lambda \,|x \in X\right\}$ và được gọi là một phủ lân cận của tập thuộc tính B.

Định nghĩa 1: [20] Cho một hệ thông tin quyết định $\mathcal{IS} = (X, F \cup D)$, một phân hoạch từ tập thuộc tính quyết định $U/D = \left\{D_1, D_2, \ldots, D_q\right\}$ và một quan hệ lân cận X/R_B^λ được tạo ra bởi tập con thuộc tính điều kiện B và bán kính lân cận λ. Khi đó, các xấp xỉ lân cận của một lớp quyết định $D_i \in X/D$ được định nghĩa như sau:

$$\underline{N_B}(D_i) = \left\{x \,\middle|\, [x]_B^\lambda \subseteq D_i, x \in X\right\} \tag{4}$$

$$\overline{N_B}(D_i) = \left\{x \,\middle|\, [x]_B^\lambda \cap D_i \neq \emptyset, x \in X\right\} \tag{5}$$

Trong công thức trên, $\underline{N_B}(D_i)$ và $\overline{N_B}(D_i)$ lần lượt là các xấp xỉ lân cận dưới và trên của lớp quyết định D_i. Rõ ràng, xấp xỉ lân cận dưới rất nhạy cảm với nhiễu do phải chứa toàn bộ các đối tượng nằm trong D_i. Ngoài ra, để thấy được những hạn chế của mô hình, chúng tôi sẽ trình bày một số trường hợp được mô tả trong Hình 1.



Hình 1. Không gian tập thô lân cận.

Hình 1 mô tả một lớp quyết định $D_i \in X \backslash D$ và ba lớp lân cận của đối tượng $x \in X$ với ba thuộc tính điều kiện f_1, f_2 và f_3. Rõ ràng, trên lớp lân cận của đối tượng x với hai thuộc tính f_1 và f_2 thì số lượng đối tượng thuộc D_i là như nhau. Dó đó, trên một số độ đo như hàm phụ thuộc hay entropy quyết định thì ý nghĩa của hai thuộc tính là như nhau vì chúng chỉ xét tới số lượng các đối tượng nằm trong lớp lân cận. Tuy nhiên, có thể thấy rằng lân cận $[x]_{\{f_1\}}^\lambda$ chứa các đối tượng có sự phân bố khác biệt so với $[x]_{\{f_2\}}^\lambda$, đặc biệt bao gồm các đối tượng nằm ở biên. Điều này chỉ ra rằng, nếu thay đổi một ngưỡng bán kính nhỏ, rõ ràng lân cận $[x]_{\{f_1\}}^\lambda$ sẽ hoàn toàn không được xét. Điển hình như lớp lân cận $[x]_{\{f_3\}}^\lambda$, mặc dù chỉ chứa một đối tượng nằm ngoài vùng D_i mà toàn bộ các đối tượng trong đó sẽ coi như không có ý nghĩa. Do đó, việc đánh giá một mức độ thuộc của các đối tượng trong một lớp lân cận là rất quan trọng. Từ đây, khái niệm tập thô lân cận mờ được hình thành.

III. TẬP THÔ LÂN CẬN MỜ

1. Lân cận mờ và phủ lân cận mờ

Trong phần này, chúng tôi sẽ trình bày một mô hình mới dựa trên sự kết hợp giữa mô hình tập thô lân cận đã được trình bày trong phần trên của bài báo và mô hình tập mờ. Mô hình này hứa hẹn sẽ mang tới những hiệu quả trong vấn đề lựa chọn các thuộc tính quan trọng trên các hệ thông tin có tính không chắc chắn.

Định nghĩa 2: [19] Cho một hệ thông tin quyết định $\mathcal{IS} = (X, F \cup D)$, một tập con thuộc tính $B \subseteq F$ và một bán kính lân cận λ thỏa mãn $0 \leq \lambda \leq 1$ xác định một quan hệ R_B^λ. Khi đó, hạt thông tin lân cận mờ của đối tượng $x \in X$ trên tập thuộc tính B ký hiệu là $[\widetilde{x}]_B^\lambda$, được biểu diễn dưới dạng $[\widetilde{x}]_B^\lambda = \left\{[\widetilde{x}]_B^\lambda(y) \,|y \in X\right\}$, trong đó:

$$[\widetilde{x}]_B^\lambda(y) = \begin{cases} [\widetilde{x}]_B(y) \text{ nếu } y \in [x]_B^\lambda \\ 0 \text{ ngược lại} \end{cases} \tag{6}$$

Rất dễ để thấy rằng việc sử dụng quan hệ lân cận với hàm khoảng cách Chebyshev sẽ tạo ra một tính chất quan trọng $[\widetilde{x}]_B^\lambda = \bigcap_{f \in B} [\widetilde{x}]_{\{f\}}^\lambda$. Từ đây, chúng tôi xây dựng các hạt thông tin lân cận mờ với độ tương tự giữa hai đối tượng x và y theo thuộc tính f như sau [21]:

$$[\widetilde{x}]_{\{f\}}(y) = 1 - |f(x) - f(y)| \tag{7}$$

Như vậy, mô hình tập thô lân cận mờ bổ sung sự mô tả về độ thuộc cho các đối tượng trong một lớp lân cận. Về mặt trực cảm, mô hình này sẽ giải quyết khá tốt việc hạn chế tầm ảnh hưởng của các đối tượng nhiễu. Những đối tượng này có giá trị hàm thuộc được loại bỏ để không cho phép đóng góp các thông tin nhiễu vào các độ đo. Từ định nghĩa trên, mọi hạt thông tin lân cận mờ trên tập đối tượng X của tập thuộc tính B sẽ tạo ra một phủ lân cận mờ ký hiệu là $\left(X/R_{\widetilde{B}}^\lambda\right) = \left\{[\widetilde{x}]_B^\lambda \,|x \in X\right\}$. Khi đó, với hai phủ lân cận $\left(X/R_{\widetilde{A}}^\lambda\right)$ và $\left(X/R_{\widetilde{B}}^\lambda\right)$ cho trước, chúng ta nói rằng $\left(X/R_{\widetilde{A}}^\lambda\right)$ mịn hơn $\left(X/R_{\widetilde{B}}^\lambda\right)$, ký hiệu là $\left(X/R_{\widetilde{A}}^\lambda\right) \preceq \left(X/R_{\widetilde{B}}^\lambda\right)$ nếu $\forall x \in X$, $[\widetilde{x}]_A^\lambda \subseteq [\widetilde{x}]_B^\lambda$.

Mệnh đề 1: [19] Cho hệ thông tin quyết định $\mathcal{IS} = (X, F \cup D)$ và các tập con thuộc tính điều kiện $A, B \subseteq F$. Nếu $A \subseteq B$ thì $[\widetilde{x}]_B^\lambda \subseteq [\widetilde{x}]_A^\lambda$.

Chứng minh: Vì $A \subseteq B$ nên $[x]_B^\lambda \subseteq [x]_A^\lambda$. Từ đây, có thể dễ dàng suy ra $[\widetilde{x}]_B^\lambda \subseteq [\widetilde{x}]_A^\lambda$. ■

Mệnh đề 2: Cho hệ thông tin quyết định $\mathcal{IS} = (X, F \cup D)$ với λ_1, λ_2 là các bán kính lân cận và một tập con thuộc tính điều kiện $B \subseteq F$. Nếu $\lambda_1 \leq \lambda_2$ thì $\left(X/R_{\widetilde{B}}^{\lambda_1}\right) \preceq \left(X/R_{\widetilde{B}}^{\lambda_2}\right)$.

Chứng minh: Dễ thấy $\forall x \in X$, nếu $\lambda_1 \leq \lambda_2$ thì $[x]_B^{\lambda_1} \subseteq [x]_B^{\lambda_2}$. Do đó, $[\widetilde{x}]_B^{\lambda_1} \subseteq [\widetilde{x}]_B^{\lambda_2}$, chúng tôi xét ba trường hợp sau:

1) Với $y \in [x]_B^{\lambda_1} \Rightarrow y \in [x]_B^{\lambda_2}$, khi đó $[\widetilde{x}]_B^{\lambda_1}(y) = [\widetilde{x}]_B^{\lambda_2}(y) = [\widetilde{x}]_B(y)$.

2) Với $y \in [x]_B^{\lambda_2} \setminus [x]_B^{\lambda_1}$, khi đó $[\widetilde{x}]_B^{\lambda_1}(y) = 0$ và $[\widetilde{x}]_B^{\lambda_2}(y) = 1$.

3) Với $y \in X \setminus [x]_B^{\lambda_2} \Rightarrow y \notin [x]_B^{\lambda_1} \cup [x]_B^{\lambda_2}$, khi đó $[\widetilde{x}]_B^{\lambda_1}(y) = [\widetilde{x}]_B^{\lambda_2}(y) = 0$.

Từ kết quả của ba trường hợp trên, rất dễ để thấy rằng $\forall x, y \in X$, ta luôn có $[\widetilde{x}]_B^{\lambda_1}(y) \leq [\widetilde{x}]_B^{\lambda_2}(y)$. Điều này chỉ ra rằng, $[\widetilde{x}]_B^{\lambda_1} \subseteq [\widetilde{x}]_B^{\lambda_2}$, $\forall x \in X$. Do đó, $\left(X/R_{\widetilde{B}}^{\lambda_1}\right) \preceq \left(X/R_{\widetilde{B}}^{\lambda_2}\right)$. Mệnh đề đã được chứng minh. ■

Dựa trên Mệnh đề 2, rất dễ để thấy rằng khi giá trị của bán kính lân cận λ càng nhỏ thì kích thước của các hạt thông tin lân cận sẽ càng nhỏ. Điều này dẫn tới phủ lân cận mờ sẽ càng thô. Tiếp theo, chúng tôi sẽ định nghĩa về một xác suất điều kiện mà một đối tượng x cho trước thuộc về một tập đối tượng $Y \subseteq X$ thông qua hạt thông tin lân cận mờ được xác định bởi tập thuộc tính $B \subseteq F$. Xác suất này được biểu diễn bởi hàm $\eta_{\widetilde{Y}}^B(x)$ và được xác định như sau:

$$\eta_{\widetilde{Y}}^B(x) = \frac{\left|\widetilde{Y} \cap [\widetilde{x}]_B^\lambda\right|}{\left|[\widetilde{x}]_B^\lambda\right|} \tag{8}$$

trong đó, $\widetilde{Y}$ là tập mờ được biến đổi từ tập đối tượng Y với các đối tượng $y \in X$ có độ thuộc được xác định dựa trên công thức sau:

$$\widetilde{Y}(y) = \begin{cases} 1 \text{ nếu } y \in Y \\ 0 \text{ ngược lại} \end{cases} \tag{9}$$

Rất dễ để thấy rằng, $0 < \eta_{\widetilde{Y}}^B(x) \leq 1$. Từ đây, chúng tôi định nghĩa các khái niệm cơ bản về mô hình tập thô lân cận mờ thông qua các xấp xỉ như sau:

1) Xấp xỉ dưới lân cận mờ của tập đối tượng Y theo tập thuộc tính B:

$$\underline{B}(Y) = \left\{x \in X \,\middle|\, \eta_{\widetilde{Y}}^B(x) = 1\right\} \tag{10}$$

2) Xấp xỉ trên lân cận mờ của tập đối tượng Y theo tập thuộc tính B:

$$\overline{B}(Y) = \left\{x \in X \,\middle|\, \eta_{\widetilde{Y}}^B(x) > 0\right\} \tag{11}$$

2. Miền dương lân cận mờ

Trong phần này, chúng tôi sẽ trình bày về khái niệm miền dương lân cận mờ. Khái niệm này là một cơ sở quan trọng nhằm xây dựng một độ đo đánh giá ý nghĩa của các thuộc tính trong hệ thông tin. Các thuộc tính có ý nghĩa theo độ đo được lựa chọn sẽ đảm bảo tính toàn vẹn của hệ thông tin và ngược lại các thuộc tính khác sẽ được cho là dư thừa và không đóng góp các thông tin quan trọng. Như đã trình bày trong phần trên của bài báo, chúng ta nhận thấy rằng với một đối tượng bất kỳ $x \in X$ và một tập con thuộc tính

điều kiện $B \subseteq F$ thì hạt thông tin lân cận mờ $[\widetilde{x}]_B^\lambda$ được chứa một phần hoặc hoàn toàn trong lớp quyết định mờ tương ứng $[\widetilde{x}]_D$, trong đó $[x]_D \in X/D$. Từ đây, khái niệm về miền dương lân cận mờ mô tả một tập các đối tượng đặc biệt trong X mà ở đó các đối tượng này sẽ thuộc hoàn toàn vào lớp quyết định tương ứng theo tập thuộc tính B. Miền dương lân cận mờ của D được đưa ra bởi tập thuộc tính $B \subseteq F$ được xác định như sau:

$$POS_{\widetilde{B}}^{\lambda}(D) = \bigcup_{D_i \in X/D} \underline{B}(D_i) \tag{12}$$

Dựa trên khái niệm miền dương lân cận mờ, xét hệ thông tin quyết định $\mathcal{IS} = (X, F \cup D)$, độ nhất quán lân cận mờ được xác định theo công thức:

$$\gamma(F, D) = \frac{\left|POS_{\widetilde{F}}^{\lambda}(D)\right|}{|X|} \tag{13}$$

Từ khái niệm này, chúng ta nói rằng $\mathcal{IS}$ là nhất quán khi $\gamma(F, D) = 1$ và không nhất quán khi $\gamma(F, D) \neq 1$. Nói cách khác nếu $\mathcal{IS}$ được rút gọn và chỉ chứa các đối tượng nằm trong $POS_{\widetilde{F}}^{\lambda}(D)$ thì $\mathcal{IS}$ là nhất quán. Đối với hệ thông tin quyết định nhất quán ta ký hiệu $\mathcal{IS}' = (X', F \cup D)$ với $X' = POS_{\widetilde{F}}^{\lambda}(D)$.

Mệnh đề 3: Cho hệ thông tin quyết định $\mathcal{IS} = (X, F \cup D)$ và hai tập con thuộc tính điều kiện $A, B \subseteq F$. Nếu $A \subseteq B$ thì $\gamma(A, D) \leq \gamma(B, D)$.

Chứng minh: Với $A \subseteq F$, $\forall D_i \in X/D$ và một đối tượng bất kỳ $x \in X$, để $x \in \underline{A}(D_i)$ thì theo Công thức 13, $\eta_{\widetilde{D}_i}^{A}(x) = 1$, điều này dẫn tới, $[\widetilde{x}]_A^\lambda \subseteq \widetilde{D}_i$.

Mặt khác, $A \subseteq B$ nên theo Mệnh đề 1, $\forall x \in X$, $[\widetilde{x}]_B^\lambda \subseteq [\widetilde{x}]_A^\lambda$. Do đó, $\forall D_i \in X/D$, $\underline{A}(D_i) \subseteq \underline{B}(D_i)$. Từ đây, $POS_{\widetilde{A}}^{\lambda}(D) \subseteq POS_{\widetilde{B}}^{\lambda}(D)$ và $\gamma(A, D) \leq \gamma(B, D)$. ■

Mệnh đề 3 mô tả tính chất đơn điệu về quan hệ giữa kích thước tập thuộc tính điều kiện và giá trị của độ nhất quán lân cận mờ. Cụ thể, khi kích thước tập thuộc tính càng lớn thì độ nhất quán này sẽ càng cao và ngược lại.

Định lý 1: Cho $\mathcal{IS} = (X, F \cup D)$ là hệ thông tin quyết định không nhất quán. Giả sử rằng, $POS_{\widetilde{F}}^{\lambda}(D)'$ là miền dương lân cận mờ của $\mathcal{IS}'$, khi đó $POS_{\widetilde{F}}^{\lambda}(D) = POS_{\widetilde{F}}^{\lambda}(D)'$.

IV. THUẬT TOÁN LỰA CHỌN THUỘC TÍNH

Phần này sẽ định nghĩa một rút gọn trong phương pháp tập thô lân cận mờ. Sau đó, chúng tôi phân tích các rút gọn này và thiết kế một thuật toán hiệu quả cho việc chọn lọc thuộc tính trên hệ thông tin quyết định. Đầu tiên, chúng tôi sẽ định nghĩa một γ-rút gọn tối ưu trên miền dương lân cận mờ.

Định nghĩa 3: Cho hệ thông tin quyết định $\mathcal{IS} = (X, F \cup D)$, một tập con thuộc tính điều kiện $B \subseteq F$ được gọi là một γ-rút gọn của F nếu B thỏa mãn:

1) $\gamma(B, D) = \gamma(F, D)$,
2) $\forall f \in B$, $\gamma(B \setminus \{f\}, D) \neq \gamma(B, D)$.

Theo Định lý 1, dễ thấy rằng một γ-rút gọn trong Định nghĩa 3 là tập con thuộc tính tối thiểu bảo toàn yếu tố nhất quán như tập hợp tất cả các thuộc tính. Nói cách khác, rút gọn này duy trì các đối tượng trong miền dương của bảng quyết định ban đầu và bỏ qua các đối tượng khác. Theo định nghĩa, các đối tượng trong miền dương có độ thành viên bằng một, tức là mỗi đối tượng chỉ có thể được phân loại vào một lớp quyết định nhất định. Tuy nhiên, trên thực tế, có thể tồn tại nhiều đối tượng không thể phân loại một cách chắc chắn, tức là các đối tượng nằm ngoài miền dương. Rõ ràng, bỏ qua các đối tượng này trong quá trình tính toán sẽ ảnh hưởng đến chất lượng của rút gọn thu được. Do đó, chúng tôi sẽ định nghĩa một rút gọn khác để xử lý tất cả các đối tượng trong vũ trụ. Trước hết, chúng tôi sẽ định nghĩa về một độ đo được gọi là độ chắc chắn trong việc phân lớp của các đối tượng trong tập vũ trụ trên một tập con thuộc tính.

Định nghĩa 4: Cho hệ thông tin quyết định $\mathcal{IS} = (X, F \cup D)$ và một tập con thuộc tính $B \subseteq F$, độ chắc chắn trong việc phân lớp của các đối tượng trong X trên B được xác định bởi công thức

$$\widetilde{\eta}(B, D) = \frac{1}{|X|} \sum_{x \in X} \frac{\left|[\widetilde{x}]_B^\lambda \cap [\widetilde{x}]_D\right|}{\left|[\widetilde{x}]_B^\lambda\right|} \tag{14}$$

Rõ ràng, $0 \leq \widetilde{\eta}(B, D) \leq 1$, khả năng phân lớp của tập thuộc tính B sẽ càng cao khi $\widetilde{\eta}(B, D)$ càng lớn và ngược lại. Trong trường hợp $\widetilde{\eta}(B, D) = 1$, thì mọi đối tượng trong X sẽ được phân lớp chắc chắn dựa trên tập thuộc tính B. Từ đây, ta dễ dàng có được tính chất quan trọng tiếp theo.

Mệnh đề 4: Nếu $\mathcal{IS}$ là một hệ thông tin quyết định nhất quán thì $\widetilde{\eta}(F, D) = 1$.

Chứng minh: Mệnh đề dễ dàng chứng minh. ■

Định lý 2: Cho hệ thông tin quyết định nhất quán $\mathcal{IS} = (X, F \cup D)$, một tập con thuộc tính B là một γ-rút gọn của F nếu B là một tập con thuộc tính tối thiểu của F thỏa mãn $\widetilde{\eta}(B, D) = \widetilde{\eta}(F, D)$.

Chứng minh: Vì $\mathcal{IS}$ là hệ thông tin nhất quán, do đó theo Mệnh đề 4 thì $\widetilde{\eta}(B, D) = \widetilde{\eta}(F, D) = 1$ và rõ ràng rằng $\gamma(B, D) = \gamma(F, D) = 1$. Tiếp theo, chúng tôi sẽ chứng minh $\forall f \in F$, nếu $\widetilde{\eta}(B \setminus \{f\}, D) < \widetilde{\eta}(B, D)$ thì $\gamma(B \setminus \{f\}, D) < \gamma(B, D)$.

Giả sử rằng, tồn tại thuộc tính $f \in F$ sao cho $\gamma(B \setminus \{f\}, D) = \gamma(B, D)$. Khi đó, các đối tượng trong tập vũ trụ sẽ nằm trong miền dương lân cận mờ $POS_{\widetilde{B \setminus \{f\}}}^{\lambda}(D)$, nói cách khác thì với mọi đối tượng $x \in X$

thì $[\widetilde{x}]^{\lambda}_{B\setminus\{f\}} \subseteq [\widetilde{x}]_D$. Do đó, $\widetilde{\eta}(B\setminus\{f\},D) = 1$ hay $\widetilde{\eta}(B\setminus\{f\},D) = \widetilde{\eta}(B,D)$. Điều này là trái ngược với giả thiết $\widetilde{\eta}(B\setminus\{f\},D) < \widetilde{\eta}(B,D)$. Do đó, $\gamma(B\setminus\{f\},D) < \gamma(B,D)$. Như vậy, B cũng là một γ-rút gọn của F. ■

Từ Định lý trên, chúng ta hoàn toàn có thể đưa ra một rút gọn mới dựa trên độ chắc chắn trong việc phân lớp, chúng tôi gọi đây là một $\widetilde{\eta}$-rút gọn.

Định nghĩa 5: Cho hệ thông tin quyết định $\mathcal{IS} = (X, F\cup D)$, một tập con thuộc tính điều kiện $B \subseteq F$ được gọi là một $\widetilde{\eta}$-rút gọn của F nếu B thỏa mãn:

1) $\widetilde{\eta}(B,D) = \widetilde{\eta}(F,D)$,
2) $\forall f \in B, \widetilde{\eta}(B\setminus\{f\},D) \neq \widetilde{\eta}(B,D)$.

Tiếp theo, để chứng minh tính hiệu quả của $\widetilde{\eta}$-rút gọn, chúng tôi sẽ đưa ra một mối quan hệ giữa $\widetilde{\eta}$-rút gọn và γ-rút gọn thông qua trường hợp hệ thông tin là không nhất quán.

Định lý 3: B là $\widetilde{\eta}$-rút gọn của F trên $\mathcal{IS}'$ nếu và chỉ nếu B là γ-rút gọn của F trên $\mathcal{IS}$.

Chứng minh: Vì B là một γ-rút gọn của F trên $\mathcal{IS}$ nên $POS^{\lambda}_{\widetilde{B}}(D) = POS^{\lambda}_{\widetilde{F}}(D)$. Mặt khác, theo Định lý 1 thì $POS^{\lambda}_{\widetilde{F}}(D) = POS^{\lambda}_{\widetilde{F}}(D)'$. Do đó, $POS^{\lambda}_{\widetilde{B}}(D) = POS^{\lambda}_{\widetilde{F}}(D)' = X'$. Như vậy, B cũng là $\widetilde{\eta}$-rút gọn của F trên $\mathcal{IS}'$. Tương tự, dễ dàng chứng minh với trường hợp ngược lại. ■

Đối với các hệ thông tin không nhất quán, từ Định lý 3 có thể dễ dàng nhận thấy rằng $\widetilde{\eta}$-rút gọn cũng là một γ-rút gọn khi loại bỏ các đối tượng nằm ngoài miền dương lân cận mờ khỏi tập hợp vũ trụ. Do đó, về mặt tổng quát, rõ ràng $\widetilde{\eta}$-rút gọn hiệu quả hơn so với γ-rút gọn. Từ đây, chúng tôi xây dựng một công thức nhằm tính toán độ quan trọng của một thuộc tính trên hệ thông tin với một tập thuộc tính cho trước.

Định nghĩa 6: Cho hệ thông tin quyết định $\mathcal{IS} = (X, F\cup D)$, một tập con thuộc tính điều kiện $B \subseteq F$ và một thuộc tính $f \in F\setminus B$, độ ý nghĩa của thuộc tính f theo B, ký hiệu là $Sig(f,B)$ và được xác định như sau:

$$Sig(f,B) = \widetilde{\eta}(B\cup\{f\},D) - \widetilde{\eta}(B,D) \qquad (15)$$

Độ ý nghĩa của một thuộc tính theo một tập con thuộc tính điều kiện được xác định bằng việc đánh giá sự thay đổi thông qua mức chắc chắn khi bổ sung thuộc tính đó vào. Tuy nhiên, độ ý nghĩa có thể mang dấu âm do độ chắc chắn trong phân lớp không thỏa mãn tính chất đơn điệu so với kích thước của tập thuộc tính. Trong trường hợp này, thuộc tính được xét sẽ coi là chứa nhiễu và mang vai trò không quan trọng trong bảng dữ liệu. Từ định nghĩa này, chúng tôi sẽ thiết kế một thuật toán nhằm trích lọc một tập con thuộc tính từ hệ thông tin quyết định. Thuật toán với ý tưởng bắt đầu từ một tập rỗng các thuộc tính và sau đó lần lượt bổ sung các thuộc tính có độ quan trọng lớn nhất sau mỗi vòng lặp đến khi điều kiện dừng xảy ra. Thuật toán được trình bày cụ thể trên bảng mã giả 1.

Thuật toán 1: Lựa chọn thuộc tính dựa trên độ chắc chắn trong phân lớp (ASFCC)

Input: hệ thông tin $\mathcal{IS} = (X, F\cup D)$, ngưỡng lân cận λ.
Output: một rút gọn red

tính toán $\left(X/R^{\lambda}_{D}\right)$ và $\left(X/R^{\lambda}_{\widetilde{\{f\}}}\right), \forall f \in F$.
tính toán $\left(X/R^{\lambda}_{\widetilde{F}}\right) = \bigcap_{f\in F}\left(X/R^{\lambda}_{\widetilde{\{f\}}}\right)$
$red = \{f_0\}$ thỏa mãn: $\widetilde{\eta}(\{f_0\},D) = \max_{f\in F}\{\widetilde{\eta}(\{f\},D)\}$
$\mathcal{W} = red$
// Giai đoạn lọc: Tìm kiếm các rút gọn ứng tuyển
while $\widetilde{\eta}(red,D) \neq \widetilde{\eta}(F,D)$ **do**
    for $f \in F\setminus red$ **do**
        tính toán $\widetilde{\eta}(B\cup\{f_0\},D)$
        tính toán $Sig(f, red)$
    end for
    lựa chọn f_0 thỏa mãn: $Sig(f_0,red) = \max_{f\in F\setminus red}\{Sig(f,red)\}$ ▷ Nếu có nhiều thuộc tính có độ quan trọng bằng nhau thì ưu tiên lựa chọn thuộc tính có chỉ số thấp nhất.
    $red := red\cup\{f_0\}$
    $\mathcal{W} := \mathcal{W}\cup red$
end while
// Giai đoạn đóng gói: Lựa chọn rút gọn có độ chính xác phân lớp cao nhất
for w in $\mathcal{W}$ **do**
    Tính độ chính xác phân lớp trên w
end for
$red = w_{best}$ với $w_{best} \in \mathcal{W}$ có độ chính xác phân lớp cao nhất.
return red

Tiếp theo chúng tôi sẽ phân tích độ phức tạp của thuật toán ASFCC. Độ phức tạp tính toán của thuật toán ASFCC bao gồm hai giai đoạn là lọc và đóng gói. Rất dễ để thấy, giai đoạn lọc có độ phức tạp là $O\left(|F|^2|X|^2\right)$, trong đó $|F|$, $|X|$ tương ứng là số lượng thuộc tính và đối tượng trong hệ thông tin quyết định. Giả sử rằng, độ phức tạp của mô hình phân lớp được sử dụng trong giai đoạn đóng gói là $O(T)$. Khi đó, giai đoạn đóng gói có độ phức tạp là $O(T|F|)$. Như vậy, độ phức tạp của ASFCC là $O\left(|F|^2|X|^2\right) + O(T|F|)$. Tiếp theo, để làm rõ quá trình xử lý của thuật toán đề xuất, chúng tôi sẽ trình bày một ví dụ minh họa.

Ví dụ 2: Dựa trên Bảng I, các bước tiến hành thuật toán ASFCC với $\lambda = 0.2$ được thực hiện như sau:
Bước 1: Tính các phủ lân cận mờ tương ứng với các thuộc tính trong hệ thông tin quyết định:

- $X/D = \{\{x_1, x_4\}, \{x_2\}, \{x_3, x_5\}\}$

$$\Rightarrow \left(X/R^{\lambda}_{D}\right) = \begin{bmatrix} 1.00 & 0.00 & 0.00 & 1.00 & 0.00 \\ 0.00 & 1.00 & 0.00 & 0.00 & 0.00 \\ 0.00 & 0.00 & 1.00 & 0.00 & 1.00 \\ 1.00 & 0.00 & 0.00 & 1.00 & 0.00 \\ 0.00 & 0.00 & 1.00 & 0.00 & 1.00 \end{bmatrix}$$

- Tính phủ lân cận mờ trên các thuộc tính điều kiện

+ Tính các lân cận mờ trên thuộc tính f_1:

$[x_1]^{\lambda}_{\{f_1\}} = \{x_1, x_2, x_5\}$

$\Rightarrow [\widetilde{x_1}]^{\lambda}_{\{f_1\}} = \{1.00, 0.85, 0.00, 0.00, 0.80\}$,

$[x_2]^{\lambda}_{\{f_1\}} = \{x_1, x_2, x_5\}$

$\Rightarrow [\widetilde{x_2}]^{\lambda}_{\{f_1\}} = \{0.85, 1.00, 0.00, 0.00, 0.95\}$,

$[x_3]^{\lambda}_{\{f_1\}} = \{x_3, x_4\}$

$\Rightarrow [\widetilde{x_3}]^{\lambda}_{\{f_1\}} = \{0.00, 0.00, 1.00, 0.95, 0.00\}$,

$[x_4]^{\lambda}_{\{f_1\}} = \{x_3, x_4\}$

$\Rightarrow [\widetilde{x_4}]^{\lambda}_{\{f_1\}} = \{0.00, 0.00, 0.95, 1.00, 0.00\}$,

$[x_5]^{\lambda}_{\{f_1\}} = \{x_1, x_2, x_5\}$

$\Rightarrow [\widetilde{x_5}]^{\lambda}_{\{f_1\}} = \{0.80, 0.95, 0.00, 0.00, 1.00\}$.

Từ đây, ta dễ dàng thu được phủ lân cận mờ trên thuộc tính f_1:

$$\left(X/R^{\lambda}_{\widetilde{\{f_1\}}}\right) = \begin{bmatrix} 1.00 & 0.85 & 0.00 & 0.00 & 0.80 \\ 0.85 & 1.00 & 0.00 & 0.00 & 0.95 \\ 0.00 & 0.00 & 1.00 & 0.95 & 0.00 \\ 0.00 & 0.00 & 0.95 & 1.00 & 0.00 \\ 0.80 & 0.95 & 0.00 & 0.00 & 1.00 \end{bmatrix}$$

+ Tương tự, ta cũng tính được các phủ lân cận mờ trên các thuộc tính còn lại.

$$\left(X/R^{\lambda}_{\widetilde{\{f_2\}}}\right) = \begin{bmatrix} 1.00 & 0.85 & 0.90 & 0.00 & 0.00 \\ 0.85 & 1.00 & 0.00 & 0.00 & 0.00 \\ 0.90 & 0.00 & 1.00 & 0.00 & 0.00 \\ 0.00 & 0.00 & 0.00 & 1.00 & 0.00 \\ 0.00 & 0.00 & 0.00 & 0.00 & 1.00 \end{bmatrix}$$

$$\left(X/R^{\lambda}_{\widetilde{\{f_3\}}}\right) = \begin{bmatrix} 1.00 & 0.90 & 0.00 & 0.00 & 0.80 \\ 0.90 & 1.00 & 0.80 & 0.85 & 0.90 \\ 0.00 & 0.80 & 1.00 & 0.95 & 0.90 \\ 0.00 & 0.85 & 0.95 & 1.00 & 0.95 \\ 0.80 & 0.90 & 0.90 & 0.95 & 1.00 \end{bmatrix}$$

$$\left(X/R^{\lambda}_{\widetilde{\{f_4\}}}\right) = \begin{bmatrix} 1.00 & 0.90 & 0.00 & 0.90 & 0.00 \\ 0.90 & 1.00 & 0.00 & 0.80 & 0.00 \\ 0.00 & 0.00 & 1.00 & 0.00 & 0.90 \\ 0.90 & 0.80 & 0.00 & 1.00 & 0.00 \\ 0.00 & 0.00 & 0.90 & 0.00 & 1.00 \end{bmatrix}$$

$$\left(X/R^{\lambda}_{\widetilde{\{f_5\}}}\right) = \begin{bmatrix} 1.00 & 0.90 & 0.85 & 0.95 & 0.00 \\ 0.90 & 1.00 & 0.95 & 0.85 & 0.85 \\ 0.85 & 0.95 & 1.00 & 0.80 & 0.90 \\ 0.95 & 0.85 & 0.80 & 1.00 & 0.00 \\ 0.00 & 0.85 & 0.90 & 0.00 & 1.00 \end{bmatrix}$$

Dựa trên độ đo khoảng cách Chebyshev, thu được phủ lân cận mờ trên tập thuộc tính F.

$$\left(X/R^{\lambda}_{\widetilde{F}}\right) = \begin{bmatrix} 1.00 & 0.85 & 0.00 & 0.00 & 0.00 \\ 0.85 & 1.00 & 0.00 & 0.00 & 0.00 \\ 0.00 & 0.00 & 1.00 & 0.00 & 0.00 \\ 0.00 & 0.00 & 0.00 & 1.00 & 0.00 \\ 0.00 & 0.00 & 0.00 & 0.00 & 1.00 \end{bmatrix}$$

- Tính độ chắc chắn trong phân lớp trên mỗi thuộc tính:

+ $\widetilde{\eta}(\{f_1\}, D) = 0.424$, $\widetilde{\eta}(\{f_2\}, D) = 0.686$, $\widetilde{\eta}(\{f_3\}, D) = 0.359$, $\widetilde{\eta}(\{f_4\}, D) = 0.750$, $\widetilde{\eta}(\{f_5\}, D) = 0.480$. Như vậy, $red = \{f_4\}$ do $\widetilde{\eta}(\{f_4\}, D)$ là lớn nhất.

+ Ta cũng tính được $\widetilde{\eta}(F, D) = 0.816$. Vì $\widetilde{\eta}(red, D) < \widetilde{\eta}(F, D)$, thuật toán tiếp tục chuyển sang bước tiếp theo.

Bước 2: Xử lý lọc với các thuộc tính còn lại:

+ Tính độ chắc chắn trong phân lớp của mỗi thuộc tính trong $F \backslash red$: $\widetilde{\eta}(red \cup \{f_1\}, D) = 0.816$, $\widetilde{\eta}(red \cup \{f_2\}, D) = 0.816$, $\widetilde{\eta}(red \cup \{f_3\}, D) = 0.690$ và $\widetilde{\eta}(red \cup \{f_5\}, D) = 0.750$.

+ Từ đây, thu được độ quan trọng của các thuộc tính còn lại:

$Sig(f_1, red) = 0.816 - 0.750 = 0.066$,

$Sig(f_2, red) = 0.816 - 0.750 = 0.066$,

$Sig(f_3, red) = 0.690 - 0.750 < 0.000$,

$Sig(f_5, red) = 0.750 - 0.750 = 0.000$.

Có thể thấy $Sig(f_3, red) < 0$ nên f_3 được coi là thuộc tính mang thông tin nhiễu. Vì $Sig(f_1, red)$ lớn nhất và f_1 có chỉ số nhỏ nhất nên $red = \{f_4, f_1\}$. Vì $\widetilde{\eta}(red \cup \{f_1\}, D) = \widetilde{\eta}(F, D)$, nên thuật toán kết thúc.

Chúng ta nhận thấy rằng, với $D_1 = \{x_1, x_4\}$, $D_2 = \{x_2\}$ và $D_3 = \{x_3, x_5\}$, các xấp xỉ lân cận dưới của thuộc tính f_1 là $\underline{N_{\{f_1\}}}(D_1) = \underline{N_{\{f_1\}}}(D_2) = \underline{N_{\{f_1\}}}(D_3) = \emptyset$. Do đó, trên một số độ đo như hàm phụ thuộc hay miền dương thì thuộc tính f_1 sẽ không được xét tới. Mặc dù thuộc tính f_1 là một thuộc tính đóng vai trò quan trọng trên hệ thông tin và được lựa chọn theo thuật toán đề xuất. Do đó, độ chắc chắn trong phân lớp được chúng tôi đề xuất là hiệu quả trong việc đánh giá các thuộc tính quan trọng.

V. THỰC NGHIỆM

1. Quá trình so sánh các thuật toán

Trong phần này, chúng tôi sẽ chứng minh hiệu quả của thuật toán đề xuất trên một số thực nghiệm. Thuật toán đề xuất được so sánh với hai thuật toán theo tiếp cận thô mờ là HANDI [22] và AFNRDE [23] theo ba tiêu chuẩn: Kích thước rút gọn, độ chính xác phân lớp SVM (RBF) với kiểm định chéo 10-fold và thời gian thực thi (giây). Các bộ dữ liệu được sử dụng được mô tả trong Bảng II được lấy từ hai kho dữ liệu là UCI [24] và OpenML [25]. Đối với mỗi bộ dữ liệu, chúng tôi tiến hành chuẩn hóa min-max để đưa miền giá trị của từng thuộc tính về đoạn [0, 1] và sau đó chia thành mười phần xấp xỉ. Ở mỗi phần, các thuật toán sẽ được thực hiện để tìm ra một rút gọn. Kết quả cuối cùng sẽ được dựa trên hiệu quả trung bình khi thực hiện trên

toàn bộ mười phần từ dữ liệu. Tất cả thực nghiệm được thực hiện trên một máy tính cá nhân với bộ xử lý Xeon(R) E3-1545M v5, 2.90 GHz và 16GB bộ nhớ.

Bảng II
DỮ LIỆU SỬ DỤNG TRONG THỰC NGHIỆM

ID	Dữ liệu	$\|X\|$	$\|F\|$	$\|D\|$	Nguồn
1	Leaf	340	15	30	UCI [24]
2	Sonar	208	60	2	UCI [24]
3	Movement	360	90	15	UCI [24]
4	Urban	675	147	9	OpenML [25]
5	Spectf	349	44	2	UCI [24]
6	Mfeat	2000	76	10	OpenML [25]

Đầu tiên, chúng tôi tiến hành thực thi các thuật toán thông qua quá trình duyệt các tham số để tìm một rút gọn có độ chính xác phân lớp cao nhất. Các tham số của cả ba thuật toán được duyệt trong khoảng giá trị $[0, 1]$ với bước nhảy cho mỗi lần duyệt là 0.05. Quá trình duyệt tham số được mô tả trong Hình 2 với việc biểu diễn độ chính xác phân lớp của các rút gọn thu được từ mỗi thuật toán theo từng giá trị tham số. Với mỗi giá trị tham số khác nhau, các thuật toán đều đưa ra các kết quả rút gọn khác nhau.

Bảng III
KÍCH THƯỚC RÚT GỌN VÀ THỜI GIAN XỬ LÝ

ID	Dữ liệu	ASFCC		HANDI		AFNRDE	
		K/t	T/g	K/t	T/g	K/t	T/g
1	Leaf	**5.6**	0.25	6.6	**0.24**	5.8	0.25
2	Sonar	**5.8**	0.16	7.3	**0.07**	11.2	0.14
3	Movement	**12**	**0.36**	12	0.55	17	1.76
4	Urban	**9.4**	**1.48**	18.7	1.70	18	2.40
5	Spectf	**8.8**	0.15	9.8	**0.13**	9.4	0.19
6	Mfeat	**13.9**	**7.63**	17.5	16.9	16.4	19.7
Trung bình		**9.25**	**1.67**	11.9	3.26	12.9	4.07

Tiếp theo, chúng tôi trình bày về kích thước tập rút gọn và thời gian thực thi của mỗi thuật toán khi thu được rút gọn tốt nhất. Bảng III mô tả kích thước rút gọn (K/t) và thời gian thực thi (T/g) của các thuật toán. Rõ ràng, cả ba thuật toán đều thu được các rút gọn có kích thước nhỏ hơn rất nhiều so với tập thuộc tính gốc. Trên toàn bộ các bộ dữ liệu, thuật toán đề xuất đều trích chọn được một tập con thuộc tính nhỏ hơn. Chẳng hạn, trên bộ dữ liệu Sonar, thuật toán của chúng tôi giảm được 97.2% số lượng thuộc tính, trong khi với HANDI là 96.5% và của AFNRDE là 94.6%. Bên cạnh đó, kích thước trung bình của ASFCC cũng nhỏ nhất với 9.25, trong khi hai thuật toán còn lại là HANDI và AFNRDE lần lượt là 11.9 và 12.9. Điều này chỉ ra rằng, thuật toán có hiệu quả rất cao khi thực hiện trên các bộ dữ liệu có số lượng thuộc tính lớn.

Chúng tôi tiếp tục phân tích về thời gian thực thi của thuật toán. Trong sáu trường hợp, có ba trường hợp thuật toán của chúng tôi có thời gian thực hiện nhanh hơn. Thời gian trung bình khi thực hiện trên toàn bộ các thuật toán là 1.67 giây, nhanh hơn rất nhiều so với hai thuật toán còn lại. Điều này có thể được hiểu rằng, ưu điểm của tập thô lân cận mờ giúp thu hẹp không gian tính toán và có khả năng tính toán nhanh trên các bộ dữ liệu có số chiều lớn. Bên cạnh đó, độ chắc chắn trong việc phân lớp được chúng tôi đưa ra cũng có tính đơn giản. Điều này giúp cho thuật toán có thể tìm kiếm thuộc tính một cách nhanh chóng trong từng vòng lặp tại giai đoạn lọc.

Cuối cùng, chúng tôi sẽ phân tích về hiệu quả phân lớp từ các rút gọn thu được của mỗi thuật toán. Như kết quả biểu diễn trong Bảng IV, trên hầu hết các bộ dữ liệu, rút gọn thu được của chúng tôi có độ chính xác cao hơn. Đặc biệt, trên các bộ dữ liệu nhiễu, thuật toán đề xuất chứng minh được hiệu quả vượt trội trong việc cải thiện độ chính xác. Chẳng hạn, trên bộ dữ liệu Movement, rút gọn của chúng tôi cải thiện được 6.45% độ chính xác so với tập thuộc tính gốc, trong khi với HANDI và AFNRDE lần lượt là 4.65% và 1.81%. Điều này nhấn mạnh rằng, độ chắc chắn trong việc phân lớp có khả năng lựa chọn rất hiệu quả các thuộc tính quan trọng. Từ các kết quả được trình bày, có thể nói rằng thuật toán đề xuất đạt hiệu quả vượt trội so với các thuật toán theo cách tiếp cận tập thô mờ. Tuy nhiên, để có cái nhìn rõ nét hơn, chúng tôi cũng trình bày một số ưu nhược điểm của thuật toán như sau:

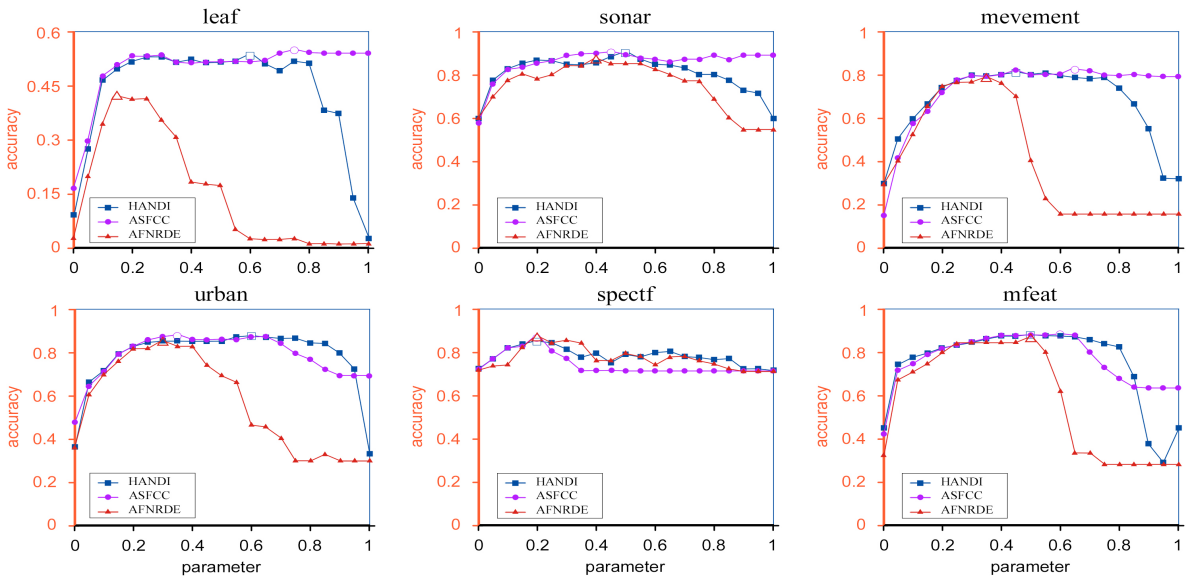
Ưu điểm:

1) Thuật toán tận dụng ưu điểm của tập thô lân cận với việc sử dụng một ngưỡng bán kính λ để loại bỏ các đối tượng có sự phân bố khác biệt so với phần lớn các đối tượng trong tập vũ trụ. Về mặt trực cảm, các đối tượng này là nguyên nhân làm giảm hiệu quả trên các mô hình phân lớp. Do đó, khi xử lý trên các bộ dữ liệu có độ chính xác phân lớp gốc thấp, thì thuật toán sẽ đạt kết quả rất tốt trong việc tìm kiếm một rút gọn tối ưu, chẳng hạn trên các bộ dữ liệu như Leaf, Sonar hay Movement.

2) Kết hợp với việc rút gọn không gian tìm kiếm khi loại bỏ các đối tượng thông qua một ngưỡng bán kính λ, độ chắc chắn trong việc phân lớp các đối tượng mà nghiên cứu đề xuất được áp dụng trên các hạt thông tin lân cận mờ đã cho thấy sự hiệu quả trong việc cải thiện cả về thời gian tính toán cũng như có khả năng đánh giá độ quan trọng của từng thuộc tính trên các hệ thông tin nhất quán và không nhất quán. Ngoài ra, với việc đo lường rất tốt lượng thông tin trong các hạt thông tin lân cận mờ thuộc vào các lớp quyết định, thuật toán đề xuất cũng có khả năng xử lý hiệu quả trên các bộ dữ liệu đa lớp, chẳng hạn như Urban, Movement, Leaf và Mfeat.

Hạn chế:

1) Thuật toán cần phải duyệt trên từng giá trị bán kính λ để lựa chọn một rút gọn tối ưu. Về cơ bản, giá trị λ sẽ phù hợp với đặc trưng của từng bộ dữ liệu. Mặc dù, trong



Hình 2. Quá trình duyệt tham số trên các thuật toán.

Bảng IV
ĐỘ CHÍNH XÁC PHÂN LỚP TRÊN CÁC THUẬT TOÁN

ID	Dữ liệu	Dữ liệu gốc	ASFCC		HANDI		AFNRDE	
		SVM (RBF)	SVM (RBF)	λ	SVM (RBF)	ϵ	SVM (RBF)	δ
1	Leaf	0.435 ± 0.114	**0.559 ± 0.112**	0.75	0.544 ± 0.122	0.60	0.435 ± 0.088	0.15
2	Sonar	0.793 ± 0.099	**0.913 ± 0.054**	0.45	0.899 ± 0.084	0.50	0.875 ± 0.072	0.40
3	Movement	0.775 ± 0.077	**0.825 ± 0.074**	0.65	0.811 ± 0.082	0.45	0.789 ± 0.062	0.35
4	Urban	0.837 ± 0.045	**0.877 ± 0.042**	0.35	0.876 ± 0.032	0.60	0.858 ± 0.033	0.30
5	Spectf	0.831 ± 0.073	0.868 ± 0.049	0.20	0.857 ± 0.071	0.20	**0.871 ± 0.079**	0.20
6	Mfeat	0.836 ± 0.035	**0.877 ± 0.026**	0.60	0.870 ± 0.024	0.50	0.872 ± 0.026	0.50
Trung bình		0.751 ± 0.074	**0.820 ± 0.059**	0.50	0.809 ± 0.069	0.47	0.783 ± 0.060	0.32

thực nghiệm, quá trình so sánh đều được tiến hành trên các thuật toán theo hướng tiếp cận mức hạt (có tham số duyệt). Tuy nhiên, trong tương lai, nghiên cứu sẽ tiếp tục được mở rộng để đưa ra một công thức xác định giá trị bán kính λ cho từng bộ dữ liệu cụ thể.

2) Thuật toán sẽ không hiệu quả khi xử lý trên các bộ dữ liệu có mật độ phân bố thấp. Cụ thể, trong các lớp quyết định có sự phân bố mật độ thấp, nếu sử dụng ngưỡng bán kính nhỏ, các lớp lân cận không thể chứa đủ các đối tượng thuộc cùng một lớp quyết định.

2. Đánh giá hiệu quả tham số

Trong phần này, chúng tôi sẽ đánh giá hiệu quả của mô hình khi có sự thay đổi của tham số lân cận δ. Dựa trên Hình 2 và Bảng V, rất dễ để thấy rằng, với mỗi một giá trị lân cận δ, thuật toán sẽ đều thu được một kết quả rút gọn khác nhau. Do đó, các tiêu chí về độ chính xác phân lớp, thời gian xử lý và kích thước rút gọn sẽ đều có sự khác nhau. Cụ thể, sự thay đổi này được trình bày như sau:

1) Thời gian thực thi: Khi ngưỡng λ càng nhỏ, số lượng đối tượng bị loại bỏ càng nhiều, điều này làm cho không gian tính toán càng thu hẹp và dẫn tới quá trình xử lý của thuật toán sẽ càng nhanh. Điều này được thể hiện rất rõ trong Bảng V. Ngược lại, khi λ càng lớn, không gian tính toán sẽ càng lớn, điều này làm cho thời gian tính toán trên cả hai giai đoạn lọc và đóng gói sẽ càng chậm.

2) Kích thước rút gọn: Đối với thuật toán đề xuất, có thể thấy rằng trên hầu hết các bộ dữ liệu, kích thước rút gọn có sự tăng dần từ 0 đến 0.6 và sau đó giảm dần. Với λ nhỏ, các phủ lân cận mờ sẽ càng thô. Điều này làm cho điều kiện dừng của thuật toán dễ dàng xảy ra. Do đó, thuật toán sẽ hội tụ sớm và thu được một rút gọn có kích thước nhỏ.

3) Độ chính xác phân lớp: Dựa trên kết quả Bảng V, độ chính xác phân lớp sẽ có sự giao động ở mức cao khi ngưỡng λ nằm trong khoảng từ 0.4 đến 0.7. Ngưỡng λ trung bình để thuật toán thu được một rút gọn tối ưu tính được là 0.5. Về mặt trực cảm, việc lựa chọn một ngưỡng δ sẽ phụ

Bảng V
ẢNH HƯỞNG CỦA THAM SỐ TỚI MÔ HÌNH

ID	Dữ liệu	Kết quả	Tham số mô hình λ										
			0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
1	Leaf	Kích thước	1.0	4.3	6.6	7.7	7.1	6.0	5.3	4.9	5.6	5.3	5.2
		Thời gian	0.02	0.05	0.09	0.12	0.14	0.19	0.21	0.23	0.27	0.29	0.32
		Độ chính xác	0.165± 0.087	0.477± 0.121	0.541± 0.113	0.547± 0.091	0.521± 0.109	0.532± 0.117	0.529± 0.120	0.550± 0.130	0.556± 0.122	0.550± 0.124	0.550± 0.124
2	Sonar	Kích thước	1.0	1.3	1.8	2.9	5.9	4.0	4.6	3.7	3.4	2.9	2.9
		Thời gian	0.01	0.06	0.08	0.11	0.14	0.17	0.20	0.25	0.28	0.30	0.30
		Độ chính xác	0.586± 0.097	0.832± 0.088	0.861± 0.052	0.870± 0.044	0.909± 0.058	0.889± 0.057	0.875± 0.040	0.875± 0.060	0.880± 0.064	0.880± 0.060	0.880± 0.060
3	Movement	Kích thước	1.0	4.0	6.7	8.9	11.0	11.7	10.4	11.5	11.6	9.5	9.1
		Thời gian	0.02	0.05	0.07	0.16	0.19	0.21	0.30	0.41	0.47	0.51	0.59
		Độ chính xác	0.158± 0.054	0.586± 0.110	0.728± 0.065	0.800± 0.060	0.808± 0.069	0.808± 0.077	0.814± 0.075	0.820± 0.077	0.797± 0.091	0.797± 0.095	0.794± 0.094
4	Urban	Kích thước	1.0	3.5	5.6	21.7	15.2	10.9	15.7	7.6	4.0	2.3	2.3
		Thời gian	0.03	0.16	0.87	1.24	1.86	2.27	2.68	3.39	3.67	4.51	4.89
		Độ chính xác	0.486± 0.109	0.723± 0.059	0.836± 0.041	0.876± 0.046	0.868± 0.024	0.868± 0.036	0.874± 0.037	0.844± 0.062	0.776± 0.099	0.701± 0.050	0.699± 0.050
5	Spectf	Kích thước	1.0	4.1	8.8	1.9	1.0	1.0	1.0	1.0	1.0	1.0	1.0
		Thời gian	0.04	0.09	0.15	0.24	0.55	0.58	0.61	0.61	0.66	0.67	0.67
		Độ chính xác	0.731± 0.087	0.825± 0.052	0.868± 0.049	0.776± 0.095	0.731± 0.087	0.728± 0.085	0.728± 0.085	0.728± 0.085	0.728± 0.085	0.728± 0.085	0.728± 0.085
6	Mfeat	Kích thước	2.0	6.3	10.1	18.5	18.1	17.5	13.9	7.0	3.3	2.8	2.8
		Thời gian	0.56	1.21	2.38	4.97	5.58	6.37	7.63	8.64	9.66	10.71	10.87
		Độ chính xác	0.432± 0.076	0.758± 0.050	0.822± 0.036	0.859± 0.027	0.873± 0.021	0.872± 0.021	0.877± 0.026	0.812± 0.050	0.688± 0.063	0.642± 0.061	0.642± 0.061

thuộc vào đặc trưng của từng dữ liệu. Trên các bộ dữ liệu có độ chính xác phân lớp thấp, chẳng hạn như Leaf hay Movement, có thể sử dụng một ngưỡng $\lambda \geq 0.5$ để loại bỏ nhiều đối tượng nhiễu và giúp cho các độ đo đạt hiệu quả cao trong việc đưa ra các thuộc tính quan trọng.

VI. KẾT LUẬN

Các phương pháp lựa chọn thuộc tính theo hướng tiếp cận tập thô mờ và tập thô lân cận giải quyết tốt trên các hệ thông tin quyết định gốc mà không cần tiến trình rời rạc hóa dữ liệu. Tuy nhiên, các phương pháp này còn gặp nhiều khó khăn khi xử lý trên các hệ thông tin nhiễu. Trước thách thức đó, nghiên cứu này ban đầu đã dựa trên một mô hình mới được gọi là mô hình tập thô lân cận mờ với mong muốn hạn chế tầm ảnh hưởng của các đối tượng nhiễu trong hệ thông tin. Tiếp theo, nghiên cứu cũng đề xuất độ đo phân lớp chắc chắn làm cơ sở cho việc xây dựng một rút gọn hiệu quả. Cuối cùng, nghiên cứu thiết kế một thuật toán theo hướng tiếp cận lai ghép để tìm kiếm một rút gọn tối ưu trên hệ thông tin quyết định. Các tính chất, ví dụ và kết quả thực nghiệm cũng được trình bày chi tiết trong bài báo để thấy được tầm quan trọng mà nghiên cứu này mang lại.

LỜI CẢM ƠN

Các tác giả cảm ơn phòng mô phỏng và tính toán hiệu năng cao (SHPC) thuộc Viện Công nghệ HaUI, Trường Đại học Công nghiệp Hà Nội đã hỗ trợ trong quá trình triển khai nghiên cứu này.

TÀI LIỆU THAM KHẢO

[1] Z. Pawlak, "Rough sets", *International Journal of Computer & Information Sciences*, vol. 11, pp. 341–356, 1982.
[2] N. L. Giang and V. D. Thi, "An attribute reduction algorithm in a decision based on improved entropy", *Journal of Computer Science and Cybernetics*, vol. 27, no. 2, pp. 166–175, 2012.
[3] N. L. Giang, N. T. Tung, and V. D. Thi, "A new method for attribute reduction to imcomplete decision table based on metric", *Journal of Computer Science and Cybernetics*, vol. 28, no. 2, pp. 129–140, 2012.
[4] P. V. Anh, V. D. Thi, and N. N. Cuong, "A novel algorithm for finding all reducts in the incomplete decision table", *Journal of Computer Science and Cybernetics*, vol. 39, no. 4, p. 313–321, 2023.
[5] L. A. Zadeh, "Fuzzy sets", *Information and Control*, vol. 8, pp. 338–353, 1965.
[6] D. Dubois and H. Prade, "Putting rough sets and fuzzy sets together", *Intelligent Decision Support*, vol. 11. Dordrecht, The Netherlands: Springer, 1992, pp. 203–232.
[7] 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.
[8] T. K. Sheeja and A. S. Kuriakose, "A novel feature selection method using fuzzy rough sets", *Computers in Industry*, vol. 97, pp. 111–116, 2018.
[9] X. Che, D. Chen, and J. Mi, "Label correlation in multi-label classification using local attribute reductions with fuzzy rough sets", *Fuzzy Sets and Systems*, vol. 426, pp. 121–144, 2022.
[10] Q. Hu, D. Yu, and Z. Xie, "Information-preserving hybrid data reduction based on fuzzy-rough techniques", *Pattern Recognition Letters*, vol. 27, no. 5, pp. 414–423, 2006.
[11] B. Liang, L. Wang, and Y. Liu, "Attribute reduction based on improved information entropy", *Journal of Intelligent & Fuzzy Systems*, vol. 36, no. 1, pp. 709–718, 2019.
[12] 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.
[13] P. M. N. Ha, N. L. Giang, N. V. Thien, and N. B. Quang, "An Incremental Algorithm for Finding Reducts of Incomplete Decision Tables", *Research, Development and Application on Information and Communication Technology*, vol. 2019, no. 1, 2019.
[14] V. V. Dinh, V. D. Thi, N. L. Giang, and N. Q. Tao, "Partition Distance Based Attribute Reduction in Incomplete Decision Tables, Research", *Development and Application on Information and Communication Technology*, vol. 2, no. 1, 2015.
[15] G. C. Y. Tsang, D. Chen, E. C. C. Tsang, J. W. T. Lee, and D. S. Yeung, "On attributes reduction with fuzzy rough sets", *Proc. IEEE Int. Conf. Syst., Man, Cybern.*, vol. 3, no. 3, pp. 2775–2780, 2005.
[16] W. L. Hung, M. S. Yang, "Similarity measures of intuitionistic fuzzy sets based on lp metric", *International Journal of Approximate Reasoning*, vol. 46, no. 1, pp. 120–136, 2007.
[17] P. V. Anh, N. L. Giang, N. N. Thuy, N. T. Thuy, and P. D. Khanh, "A Novel Incremental Algorithm for Finding the Reduct on the Decision Table when Deleting the Object Set", *Research, Development and Application on Information and Communication Technology*, vol. 2023, no. 2, 2023.
[18] P. V. Anh, N. N. Thuy, V. D. Thi, and N. L. Giang, "On distance-based attribute reduction with α, β-level intuitionistic fuzzy sets", *IEEE Access 11*, pp. 138095–138107, 2023.
[19] C. Wang, M. Shao, Q. He, Y. Qian, and Y. Qi, "Feature subset selection based on fuzzy neighborhood rough sets", *KBS*, vol. 111, pp. 173-179, 2016.
[20] Q. Hu, D. Yu, J. Liu, and C. Wu, "Neighborhood rough set based heterogeneous feature subset selection", *KBS*, vol. 178, pp. 3577-3594, 2008.
[21] N. L. Giang, L. H. Son, T. T. Ngan, T. M. Tuan, H. T. Phuong and 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, 2020.
[22] C. Wang, Q. Hu, X. Wang, D. Chen, Y .Qian and Z. Dong, "Feature Selection Based on Neighborhood DiscriminationIndex Index", *IEEE Transactions on Neural Networks and Learning Systems*, vol. 29, 2017.
[23] X. Zhang, Y. Fan and J. Yang, "Feature selection based on fuzzy-neighborhood relative decision entropy", *Pattern Recognition Letters*, vol. 146, 2021.
[24] UCI Machine learning repository, "Data," 2024. https://archive.ics.uci.edu/
[25] Openml machine learning platform, "Data," 2024. https://www.openml.org/

Phạm Việt Anh tốt nghiệp Đại học chuyên ngành Hệ thống Thông tin Quản lý tại Đại học Kinh tế Quốc dân năm 2016 và nhận bằng thạc sĩ chuyên ngành Khoa học máy tính tại Đại học Công nghệ Thông tin và Truyền thông, Thái Nguyên năm 2019. Hiện đang là nghiên cứu sinh tại Học viện khoa học và Công nghệ, Viện Hàn lâm Khoa học và Công nghệ Việt Nam và là giảng viên thuộc Viện Công nghệ HaUI, Đại học Công nghiệp Hà Nội. Các lĩnh vực nghiên cứu chính bao gồm Trí tuệ nhân tạo, tính toán tối ưu, học máy, khai phá dữ liệu và tính toán mềm.
Email: anhpv@haui.edu.vn, vietanh.hn.4078@gmail.com

Đặng Trọng Hợp tốt nghiệp Thạc sĩ tại Đại học Bách khoa Hà Nội năm 2007, nhận bằng Tiến sĩ tại khoa Công nghệ thông tin, Đại học Kỹ thuật Lê Quý Đôn năm 2019. Hiện là Trưởng khoa Công nghệ thông tin, trường Đại học Công nghiệp Hà Nội. Các lĩnh vực nghiên cứu chính bao gồm Machine Learning, Deep Learning, Computer vision, Fuzzy Sets.
Email: dangtronghop@gmail.com

Ngô Quang Huy tốt nghiệp Đại học chuyên ngành Công nghệ thông tin, trường Đại học Công nghiệp Hà Nội năm 2021. Hiện là chuyên viên Trung tâm Truyền thông và Quan hệ công chúng, trường Đại học Công nghiệp Hà Nội. Các lĩnh vực nghiên cứu chính bao gồm machine learning, soft computing, fuzzy sets.
Email: huynq1@haui.edu.vn

Lê Xuân Hùng tốt nghiệp Đại học Sư phạm Toán, Trường Đại học Sư phạm Thái Nguyên. Năm 2000 tốt nghiệp Thạc sĩ Toán học tại Viện Toán học Việt Nam. Năm 2006 bảo vệ luận án Tiến sĩ Toán học tại Viện Toán học Việt Nam, chuyên ngành Đảm bảo toán học cho máy tính và hệ thống tính toán. Hiện nay công tác tại Khoa Công nghệ thông tin Trường Đại học Công nghiệp Hà Nôi. Hướng nghiên cứu chính: lý thuyết đồ thị.
Email: hungmath68@gmail.com

Trần Phi Lực tốt nghiệp Đại học ngành Công nghệ thông tin tại Đại học Công nghiệp Hà Nội năm 2022. Năm 2024 tốt nghiệp thạc sĩ Hệ thống thông tin tại Đại học Công nghiệp Hà Nội. Hiện nay đang công tác tại phòng Khoa học và Công nghệ, Đại học Công nghiệp Hà Nội. Các lĩnh vực nghiên cứu chủ yếu bao gồm Machine Learning, Deep Learning, Clustering, Fuzzy Sets
Email: luctp@haui.edu.vn

Đỗ Đình Lực tốt nghiệp Đại học ngành Công nghệ thông tin tại Đại học Công nghệ thông tin và Truyền thông, Thái Nguyên năm 2011. Năm 2015 tốt nghiệp thạc sĩ Công nghệ thông tin tại Trường Đại học Công nghệ, Đại học Quốc Gia Hà Nội. Hiện nay công tác tại Khoa Công nghệ thông tin, Trường Đại học Công nghệ thông tin và Truyền thông Thái Nguyên. Các lĩnh vực nghiên cứu chính bao gồm Machine Learning, Deep Learning, Cyber security.
Email: ddluc@ictu.edu.vn