Một phương pháp thiết kế ngữ nghĩa tính toán của các từ ngôn ngữ giải bài toán phân lớp dựa trên luật mờ

Nguyễn Đức Dư[1], Phạm Đình Phong[1], Phạm Đình Vũ[2], Nguyễn Đức Thảo[3]
[1]Khoa Công nghệ thông tin, Trường Đại học Giao thông vận tải
[2]Cục Công nghệ thông tin và Thống kê hải quan, Tổng cục Hải quan
[3]Viện Khoa học và Công nghệ quân sự

Tác giả liên hệ: Phạm Đình Phong, phongpd@utc.edu.vn
Ngày nhận bài: 20/01/2020, ngày sửa chữa: 17/06/2020
Định danh DOI: 10.32913/mic-ict-research-vn.v2020.n1.914

Tóm tắt: Thiết kế ngữ nghĩa tính toán của các từ ngôn ngữ trong cơ sở luật và biểu diễn cấu trúc của chúng đóng vai trò quan trọng trong việc nâng cao hiệu suất cũng như tính giải nghĩa được của hệ dựa trên luật mờ. Bài báo này trình bày phương pháp thiết kế ngữ nghĩa tính toán dựa trên tập mờ dạng hàm *S* được sinh bởi đại số gia tử mở rộng và được biểu diễn dưới dạng cấu trúc phân hoạch mờ đảm bảo tính giải nghĩa được của hệ phân lớp dựa trên luật mờ. Kết quả thực nghiệm với 23 tập dữ liệu chuẩn cho thấy hệ phân lớp với ngữ nghĩa tính toán dựa trên tập mờ dạng hàm *S* cho độ chính xác phân lớp tốt hơn so với ngữ nghĩa tính toán dựa trên tập mờ tam giác và hình thang cũng như chỉ ra tính hiệu quả của biểu diễn cấu trúc phân hoạch mờ đảm bảo tính giải nghĩa được của hệ phân lớp so với cấu trúc phân hoạch đã được đề xuất trước đó.

Từ khóa: *đại số gia tử, thứ tự ngữ nghĩa, hàm thuộc, hệ phân lớp dựa trên luật mờ.*

Title: **A Design Method of Computational Semantics of Linguistic Words for Fuzzy Rule-based Classifier**

Abstract: The design of computational semantics of linguistic terms in the fuzzy rule bases and structural representation of them play important roles in improving the performance and the interpretability of fuzzy rule-based systems. This paper presents a method of designing computational fuzzy sets-based semantics in form of *S*-shape membership function generated by the enlarged hedge algebras and represented as fuzzy partition structure to ensure the interpretability of the fuzzy rule-based classifiers. Experimental results over 23 real-world datasets have shown that the classifier with the fuzzy set-based computational semantics in form of *S*-shape membership function gives better classification accuracy than the ones previously proposed with triangular and trapezoidal fuzzy sets based semantics as well as shown the efficiency of the fuzzy partition structure representation which ensures the interpretability of the fuzzy rule-based classifiers in comparison with the existing ones.

Keywords: *hedge algebras, order-based semantics, membership function, fuzzy rule-based classifier.*

I. GIỚI THIỆU

Hệ phân lớp dựa trên luật mờ (Fuzzy Rule Based Classifier – FRBC) có nhiều ứng dụng trong lĩnh vực khai phá dữ liệu [1–4, 18–22] do mô hình phân lớp này có ưu điểm là dễ hiểu với người dùng và có thể sử dụng các tri thức dạng luật if-then được trích rút tự động từ dữ liệu như là tri thức của họ.

Trong [4, 5], Ishibuchi và Yamamoto đề xuất phương pháp trích rút hệ luật mờ tối giản cho FRBC từ cấu trúc phân hoạch mờ đa thể hạt được thiết kế sẵn bằng cách áp dụng một số kỹ thuật trong khai phá dữ liệu như độ tin cậy, độ hỗ trợ và trọng số luật kết hợp với thuật toán di truyền đa mục tiêu. Alcalá và các cộng sự đề xuất trong [1] một số phương pháp lựa chọn một đơn thể hạt tốt nhất trong số các thể hạt được thiết kế sẵn ban đầu do họ quan niệm rằng cấu trúc phân hoạch mờ đa thể hạt không giải nghĩa được. Sau đó thuật toán di truyền được áp dụng để lựa chọn hệ luật tối ưu đồng thời với tối ưu các tham số của các hàm thuộc. Một giản đồ tiến hóa đa mục tiêu nhanh và hiệu quả được Antonelli và các cộng sự đề xuất trong [2] có tên là PAES-RCS. Đây là một tiếp cận tiến hóa đa mục tiêu thực

hiện huấn luyện đồng thời cơ sở luật và cơ sở dữ liệu của FRBC. Trong pha đầu, tập luật mờ ứng cử viên được sinh từ các phân hoạch mờ được thiết kế sẵn bằng thuật toán C4.5. Sau đó, thuật toán tiến hóa đa mục tiêu được thực hiện để lựa chọn một tập luật mờ từ tập luật ứng cử viên đồng thời với lựa chọn các điều kiện của luật mờ cũng như hiệu chỉnh các tham số của hàm thuộc. Trong [5], Rey và các cộng sự đề xuất thêm một mục tiêu là tính thích hợp của luật (rule relevance) bên cạnh hai mục tiêu là tính chính xác (accuracy) và tính giải nghĩa được (interpretability) cho giải thuật tiến hóa đa mục tiêu lựa chọn hệ luật tối ưu cho hệ dựa trên luật mờ. Trong [18], Rudzinski đề xuất thuật toán tiến hóa đa mục tiêu thiết kế hệ phân lớp dựa trên luật mờ hướng tính giải nghĩa được. Trong quá trình huấn luyện, các tham số của các hàm thuộc và cấu trúc của cơ sở luật được tiến hóa đồng thời. Các độ đo về số tập mờ hoạt động và số biến đầu vào hoạt động (tức được sử dụng bởi ít nhất một luật) cùng với độ dài trung bình của luật được sử dụng để đánh giá tính giải nghĩa được của hệ phân lớp. Một mở rộng của giải thuật Chi nổi tiếng thiết kế hệ phân lớp dựa trên luật mờ phân tán cho phân lớp dữ liệu lớn bằng cách áp dụng khung làm việc dữ liệu lớn phổ biến Apache Hadoop, được đề xuất bởi Elkanoa và các cộng sự trong [19]. Trong [20] và [21] các tác giả đề xuất xây dựng các hệ phân lớp dựa trên luật mờ đặc thù áp dụng trong các lĩnh vực y tế và đánh giá rủi ro tín dụng. Một phương pháp thiết kế FRBC sử dụng giải thuật tiến hóa lượng tử đa dân số (Multi-population quantum evolutionary algorithm) với sự tái tạo lại luật mâu thuẫn được Zhang và các cộng sự đề xuất trong [22].

Như đã được trình bày ở trên, các phương pháp thiết kế FRBC trên cơ sở lý thuyết tập mờ [1–4, 18–22] trích rút các luật mờ từ các phân hoạch mờ được thiết kế sẵn trên miền giá trị của các thuộc tính sử dụng các tập mờ. Để nâng cao hiệu quả phân lớp, giá trị của các tham số hàm thuộc được hiệu chỉnh thích nghi bằng giải thuật tối ưu. Do không có cơ sở hình thức kết nối giữa ngữ nghĩa của các từ ngôn ngữ với các tập mờ nên ngữ nghĩa tính toán dựa trên tập mờ không phản ảnh đúng ngữ nghĩa thực của các ngôn ngữ sau quá trình tối ưu và làm ảnh hưởng đến tính giải nghĩa được của hệ luật phân lớp.

Đại số gia tử (ĐSGT) [6–8] đã có những ứng dụng hiệu quả trong khai phá dữ liệu [9–12], điều khiển mờ [13], xử lý ảnh [14], lập lịch [15], ... ĐSGT khai thác tính thứ tự về ngữ nghĩa của các từ trong miền giá trị ngôn ngữ của biến ngôn ngữ để hình thành một cơ sở hình thức toán học cho việc liên kết ngữ nghĩa tính toán dựa trên tập mờ với ngữ nghĩa vốn có của các từ ngôn ngữ. Trên cơ sở đó, ĐSGT đã được ứng dụng hiệu quả để thiết kế tối ưu các từ ngôn ngữ cùng với ngữ nghĩa tính toán dựa trên tập mờ hình tam giác [9] và hình thang [10] cho các FRBC. Ngữ nghĩa tính toán dựa trên tập mờ hình thang có ưu điểm so với hình tam giác là biểu diễn được lõi ngữ nghĩa khoảng của các từ ngôn ngữ. Tuy nhiên, cả hai dạng tập mờ này đều có các cạnh được biểu diễn bởi các hàm tuyến tính có độ dốc lớn nên chưa thật mềm dẻo và gây mất mát thông tin lớn. Một phương pháp thiết kế ngữ nghĩa tính toán dựa trên tập mờ dạng hàm S và được sinh bởi ĐSGT mở rộng [10] cho các FRBC được trình bày trong bài báo này. Do hàm S là hàm phi tuyến nên phù hợp với sự biến thiên về ngữ nghĩa vốn có của các từ ngôn ngữ trong khi vẫn biểu diễn được lõi ngữ nghĩa khoảng của các từ ngôn ngữ.

Mặt khác, để đảm bảo tính giải nghĩa được của hệ dựa trên luật mờ được thiết kế theo tiếp cận ĐSGT, trong [11] các tác giả đã đưa ra bốn ràng buộc trên ngữ nghĩa tính toán của các từ ngôn ngữ. Các phương pháp thiết kế ngữ nghĩa tính toán dựa trên tập mờ cho FRBC của các công bố trong [9, 10] đều chưa thỏa tất cả bốn ràng buộc này. Cụ thể, một từ ngôn ngữ hx được sinh ra từ từ ngôn ngữ x bởi gia tử h có ngữ nghĩa cụ thể hơn x nhưng vẫn giữ nguyên ngữ nghĩa gốc của x. Ví dụ, từ ngôn ngữ "*rất trẻ*" được sinh ra từ từ ngôn ngữ "*trẻ*" bởi gia tử *rất* có ngữ nghĩa cụ thể hơn "*trẻ*" nhưng vẫn giữ được ngữ nghĩa gốc của "*trẻ*". Do đó, để thỏa ràng buộc thứ ba trong [11], trong biểu diễn cấu trúc phân hoạch mờ sử dụng các tập mờ thì độ hỗ trợ của tập mờ ứng với từ ngôn ngữ hx phải nằm trọn trong độ hỗ trợ của tập mờ ứng với từ ngôn ngữ x. Tuy nhiên, các thiết kế phân hoạch mờ trong [9, 10] không thỏa tính chất này. Bài báo này trình bày một phương pháp biểu diễn cấu trúc phân hoạch mờ sử dụng các tập mờ có dạng hàm S thỏa tất cả bốn ràng buộc trong [11], tức đảm bảo tính giải nghĩa được của hệ phân lớp dựa trên luật mờ.

Phần còn lại của bài báo được bố cục như sau: mục 2 trình bày tóm tắt ĐSGT mở rộng và hệ phân lớp dựa trên luật mờ với ngữ nghĩa dựa trên tập mờ dạng hàm S; mục 3 trình bày kết quả thực nghiệm và thảo luận; một số kết luận rút ra trong mục 4.

II. NỘI DUNG NGHIÊN CỨU

1. Một số khái niệm cơ bản về đại số gia tử mở rộng

ĐSGT mở rộng [10] được xây dựng bằng việc bổ sung một gia tử nhân tạo h_0 nhằm mô hình hóa lõi ngữ nghĩa của các từ ngôn ngữ.

Một cấu trúc $\mathcal{AX}^{en} = (X_{en}, G, C, H_{en}, \leq)$ được gọi là ĐSGT mở rộng (ĐSGTMR) của ĐSGT tuyến tính và sinh tự do $\mathcal{AX}$ nếu thỏa các tiên đề bổ sung sau:

(A1) $h_0x \notin H(G) = \{\sigma c \mid c \in G\}$ và $hh_0x = h_0x$ luôn là điểm bất động.

(A2) $h_px \geq x \Rightarrow h_{-q}x \leq \ldots \leq h_{-1}x \leq h_0x \leq h_1x \leq \ldots \leq h_px$

$$h_px \leq x \Rightarrow h_px \leq \ldots \leq h_1x \leq h_0x \leq h_{-1}x \leq \ldots \leq h_{-q}x.$$

Một hàm fm: $X_{en} \rightarrow [0, 1]$ được gọi là độ đo tính mờ của ĐSGTMR $\mathcal{AX}^{en}$ nếu nó thỏa các tính chất sau:

(F1): $fm(\mathbf{0}) + fm(c^-) + fm(W) + fm(c^+) + fm(\mathbf{1}) = 1$;

(F2): $\sum_{h \in H_{en}} fm(hx) = fm(x)$ với $\forall x \in H(G)$;

(F3): $\forall x, y \in H(G), \forall h \in H_{en}$ tỷ lệ $fm(hx)/fm(hy) = fm(x)/fm(y)$ không phụ thuộc vào bất kỳ từ ngôn ngữ nào trong X_{en} được gọi là độ đo tính mờ của gia tử h và được ký hiệu là $\mu(h)$.

Độ đo tính mờ của một từ ngôn ngữ của ĐSGTMR X_{en} thỏa các tính chất sau:

(1) $\sum_{x \in X_{(k)}} fm(x) = 1, k > 0$. Với $k = 1$ thì $fm(\mathbf{0}) + fm(c^-) + fm(W) + fm(c^+) + fm(\mathbf{1}) = 1$;

(2) $\sum_{h \in H_{en}} \mu(h) = 1$

(3) $fm(hx) = \mu(h) fm(x)$, với $\forall h \in H_{en}$, $\forall x \in H(\{c^-, c^+\})$ và $hx \neq x$;

(4) $fm(x) = \mu(h_n)...\mu(h_1) fm(c)$, trong đó $x = h_n...h_1 c$, $c \in \{c^-, c^+\}$, là biểu diễn chính tắc của $x \in X_{en}$.

Cho độ đo tính mờ $fm : X_{en} \to [0, 1]$ của một ĐSGTMR $\mathcal{A}X^{en}$ của biến một ngữ $\mathcal{X}$ và mỗi từ $x \in X_{en}$ được liên kết với một khoảng $\mathfrak{I}(x) \subseteq [0, 1]$. Các khoảng này được gọi là các khoảng tính mờ ứng với các từ của $\mathcal{X}$ nếu thỏa các điều kiện sau:

(FI1): $|\mathfrak{I}(x)| = fm(x)$ với $\forall x \in X_{en}$ và $|\mathfrak{I}(x)|$ biểu thị độ dài của khoảng $\mathfrak{I}(x)$;

(FI2): Tập $\{\mathfrak{I}(hx) | x \in X_{en}\}$ tạo thành một phân hoạch của $\mathfrak{I}(x)$ và có thứ tự tương đồng với thứ tự của các từ ngôn ngữ liên kết với chúng.

Khoảng tính mờ mức k của x được ký hiệu là $\mathfrak{I}_k(x)$. Quy ước rằng các khoảng tính mờ là mở phải và đóng trái, khoảng tính mờ của hằng tử **1** là đóng cả hai phía.

Ánh xạ ngữ nghĩa định lượng khoảng $f(x)$ của từ ngôn ngữ x được xác định là hàm $f(x) = \mathfrak{I}(h_0 x), x \in X_{en}$ và khẳng định này đã được chứng minh trong [10].

2. Thiết kế FRBC với ngữ nghĩa tính toán dựa trên tập mờ dạng hàm S

Bài toán thiết kế hệ phân lớp dựa trên luật mờ $\mathcal{P}$ được định nghĩa như sau: Một tập $\boldsymbol{P} = \{(\boldsymbol{d}_p, C_p) | \boldsymbol{d}_p \in \boldsymbol{D}, C_p \in \boldsymbol{C}, p = 1, \ldots, m\}$ gồm m mẫu dữ liệu, trong đó $\boldsymbol{d}_p = [d_{p,1}, d_{p,2}, ..., d_{p,n}]$ là dòng thứ p^{th}, $\boldsymbol{C} = \{C_s | s = 1, \ldots, M\}$ là tập gồm M nhãn lớp, n là số thuộc tính.

Hệ cơ sở luật cho bài toán phân lớp được sử dụng trong bài báo này là tập luật có trong số dưới dạng:

$$\text{Luật } R_q \ : \textbf{If } \mathcal{X}_1 \text{ is } A_{q,1} \textbf{ and } ... \textbf{ and } \mathcal{X}_n \text{ is } A_{q,n} \quad \textbf{then} C_q \, with \ CF_q, for \ q = 1, \ldots, N \tag{1}$$

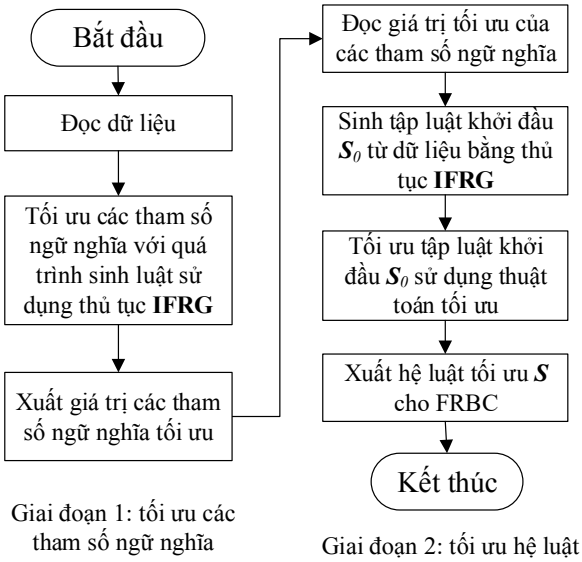
trong đó $\chi = \{\mathcal{X}_j, j = 1, \ldots, n\}$ là tập n biến ngôn ngữ ứng với n thuộc tính của tập dữ liệu $\boldsymbol{P}$; $A_{q,j}$ là các giá trị ngôn ngữ của thuộc tính thứ $j, F_j; C_q$ là nhãn lớp và CF_q là trọng số của luật R_q. Luật R_q được viết gọn lại như sau:

$$\boldsymbol{A}_q \Rightarrow C_q \text{ with } CF_q, \text{ với } q = 1, \ldots, N \tag{2}$$

trong đó $\boldsymbol{A}_q$ là tiền đề của luật thứ q.

Giải bài toán $\mathcal{P}$ là trích xuất từ tập dữ liệu **P** một tập luật **S** có dạng (1) nhỏ gọn, dễ hiểu với người dùng và có độ chính xác phân lớp cao. Phương pháp thiết kế hệ phân lớp dựa trên luật mờ theo tiếp cận ĐSGT gồm hai bước (xem Hình 1):

1) Thiết kế tối ưu các từ ngôn ngữ cùng với ngữ nghĩa tính toán dựa trên tập mờ của chúng sử dụng giải thuật tối ưu. Sau bước này ta thu được bộ tham số ngữ nghĩa tối ưu.
2) Trích xuất từ tập dữ liệu huấn luyện tập luật tối ưu cho hệ phân lớp trên cơ sở thỏa hiệp giữa tính dễ hiểu và độ chính xác của hệ phân lớp sử dụng giải thuật tối ưu.



Hình 1. Phương pháp hai bước thiết kế FRBC

ĐSGTMR là cung cấp một cơ sở hình thức cho phép ngữ nghĩa định tính xác định giá trị ngữ nghĩa định lượng khoảng của các từ ngôn ngữ, và trên cơ sở đó ngữ nghĩa dựa trên tập mờ có lõi là một khoảng của chúng được xây dựng. Trong bài báo này chúng tôi sử dụng ĐSGTMR để sinh ngữ nghĩa dựa trên tập mờ có dạng hàm S có lõi là một khoảng cho hệ phân lớp dựa trên luật mờ.

Mỗi ĐSGT $\mathcal{A}X_j^{en}$ được liên kết với một thuộc tính thứ j của tập dữ liệu cảm sinh các từ ngôn ngữ $X_{j,(k_j)}$ có độ dài lớn nhất k_j theo thứ tự ngữ nghĩa của chúng. Vì ngữ nghĩa định lượng khoảng $f(x_{j,i}) = \mathfrak{I}(h_0 x_{j,i}) \subseteq \mathfrak{I}(x_{j,i})$ biểu thị lõi ngữ nghĩa của từ ngôn ngữ $x_{j,i}$ nên được dùng để biểu diễn đỉnh của tập mờ dạng hàm S ứng với từ $x_{j,i}$. Các giá

trị trong khoảng đỉnh của tập mờ phù hợp với ngữ nghĩa định tính của từ nhất nên có giá trị là 1.

Ký kiệu $\mathcal{L}(\bullet)$ và $\mathcal{R}(\bullet)$ lần lượt là điểm mút trái và mút phải của một khoảng bất kỳ. Giả sử đặt $a = \mathcal{R}(f(x_{j,i-1}))$, $c = \mathcal{L}(f(x_{j,i}))$, $d = \mathcal{R}(f(x_{j,i}))$, $g = \mathcal{L}(f(x_{j,i+1}))$, khi đó $b = a + (c - a)/4, e = d + (g - e)/4$ và v là một điểm dữ liệu. Ta có hàm biểu diễn độ thuộc của v vào nửa trái của hàm S, S_{left} như sau:

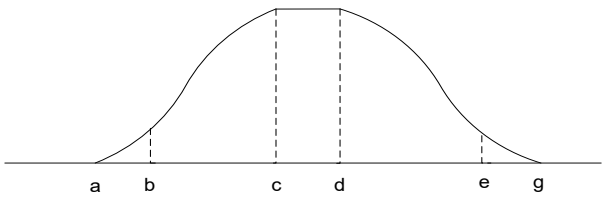
$$S_{left} = \begin{cases} 0, & 0 \le v \le a \\ \frac{(v-a)^2}{(b-a)(c-a)}, & a \le v \le b \\ 1 - \frac{(v-c)^2}{(c-b)(c-a)}, & b \le v \le c \\ 1, & v \ge c \end{cases}$$

và hàm biểu diễn độ thuộc của v vào nửa phải của hàm S, S_{right} như sau:

$$S_{right} = \begin{cases} 1, & 0 \le v \le d \\ 1 - \frac{(v-d)^2}{(d-e)(d-g)}, & d \le v \le e \\ \frac{(v-g)^2}{(e-d)(g-d)}, & e \le v \le g \\ 0, & v \ge g \end{cases}$$

Tập mờ dạng hàm S được biểu diễn như Hình 2.

Trong bài báo này, tập mờ dạng hàm S được sử dụng để phân hoạch miền giá trị thuộc tính của tập dữ liệu dưới dạng cấu trúc đa thể hạt được đề xuất trong [10], được gọi là phân hoạch k_1 và dưới dạng cấu trúc đa thể hạt được đề xuất trong [11] với mức $k = 1$ được tách thành hai mức 0 và 1, được gọi là phân hoạch k_0.



Hình 2. Biểu diễn tập mờ dạng hàm S

Trong cấu trúc phân hoạch k_1 mỗi thể hạt được phân hoạch bởi các tập mờ ứng với các từ ngôn ngữ có độ dài bằng nhau và hai phần tử 0 và 1, và theo thứ tự ngữ nghĩa của các từ ngôn ngữ tương ứng. Cấu trúc phân hoạch k_0 khác với k_1 là mức $k = 1$ gồm các từ ngôn ngữ có độ dài bằng 1 được tách thành hai thể hạt: thể hạt thứ nhất (mức $k = 0$) gồm các hằng tử $\mathbf{0}_0$, W và $\mathbf{1}_0$, và thể hạt thứ hai (mức $k = 1$) gồm 4 từ ngôn ngữ $\mathbf{0}_1$, c^-, c^+ và $\mathbf{1}_1$. Với cách thiết kế này, độ hỗ trợ của tập mờ ứng với từ ngôn ngữ x hoàn toàn chứa độ hỗ trợ của từ ngôn ngữ hx và trong [11] đã chứng minh phân hoạch k_0 đảm bảo tính giải nghĩa được của hệ dựa trên luật mờ.

Với các từ ngôn ngữ không phải là các hằng tử $\mathbf{0}$ và $\mathbf{1}$, giá trị của a là giá trị đầu mút phải của giá trị định lượng khoảng của từ gần nhất bên trái có cùng độ dài và giá trị g là đầu mút trái của giá trị định lượng khoảng của từ gần nhất bên phải có cùng độ dài. Ví dụ, Hình 3 biểu diễn cấu trúc phân hoạch k_1 và Hình 4 biểu diễn cấu trúc phân hoạch k_0, sử dụng các tập mờ dạng hàm S với độ dài tối đa của các từ ngôn ngữ $k_j = 2$. Trong đó, tập mờ ứng với từ Lc^+ có mút trái $a = \mathcal{R}(f(Lc^-))$ và mút phải $g = \mathcal{L}(f(Vc^+))$, tương tự với các tập mờ khác.

Với giá trị cụ thể của các tham số ngữ nghĩa bao gồm $fm(c^-)$, $fm(W_j)$, $fm(\mathbf{0}_j)$, $fm(\mathbf{1}_j)$, $\mu(h_{j,i})$, $\mu(h_{j,0})$ là độ đo tính mờ tương ứng của c_j^- , W_j, $\mathbf{0}_j$, $\mathbf{1}_j$, $h_{j,i}$, $h_{j,0}$ và với giá trị cụ thể của k_j, các khoảng tính mờ $\Im_k(x_{j,i})$, $x_{j,i} \in X_{j,k}, k \le k_j$ và các ngữ nghĩa định lượng khoảng $f(x_{j,i})$ được tính toán. Các khoảng tính mờ $\Im_{k_j}(x_{j,i})$ tạo thành phân hoạch mức k_j trên miền giá trị của thuộc tính j. Có duy nhất một khoảng tính mờ trong số các khoảng tính mờ $\Im_{k_j}(x_{j,i})$ chứa điểm dữ liệu $d_{p,j}$ của mẫu dữ liệu d_p. Tất cả các khoảng tính mờ mức k_j chứa $d_{p,j}$ $(0 \le j \le n)$ tạo thành một siêu hộp $\mathcal{H}_p$ và chỉ sinh các luật mờ từ các siêu hộp loại này. Luật mờ cơ sở có độ dài n được sinh từ $\mathcal{H}_p$ với nhãn lớp C_p của mẫu dữ liệu d_p có dạng sau:

if $\boldsymbol{X}_1$ is $x_{1,i(1)}$ **and** ... **and** $\boldsymbol{X}_n$ is $x_{n,i(n)}$ **then** C_p (R_b)

Các luật mờ thứ cấp có độ dài $L \le n$ thu được bằng cách bỏ bớt $n - L$ thuộc tính có dạng sau:

if $\boldsymbol{X}_{j1}$ is $x_{j1,i(j1)}$ **and**...**and** $\boldsymbol{X}_{jt}$ is $x_{jt,i(jt)}$**then** C_q (R_{snd})

trong đó $1 \le j_1 \le ... \le j_t \le n$. Nhãn lớp C_q của luật R_q được xác định bởi độ tin cậy $c(\boldsymbol{A}_q \Rightarrow C_h)$ [3, 4] của R_q:

$$C_q = \text{argmax}(c(\boldsymbol{A}_q \Rightarrow C_h) \mid h = 1, \ldots, M) \qquad (3)$$

Độ tin cậy của luật mờ được tính như sau:

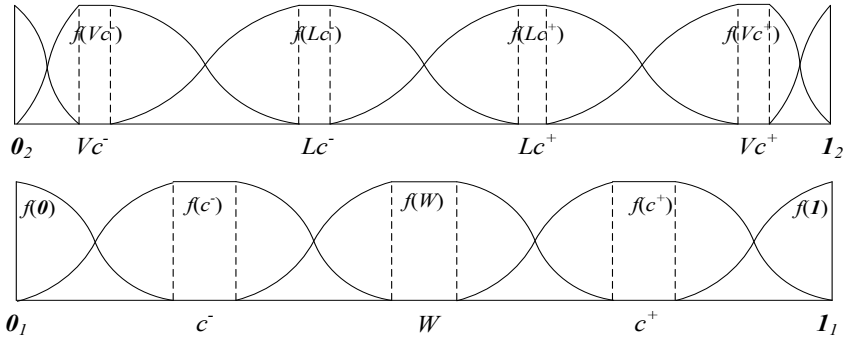
$$c(\boldsymbol{A}_q \Rightarrow C_h) = \sum_{d_p \in C_h} \mu_{\boldsymbol{A}_q}(d_p) / \sum_{p=1}^{m} \mu_{\boldsymbol{A}_q}(d_p) \qquad (4)$$

trong đó $\mu_{\boldsymbol{A}_q}(d_p)$ là độ đốt cháy của mẫu dữ liệu d_p đối với tiền đề luật của R_q và thường được tính bằng biểu thức toán tử nhân theo công thức sau:

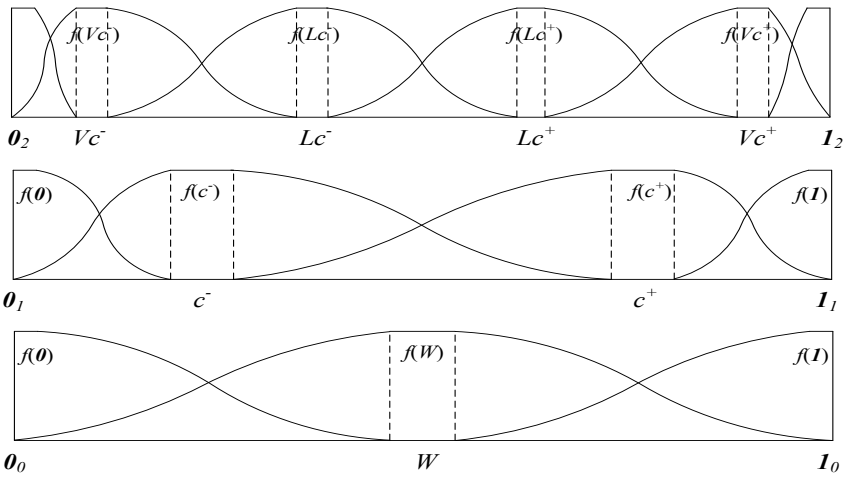
$$\mu_{\boldsymbol{A}_q}(d_p) = \prod_{j=1}^{n} \mu_{q,j}(d_{p,j}) . \qquad (5)$$

với $\mu_{q,j}(d_{p,j})$ là độ thuộc của điểm dữ liệu $d_{p,j}$ vào tập mờ $A_{q,j}$.

Tập luật ứng viên thu được sau khi sàng lọc các luật không nhất quán bằng độ hỗ trợ. Tiếp theo, một tiêu chuẩn sàng được sử dụng để chọn ra tập luật khởi đầu $\boldsymbol{S}_0$ gồm $NR_0 = NB_0 \times M$ luật với M là số nhãn lớp và NB_0 là số luật dành cho mỗi lớp. Tiêu chuẩn sàng được chọn có thể là độ tin cậy c (công thức (4)), độ hỗ trợ s hoặc tích $c \times s$. Độ hỗ trợ được tính theo công thức sau [3]:



Hình 3. Cấu trúc phân hoạch k_1 với tập mờ dạng hàm S và $k_j = 2$



Hình 4. Cấu trúc phân hoạch k_0 với tập mờ dạng hàm S và $k_j = 2$

$$s(\boldsymbol{A}_q \Rightarrow C_h) = \sum_{d_p \in C_h} \mu_{\boldsymbol{A}_q}(d_p)/m. \quad (6)$$

Mỗi luật được gán một trọng số để nâng cao độ chính xác phân lớp. Trong bài báo này, trọng số luật được tính theo công thức [3]:

$$CF_q = c\left(\boldsymbol{A}_q \Rightarrow C_q\right) - c_{q,2nd}, \quad (7)$$

trong đó $c_{q,2nd}$ là độ tin cậy lớn nhất của các luật có cùng tiền đề A_q nhưng khác kết luận C_q:

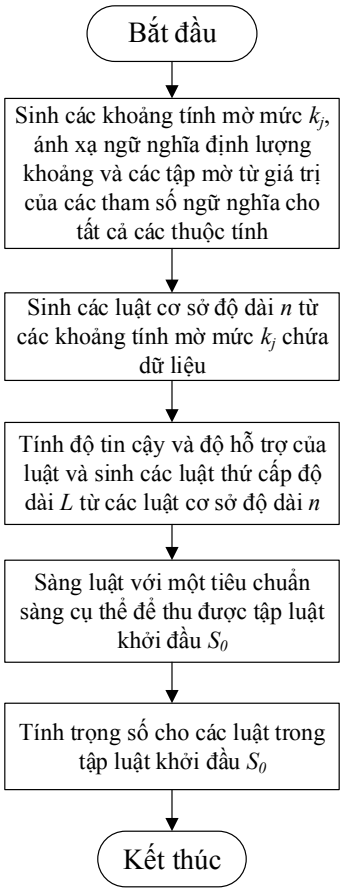
$$c_{q,2nd} = \max(c(\boldsymbol{A}_q \Rightarrow Class\ h) \mid h = 1, \ldots, M;\ h \neq C_q) \quad (8)$$

Quá trình sinh luật trên là thủ tục sinh tập luật khởi đầu ***IFRG***(Π, ***P***, NR_0, L) [9], trong đó Π là tập giá trị của các tham số ngữ nghĩa và L là số tiền đề tối đa của mỗi luật. Thủ tục này được trực quan hóa như được thể hiện trong Hình 5. Độ phức tạp của thủ tục sinh tập luật khởi đầu **IFRG** là đa thức đối với số mẫu và số thuộc tính của tập dữ liệu ***D*** và đã được chứng minh trong [9].

Mỗi loại dữ liệu có sự phân bố dữ liệu khác nhau cần bộ tham số ngữ nghĩa phù hợp để nâng cao hiệu suất phân lớp. Do đó, một thuật toán tối ưu được áp dụng để tìm bộ tham số ngữ nghĩa tối ưu và chúng được sử dụng để sinh tập luật khởi đầu làm đầu vào cho thủ tục lựa chọn tập luật nhỏ gọn và dễ hiểu cho hệ phân lớp trên cơ sở thỏa hiệp giữa độ chính xác và độ phức tạp của hệ phân lớp.

III. KẾT QUẢ THỰC NGHIỆM VÀ THẢO LUẬN

Mục này trình bày các kết quả thực nghiệm của các hệ phân lớp dựa trên luật mờ sử dụng cấu trúc phân hoạch k_0 và k_1 với ngữ nghĩa tính toán dựa trên tập mờ có dạng hàm S của các từ ngôn ngữ và so sánh đánh giá với các hệ phân lớp khác để minh chứng tính hiệu quả của các hệ phân lớp được đề xuất.



Hình 5. Lưu đồ thủ tục sinh tập luật khởi đầu

1. Cài đặt thực nghiệm

Các thực nghiệm được được cài đặt bằng ngôn ngữ C# chạy trên Windows 7. Các tập dữ liệu thực nghiệm được lấy từ nguồn KEEL-Dataset tại địa chỉ http://sci2s.ugr.es/keel/datasets.php. Phương pháp kiểm tra chéo 10 nhóm được áp dụng để huấn luyện và kiểm tra. Để đảm bảo sự khác biệt của các kết quả thực nghiệm của các hệ phân lớp được so sánh là có ý nghĩa, phương pháp kiểm định giả thuyết thống kê Wilcoxon [16] được sử dụng để kiểm tra giả thuyết H_0 (null hypothesis) có độ tin cậy là 90% ($\alpha = 0,1$) với giả định rằng các kết quả của các phương pháp được so sánh là tương đương nhau.

Nhằm giảm không gian tìm kiếm trong quá trình huấn luyện, các ràng buộc về giá trị của các tham số ngữ nghĩa được áp dụng như sau: số gia tử âm và số gia tử dương là 1, gia tử âm là "*Less*" (L) và gia tử dương là "*Very*" (V); $1 \leq k_j \leq 3$; $0,2 \leq \left\{fm\left(c_j^-\right), fm\left(c_j^+\right)\right\} \leq 0,7$; $0,00001 \leq \{fm(\mathbf{0}_j), fm(\mathbf{1}_j)\} \leq 0,01$; $0,0001 \leq fm(W_j) \leq 0,2$; $fm\left(0_j\right) + fm\left(c_j^-\right) + fm\left(W_j\right) + fm\left(c_j^+\right) + fm\left(1_j\right) = 1$; $0,2 \leq \{\mu(L_j), \mu(V_j)\} \leq 0,7; 0,01 \leq \mu(h_{0,j}) \leq 0,5$; and $\mu(L_j) + \mu(V_j) + \mu(h_{0,j}) = 1$.

Thuật toán tối ưu bầy đàn đa mục tiêu (PSO) [17] được sử dụng cho các bài toán tối ưu. Trong tối ưu các tham số ngữ nghĩa: số thế hệ là 250; số cá thể mỗi thế hệ là 600; hệ số Inertia là 0.4; hệ số nhận thức cá nhân là 0,2; hệ số nhận thức xã hội là 0,2; số luật khởi tạo bằng số thuộc tính; độ dài tối đa của luật là 1. Trong tối ưu hệ luật: số thế hệ là 1500; số luật khởi tạo là $|S_0| = 300 \times$ số lớp; độ dài tối đa của luật là 3. Phương pháp lập luận phân lớp được sử dụng trong tất cả các thực nghiệm là *single winner rule* [3, 4], tiêu chuẩn sàng luật là $c \times s$ và trọng số luật được tính toán theo công (7).

2. Kết quả thực nghiệm

Ký hiệu hệ phân lớp với ngữ nghĩa tính toán dựa trên tập mờ dạng hàm S với phân hoạch k_0 [11] và phân hoạch k_1 [10] tương ứng là **FRBC_S_k0** và **FRBC_S**, hệ phân lớp với ngữ nghĩa tính toán dựa trên tập mờ hình thang với phân hoạch k_0 và phân hoạch k_1 tương ứng là **FRBC_TRA_k0** và **FRBC_TRA**, hệ phân lớp với ngữ nghĩa tính toán dựa trên tập mờ tam giác trong [9] là **FRBC_TRI**. Bảng I thể hiện các kết quả thực nghiệm và so sánh giữa các hệ phân lớp nêu trên, trong đó chữ đậm thể hiện kết quả tốt hơn so với các hệ phân lớp còn lại. Ký hiệu #R×C là độ phức tạp của hệ phân lớp (tích của số luật trung bình và số điều kiện trung bình của các luật), P_{te} là độ chính xác phân lớp trung bình trên tập kiểm tra.

Các kết quả thực nghiệm trong Bảng I cho thấy, hệ phân lớp **FRBC_S_k0** có độ chính xác phân lớp trên tập kiểm tra cao hơn so với các hệ phân lớp **FRBC_S**, **FRBC_TRA_k0**, **FRBC_TRA** và **FRBC_TRI** tương ứng đối với 15, 18, 17 và 20 trong số 23 tập dữ liệu được thực nghiệm. So sánh dựa trên độ chính xác phân lớp trung bình của 23 tập dữ liệu được thực nghiệm, hệ phân lớp **FRBC_S_k0** có độ chính xác phân lớp trung bình là 83,04%, cao nhất so với các hệ phân lớp còn lại. So sánh dựa trên độ phức tạp của hệ phân lớp, các hệ phân lớp không có sự chênh lệch nhiều. Ngoài ra, hệ phân lớp **FRBC_S** có độ chính xác phân lớp trung bình là 82,79%, cao hơn so với hệ phân lớp **FRBC_TRA** và **FRBC_TRI** lần lượt có độ chính xác phân lớp trung bình là 82,67% và 81,92%. Các kết quả kiểm định giả thuyết thống kê Wilcoxon [16] với độ tin cậy 90% ($\alpha = 0,1$) sử dụng dữ liệu trong Bảng I với giả thiết độ chính xác phân lớp và độ phức tạp tương ứng của hai hệ phân lớp là tương đương nhau (Giả thuyết H_0) được thể hiện trong Bảng II và Bảng III. Các giá trị *Exact p-value* trong Bảng II đều nhỏ hơn $\alpha = 0,1$ cho biết rằng giả thuyết tương đương H_0 về độ chính xác phân lớp giữa các hệ phân lớp được so sánh bị bác bỏ. Điều này có nghĩa là hệ phân lớp **FRBC_S_k0** có độ chính xác phân lớp cao hơn so với các hệ phân lớp **FRBC_S**, **FRBC_TRA_k0**, **FRBC_TRA** và **FRBC_TRI**; hệ phân lớp **FRBC_S** có độ chính xác phân lớp cao hơn so

Bảng I
KẾT QUẢ THỰC NGHIỆM CỦA CÁC HỆ PHÂN LỚP **FRBC_S_K0**, **FRBC_S**, **FRBC_TRA_K0**, **FRBC_TRA** VÀ **FRBC_TRI**

STT	**Tập dữ liệu**	**FRBC_S_k0**		**FRBC_S**		**FRBC_TRA_k0**		**FRBC_TRA**		**FRBC_TRI**	
		$\#R\times C$	P_{te}	$\#R\times C$	P_{te}	$\#R\times C$	P_{te}	$\#R\times C$	P_{te}	$\#R\times C$	P_{te}
1	Appendicitis	23,30	**88,73**	17,35	88,48	19,90	88,64	**16,77**	88,15	21,32	87,55
2	Australian	46,23	**87,54**	**35,93**	87,25	46,16	87,49	46,50	87,15	36,20	86,38
3	Bands	59,40	73,00	55,80	73,40	61,80	72,95	58,20	**73,46**	**52,20**	72,80
4	Bupa	**177,72**	72,03	221,65	72,19	186,05	71,97	181,19	**72,38**	187,20	68,09
5	Cleveland	509,54	61,73	**433,16**	61,86	703,17	61,14	468,13	**62,39**	657,43	62,19
6	Dermatology	240,11	**96,26**	254,98	94,50	216,50	96,17	**182,84**	94,40	198,05	96,07
7	Glass	467,18	**72,97**	364,08	72,30	400,20	72,32	474,29	72,24	**343,60**	72,09
8	Haberman	12,00	77,42	16,00	**77,43**	12,00	77,41	10,80	77,40	**10,20**	75,76
9	Hayes-roth	117,14	**85,21**	136,65	84,36	128,44	84,58	**114,66**	84,17	122,27	84,17
10	Heart	117,24	84,94	**95,25**	84,69	124,75	**85,43**	123,29	84,57	122,72	84,44
11	Hepatitis	26,10	**91,22**	36,63	89,99	25,95	**91,22**	**25,53**	89,28	26,16	88,44
12	Ionosphere	98,81	**92,32**	92,83	91,65	96,91	92,22	**88,03**	91,56	90,33	90,22
13	Iris	**16,52**	**98,00**	17,76	97,33	21,73	97,78	30,37	97,33	26,29	96,00
14	Mammogr.	77,87	**84,36**	76,84	84,25	**49,67**	84,33	73,84	84,2	92,25	84,20
15	Newthyroid	44,55	**96,59**	49,98	95,84	41,50	96,00	**39,82**	95,67	45,18	94,42
16	Pima	62,11	76,45	**47,55**	**77.17**	57,70	77,09	56,12	77,01	60,89	76,18
17	Saheart	95,24	**71,07**	68,13	70,42	89,79	70,71	**59,28**	70,05	86,75	69,33
18	Sonar	59,29	77,98	62,32	**79,43**	53,86	77,95	**49,31**	78,61	79,76	76,80
19	Tae	**163,80**	61,22	176,48	**61,44**	176,06	61,43	210,70	61,00	261,00	59,47
20	Vehicle	177,29	68,48	207,91	**68,88**	**163,80**	68,41	195,07	68,20	242,79	67,62
21	Wdbc	27,88	96,19	35,85	95,90	28,00	96,72	**25,04**	96,78	37,35	**96,96**
22	Wine	36,73	**98,87**	46,79	98,51	36,37	98,50	40,39	98,49	**35,82**	98,30
23	Wisconsin	91,27	97,34	73,66	96,80	79,82	97,05	**69,81**	**96,95**	74,36	96,74
Trung bình		**119,45**	**83,04**	**114,07**	**82,79**	**122,61**	**82,94**	**114,78**	**82,67**	**126,53**	**81,92**

Bảng II
SO SÁNH ĐỘ CHÍNH XÁC GIỮA CÁC HỆ PHÂN LỚP **FRBC_S_K0**, **FRBC_S**, **FRBC_TRA_K0**, **FRBC_TRA** VÀ **FRBC_TRI** BẰNG PHƯƠNG PHÁP KIỂM ĐỊNH WILCOXON VỚI $\alpha = 0,1$

So sánh (α = 0,1)	R^+	R^-	Exact P-value	Giả thuyết H_0
FRBC_S_k0 vs **FRBC_S**	196,0	80,0	0,0802	Bị bác bỏ
FRBC_S_k0 vs **FRBC_TRA_k0**	188,0	65,0	0,04616	Bị bác bỏ
FRBC_S_k0 vs **FRBC_TRA**	208,0	68,0	0,03266	Bị bác bỏ
FRBC_S vs **FRBC_TRA**	188,5	64,5	0,04433	Bị bác bỏ
FRBC_S vs **FRBC_TRI**	240,0	36,0	0,0011184	Bị bác bỏ

Bảng III
SO SÁNH ĐỘ PHỨC TẠP GIỮA CÁC HỆ PHÂN LỚP **FRBC_S_K0**, **FRBC_S**, **FRBC_TRA_K0**, **FRBC_TRA** VÀ **FRBC_TRI** BẰNG PHƯƠNG PHÁP KIỂM ĐỊNH WILCOXON VỚI $\alpha = 0,1$.

So sánh (α = 0,1)	R^+	R^-	Exact P-value	Giả thuyết H_0
FRBC_S_k0 vs **FRBC_S**	133,0	143,0	$\geq 0,2$	Không bị bác bỏ
FRBC_S_k0 vs **FRBC_TRA_k0**	126,0	150,0	$\geq 0,2$	Không bị bác bỏ
FRBC_S_k0 vs **FRBC_TRA**	99,0	177,0	$\geq 0,2$	Không bị bác bỏ
FRBC_S vs **FRBC_TRA**	115,0	161,0	$\geq 0,2$	Không bị bác bỏ
FRBC_S vs **FRBC_TRI**	161,0	115,0	$\geq 0,2$	Không bị bác bỏ

với hai hệ phân lớp sử dụng cùng cấu trúc phân hoạch k_1 là **FRBC_TRA** và **FRBC_TRI**. Các giá trị *Exact p-value* trong Bảng III đều lớn hơn $\alpha = 0,1$ nên giả thuyết tương đương H_0 về độ phức tạp của các hệ phân lớp không bị bác bỏ. Do đó, ta có thể khẳng định rằng, với cùng một cách biểu diễn phân hoạch mờ thì các hệ phân lớp dựa trên luật mờ với ngữ nghĩa tính toán dựa trên tập mờ của các từ ngôn ngữ có dạng hàm S được sinh bởi ĐSGT mở rộng cho độ chính xác phân lớp cao hơn so với dạng hình tam giác và hình thang do hàm S biểu diễn sự biến thiên về ngữ nghĩa tốt hơn. Ngoài ra, cấu trúc phân hoạch k_0 cho hiệu suất phân lớp tốt hơn cấu trúc phân hoạch k_1 đồng thời đảm bảo tính giải nghĩa được của hệ phân lớp như đã được chứng minh trong [11].

Nhằm thể hiện tính hiệu quả của hệ phân lớp với ngữ nghĩa tính toán dựa trên tập mờ dạng hàm S được sinh bởi ĐSGT mở rộng được đề xuất so với tiếp cận lý thuyết tập mờ, các kết quả thực nghiệm của hệ phân lớp **FRBC_S** được so sánh với các kết quả của hai hệ phân lớp **PAES-RCS** và **FURIA** [2]. Kết quả so sánh trong Bảng IV cho thấy, hệ phân lớp **FRBC_S** cho độ chính xác phân lớp trên

Bảng IV
KẾT QUẢ THỰC NGHIỆM CỦA HỆ PHÂN LỚP **FRBC_S**, **PAES-RCS** VÀ **FURIA**

STT	**Tập dữ liệu**	**FRBC_S**		**PAES-RCS**		$\neq P_{te}$	$\neq R\times C$	**FURIA**		$\neq P_{te}$	$\neq R\times C$
		$\#R\times C$	P_{te}	$\#R\times C$	P_{te}			$\#R\times C$	P_{te}		
1	Appendicitis	17,35	88,48	35,28	85,09	3,39	-17,93	19,00	85,18	3,30	-1,65
2	Australian	35,93	87,25	329,64	85,80	1,45	-293,71	89,60	85,22	2,03	-53,67
3	Bands	55,80	73,40	756,00	67,56	5,84	-700,20	535,15	64,65	8,75	-479,35
4	Bupa	221,65	72,19	256,20	68,67	3,52	-34,55	324,12	69,02	3,17	-102,47
5	Cleveland	433,16	61,86	1140,00	59,06	2,80	-706,84	134,67	56,20	5,66	298,49
6	Dermatology	254,98	94,50	389,40	95,43	-0,93	-134,42	303,88	95,24	-0,74	-48,90
7	Glass	364,08	72,30	487,90	72,13	0,17	-123,82	474,81	72,41	-0,11	-110,73
8	Haberman	16,00	77,43	202,41	72,65	4,78	-186,41	22,04	75,44	1,99	-6,04
9	Hayes-roth	136,65	84,36	120,00	84,03	0,33	16,65	188,10	83,13	1,23	-51,45
10	Heart	95,25	84,69	300,30	83,21	1,48	-205,05	193,64	80,00	4,69	-98,39
11	Hepatitis	36,63	89,99	300,30	83,21	6,78	-263,67	52,38	84,52	5,47	-15,75
12	Ionosphere	92,83	91,65	670,63	90,40	1,25	-577,80	372,68	91,75	-0,10	-279,85
13	Iris	17,76	97,33	69,84	95,33	2,00	-52,08	31,95	94,66	2,67	-14,19
14	Mammogr.	76,84	84,25	132,54	83,37	0,88	-55,70	16,83	83,89	0,36	60,01
15	Newthyroid	49,98	95,84	97,75	95,35	0,49	-47,77	100,82	96,30	-0,46	-50,84
16	Pima	47,55	77.17	270,64	74,66	2,51	-223,09	127,50	74,62	2,55	-79,95
17	Saheart	68,13	70,42	525,21	70,92	-0,50	-457,08	50,88	69,69	0,73	17,25
18	Sonar	62,32	79,43	524,60	77,00	2,43	-462,28	309,96	82,14	-2,71	-247,64
19	Tae	176,48	61,44	323,14	60,81	0,63	-146,66	43,00	43,08	18,36	133,48
20	Vehicle	207,91	68,88	555,77	64,89	3,99	-347,86	2125,97	71,52	-2,64	-1918,06
21	Wdbc	35,85	95,90	183,70	95,14	0,76	-147,85	356,12	96,31	-0,41	-320,27
22	Wine	46,79	98,51	170,94	93,98	4,53	-124,15	80,00	96,60	1,91	-33,21
23	Wisconsin	73,66	96,80	328,02	96,46	0,34	-254,36	521,10	96,35	0,45	-447,44
Trung bình		**114,07**	**82,79**	**355,23**	**80,66**			**281,49**	**80,34**		

Bảng V
SO SÁNH ĐỘ CHÍNH XÁC CỦA HỆ PHÂN LỚP **FRBC_S** SO VỚI **PAES-RCS** VÀ **FURIA** BẰNG PHƯƠNG PHÁP KIỂM ĐỊNH WILCOXON VỚI $\alpha = 0,1$

So sánh (α = 0,1)	R^+	R^-	Exact P-value	Giả thuyết H_0
FRBC_S vs **PAES-RCS**	275,0	1,0	2,622E-5	Bị bác bỏ
FRBC_S vs **FURIA**	227,0	49,0	0,005414	Bị bác bỏ

Bảng VI
SO SÁNH ĐỘ PHỨC TẠP CỦA HỆ PHÂN LỚP **FRBC_S** SO VỚI **PAES-RCS** VÀ **FURIA** BẰNG PHƯƠNG PHÁP KIỂM ĐỊNH WILCOXON VỚI $\alpha = 0,1$

So sánh (α = 0,1)	R^+	R^-	Exact P-value	Giả thuyết H_0
FRBC_S vs **PAES-RCS**	275,0	1,0	4,768E-7	Bị bác bỏ
FRBC_S vs **FURIA**	225,0	51,0	0,00671	Bị bác bỏ

tập kiểm tra cao hơn hệ phân lớp **PAES-RCS** và **FURIA** lần lượt là 21 và 15 trên 23 tập dữ liệu được thử nghiệm. Xét trên giá trị trung bình của độ chính xác phân lớp, hệ phân lớp **FRBC_S** có giá trị trung bình là 82,79%, cao hơn lần lượt là 2,13% và 2,45% so với hệ phân lớp **PAES-RCS** và **FURIA** có giá trị trung bình lần lượt là 80,66% và 80,34%. Phân tích trên độ phức tạp của hệ phân lớp, hệ phân lớp **FRBC_S** có độ phức tạp phân lớp thấp hơn rất nhiều so với hai hệ phân lớp còn lại, tương ứng là 114,07 so với 355,23 và 281,49.

Các kết quả kiểm định giả thuyết thống kê Wilcoxon với độ tin cậy 90% ($\alpha = 0,1$) sử dụng dữ liệu trong Bảng IV đối với độ chính xác phân lớp và độ phức tạp của hệ phân lớp được thể hiện tương ứng trong Bảng V và Bảng VI. Ta thấy rằng, các giá trị giá trị *Exact p-value* đều nhỏ hơn $\alpha = 0,1$ nên giả thuyết tương đương về độ chính xác phân lớp và độ phức tạp của hệ phân lớp của **FRBC_S** tương ứng so với hai hệ phân lớp được đối sánh **PAES-RCS** và **FURIA** bị bác bỏ. Do đó, ta có thể khẳng định rằng hệ phân lớp **FRBC_S** tốt hơn hai hệ phân lớp còn lại trên cả hai tiêu chí độ chính xác phân lớp và độ phức tạp của hệ phân lớp. Do hệ phân lớp **FRBC_S_k0** sử dụng cấu trúc phân hoạch k_0 tốt hơn so với hệ phân lớp **FRBC_S** sử dụng phân hoạch k_1 như đã được so sánh ở trên nên ta có thể kết luận rằng hệ phân lớp **FRBC_S_k0** tốt hơn hai hệ phân lớp **PAES-RCS** và **FURIA**.

IV. KẾT LUẬN

Ngữ nghĩa định tính của các từ ngôn ngữ trong cơ sở luật của hệ phân lớp dựa trên luật mờ không dùng để tính

toán được. Do đó, việc biểu diễn ngữ nghĩa tính toán phù hợp với ngữ nghĩa định tính của các từ ngôn ngữ đóng vai trò quan trọng. Bài báo này trình bày phương pháp biểu diễn ngữ nghĩa tính toán dựa trên tập mờ dạng hàm S được sinh ra bởi ĐSGTMR cho các từ ngôn ngữ được sử dụng để biểu diễn cấu trúc phân hoạch đa thể hạt dạng k_0 và k_1. Các kết quả thực nghiệm và kiểm định giả thuyết thống kê Wilcoxon cho thấy tính hiệu quả của các phương pháp được đề xuất khi áp dụng cho hệ phân lớp dựa trên luật mờ.

LỜI CẢM ƠN

Nghiên cứu này được tài trợ bởi Trường Đại học Giao thông vận tải trong đề tài mã số T2020-CN-002.

TÀI LIỆU THAM KHẢO

[1] R. Alcalá, Y. Nojima, F. Herrera, H. Ishibuchi, "Multi-objective genetic fuzzy rule selection of single granularity-based fuzzy classification rules and its interaction with the lateral tuning of membership functions," *Soft Computing*, vol. 15, no. 12, pp. 2303–2318, 2011.

[2] M. Antonelli, P. Ducange, F. Marcelloni, "A fast and efficient multi-objective evolutionary learning scheme for fuzzy rule-based classifiers," *Information Sciences*, vol. 283, pp. 36–54, 2014.

[3] H. Ishibuchi, T. Yamamoto, "Fuzzy Rule Selection by Multi-Objective Genetic Local Search Algorithms and Rule Evaluation Measures in Data Mining," *Fuzzy Sets and Systems*, vol. 141, no. 1, pp. 59-88, 2014.

[4] H. Ishibuchi, T. Yamamoto, "Rule weight specification in fuzzy rule-based classification systems," *IEEE Transactions on Fuzzy Systems*, vol. 13, no. 4, pp. 428–435, 2005.

[5] M. I. Rey, M. Galende, M. J. Fuente, G. I. Sainz-Palmero, "Multi-objective based Fuzzy Rule Based Systems (FRBSs) for trade-off improvement in accuracy and interpretability: A rule relevance point of view," *Knowledge-Based Systems*, vol. 127, pp. 67–84, 2017.

[6] N. C. Ho, W. Wechler, "Hedge algebras: an algebraic approach to structures of sets of linguistic domains of linguistic truth variables," *Fuzzy Sets and Systems*, vol. 35, no. 3, pp. 281-293, 1990.

[7] N. C. Ho, W. Wechler, "Extended hedge algebras and their application to fuzzy logic," *Fuzzy Sets and Systems*, vol. 52, pp. 259–281, 1992.

[8] N. C. Ho, N. V. Long, "Fuzziness measure on complete hedges algebras and quantifying semantics of terms in linear hedge algebras," *Fuzzy Sets and Systems*, vol. 158, pp. 452-471, 2007.

[9] N. C. Ho, W. Pedrycz, D. T. Long, T. T. Son, "A genetic design of linguistic terms for fuzzy rule based classifiers," *International Journal of Approximate Reasoning*, vol 54, no. 1, pp. 1-21, 2013.

[10] N. C. Ho, T. T. Son, P. D. Phong, "Modeling of a semantics core of linguistic terms based on an extension of hedge algebra semantics and its application," *Knowledge-Based Systems*, vol. 67, pp. 244–262, 2014.

[11] N. C. Ho, H. V. Thong, N. V. Long, "A discussion on interpretability of linguistic rule based systems and its application to solve regression problems," *Knowledge-Based Systems*, vol. 88, pp. 107–133, 2015.

[12] T. T. Son, N. T. Anh, "Partition fuzzy domain with multi-granularity representation of data based on hedge algebra approach," *Journal of Computer Science and Cybernetics*, vol. 34, no. 1, pp. 63–75, 2018.

[13] B. H. Le, L. T. Anh, B. V. Binh, "Explicit formula of hedge-algebras-based fuzzy controller and applications in structural vibration control," *Applied Soft Computing*, vol. 60, pp. 150–166, 2017.

[14] N. H. Huy, N. C. Ho, N. V. Quyen, "Multichannel image contrast enhancement based on linguistic rule-based intensificators," *Applied Soft Computing Journal*, vol. 76, pp. 744–762, 2019.

[15] D. T. Long, "A genetic algorithm based method for timetabling problems using linguistics of hedge algebra in constraints," *Journal of Computer Science and Cybernetics*, vol. 32, no. 4, pp. 285–301, 2016.

[16] J. Demˇsar, "Statistical Comparisons of Classifiers over Multiple Data Sets," *Journal of Machine Learning Research*, vol. 7, pp. 1–30, 2006.

[17] P. D. Phong, N. C. Ho, N. T. Thuy, "Multi-objective Particle Swarm Optimization Algorithm and its Application to the Fuzzy Rule Based Classifier Design Problem with the Order Based Semantics of Linguistic Terms," *In Proceedings of The 10th IEEE RIVF International Conference on Computing and Communication Technologies (RIVF-2013), Hanoi, Vietnam*, pp. 12–17, 2013.

[18] F. Rudzinski, "A multi-objective genetic optimization of interpretability-oriented fuzzy rule-based classifiers," *Applied Soft Computing*, vol. 38, pp. 118–133, 2016.

[19] M. Elkanoa, M. Galara, J. Sanza, H. Bustince, "CHI-BD: A fuzzy rule-based classification system for Big Data classification problems," *Fuzzy Sets and Systems*, vol. 348, pp. 75–101, 2018.

[20] M. Pota, M. Esposito, G. D. Pietro, "Designing rule-based fuzzy systems for classification in medicine," *Knowledge-Based Systems*, vol. 124, pp. 105–132, 2017.

[21] M. Soui, I. Gasmi, S. Smiti, K. Ghédira, "Rule-based credit risk assessment model using multi-objective evolutionary algorithms," *Expert Systems With Applications*, vol. 126, pp. 144–157, 2019.

[22] Y. Zhang, X. Qian, J. Wang, M. Gendeel1, "Fuzzy rule-based classification system using multi-population quantum evolutionary algorithm with contradictory rule reconstruction," *Applied Intelligence*, vol. 49, pp. 4007–4021, 2019.

SƠ LƯỢC VỀ TÁC GIẢ

Nguyễn Đức Dư

Nhận bằng Cử nhân Toán tin ứng dụng, Thạc sĩ Toán ứng dụng tại Trường Đại học khoa học tự nhiên, Đại học Quốc gia Hà Nội lần lượt các năm 2001, 2005. Hiện là giảng viên Khoa Công nghệ thông tin, Trường Đại học Giao thông vận tải. Lĩnh vực nghiên cứu: khai phá dữ liệu, lô gic mờ, hệ mờ, tính toán mềm, tính toán với từ, học máy.

Phạm Đình Phong

Nhận bằng Thạc sĩ Công nghệ thông tin và Tiến sĩ Khoa học máy tính tại Trường Đại học Công nghệ, Đại học Quốc gia Hà Nội lần lượt các năm 2011, 2018. Hiện là giảng viên Khoa Công nghệ thông tin, Trường Đại học Giao thông vận tải. Lĩnh Vực nghiên cứu: khai phá dữ liệu, các hệ mờ, tính toán mềm, học máy.

Phạm Đình Vũ

Nhận bằng kỹ sư Công nghệ thông tin tại Trường Đại học Bách khoa Hà Nội năm 2003, Thạc sỹ Hệ thống thông tin Học viện Công nghệ Bưu chính Viễn thông năm 2015. Hiện đang công tác tại Cục Công nghệ thông tin và Thống kê hải quan, Tổng cục Hải quan. Lĩnh vực nghiên cứu: khai phá dữ liệu, các hệ mờ, tính toán mềm, học máy.

Nguyễn Đức Thảo

Nhận bằng Kỹ sư, Thạc sĩ và Tiến sỹ Công nghệ thông tin lần lượt tại Trường Đại học Tổng hợp Nga năm 1996, 2001. Hiện là cán bộ nghiên cứu tại Viện Khoa học và Công nghệ quân sự/ Bộ Quốc phòng. Lĩnh vực nghiên cứu: khai phá dữ liệu, lô gic mờ, hệ mờ, tính toán mềm, tính toán với từ, học máy, trí tuệ nhân tạo, hệ thống thông tin, hệ chuyên gia.