Transport and Communications Science Journal, Vol 75, Issue 03 (04/2024), 1335-1347

Transport and Communications Science Journal

A METHOD FOR DESIGNING FUZZY RULE-BASED CLASSIFIER USING S-FUNCTION BASED FUZZY SET GUARANTEE INTERPRETABILITY

Duc Du Nguyen

University of Transport and Communications, No 3 Cau Giay Street, Hanoi, Vietnam

ARTICLE INFO

TYPE: Research Article
Received: 10/03/2024
Revised: 03/04/2024
Accepted: 05/04/2024
Published online: 15/04/2024
https://doi.org/10.47869/tcsj.75.3.2

* *Corresponding author*
Email: nducdu@utc.edu.vn; Tel: 0912363245

Abstract. Fuzzy rule-based systems have many practical applications. However, to extract a compact, highly accurate rule-based system and ensure its interpretability requires different methodologies and techniques, including methods of representing the semantics of linguistic words in the rule bases. Hedge algebras is utilized to create a formal basis for designing computational fuzzy-set-based semantics of linguistic words from their inherent semantics. Many studies on computational fuzzy-set-based semantic representation methods to ensure the interpretability of fuzzy rule-based systems have been proposed and applied to solve classification and regression problems. Those studies also showed that the shape of the fuzzy set affects the accuracy of the fuzzy rule-based systems. A method to design a multi-semantic structure with a fuzzy set of the form *S* function to ensure the interpretability of the classification systems according to Tarski's point is presented in this paper. Experimental results with 15 real standard data sets show that the proposed method obtains better classification accuracy while not increasing the complexity of the rule bases compared to the ones of the existing methods.

Keywords: Hedge algebras, Order-based semantics, Classifier, Interpretability, Fuzzy rule-based systems.

Tạp chí Khoa học Giao thông vận tải

MỘT PHƯƠNG PHÁP THIẾT KẾ HỆ PHÂN LỚP DỰA TRÊN LUẬT MỜ SỬ DỤNG TẬP MỜ DẠNG HÀM S ĐẢM BẢO TÍNH GIẢI NGHĨA ĐƯỢC

Nguyễn Đức Dư

Trường Đại học Giao thông vận tải, Số 3 Cầu Giấy, Hà Nội, Việt Nam

THÔNG TIN BÀI BÁO

CHUYÊN MỤC: Công trình khoa học
Ngày nhận bài: 10/03/2024
Ngày nhận bài sửa: 03/04/2024
Ngày chấp nhận đăng: 05/04/2024
Ngày xuất bản Online: 15/04/2024
https://doi.org/10.47869/tcsj.75.3.2

** Tác giả liên hệ*
Email: nducdu@utc.edu.vn; Tel: 091233245

Tóm tắt. Hệ dựa trên luật mờ có nhiều ứng dụng trong thực tiễn. Tuy nhiên, để trích rút được hệ luật ngắn gọn, có độ chính xác cao và đảm bảo tính giải nghĩa cần có phương pháp luận và kỹ thuật khác nhau, trong đó có phương pháp biểu diễn ngữ nghĩa của các từ ngôn ngữ trong cơ sở luật. Đại số gia tử cho phép tạo ra một cơ sở hình thức thiết kế ngữ nghĩa tính toán dựa trên tập mờ của các từ ngôn ngữ trong cơ sở luật từ ngữ nghĩa vốn có của chúng. Đã có nhiều nghiên cứu về phương pháp biểu diễn ngữ nghĩa tính toán dựa trên tập mờ đảm bảo tính giải nghĩa của hệ dựa trên luật mờ được đề xuất và được ứng dụng giải bài toán phân lớp và hồi quy. Các nghiên cứu đó cũng chỉ ra rằng, hình dạng của tập mờ ảnh hưởng đến độ chính xác hệ dựa trên luật mờ. Một phương pháp thiết kế cấu trúc đa ngữ nghĩa với tập mờ dạng hàm *S* đảm bảo tính giải nghĩa của hệ phân lớp theo qua điểm của Tarski được trình bày trong bài báo này. Kết quả thực nghiệm với 15 tập dữ liệu chuẩn phát sinh trong thực tiễn cho thấy phương pháp được đề xuất cho độ chính xác phân lớp tốt hơn trong khi không làm tăng độ phức tạp của hệ luật so với các phương pháp đã được công bố.

Từ khóa: Đại số gia tử, Thứ tự ngữ nghĩa, Hệ phân lớp, Tính giải nghĩa được, Hệ dựa trên luật mờ.

1. ĐẶT VẤN ĐỀ

Các hệ dựa trên luật mờ (Fuzzy rule-based systems – FRBS) với ngữ nghĩa của các từ ngôn trong cơ sở luật được biểu diễn bằng các tập mờ là một trong các công cụ được dùng để

mô phỏng khả năng lập luận của con người và được ứng dụng để giải nhiều bài toán ứng dụng thực tế như phân lớp [1-4], hồi quy [5-8], y tế [9], ... Một trong các mục tiêu quan trọng của các FRBS là tính giải nghĩa được, tức là chuyên gia có thể đọc và hiểu được hệ luật mờ, từ đó có thể sử dụng nó để lập luận như những tri thức mà họ có.

Một trong những khả năng đặc biệt của con người là xử lý trực tiếp trên tri thức ngôn ngữ của họ để giải một bài toán thực tế. Để mô phỏng khả năng của con người trong việc xử lý tính toán trực tiếp các từ của ngôn ngữ, chúng ta cần phải thiết lập một cấu trúc tính toán thích hợp trong đó các đối tượng tính toán của các biến có thể được coi như là ngữ nghĩa tính toán của các từ của các FRBS. Tuy nhiên, các FRBS được thiết kế theo hướng tiếp cận lý thuyết tập mờ do không có cơ sở hình thức để đảm bảo rằng các tập hợp mờ đó biểu diễn chính xác ngữ nghĩa của các từ ngôn ngữ được gán cho chúng, nhất là sau quá trình hiệu chỉnh các tham số của các hàm thuộc, do đó chúng không được xem là các công cụ có thể xử lý trực tiếp trên các từ ngôn ngữ. Vì vậy, chúng vẫn chưa thể mô phỏng chính xác cách mà các chuyên gia lập luận hay nói các khác là chúng khó giải nghĩa được. Do đó, Mencar và Fanelli đã đưa ra một số ràng buộc mức phân hoạch mờ và cơ sở luật để đảm tính giải nghĩa được [10].

Phương pháp luận tính toán trực tiếp trên các từ ngôn ngữ theo tiếp cận Đại số gia tử [11-13] để phát triển các thuật toán tiến hóa thiết kế các hệ dựa trên luật mờ cho bài toán hồi quy [6, 7] và bài toán phân lớp, được gọi là hệ phân lớp dựa trên luật mờ (Fuzzy rule-based classifier - FRBC) có tính giải nghĩa được theo quan điểm của Tarski [14] được đề xuất trong [15]. Theo phương pháp luận này, khi thiết kế các FRBC cần có một cơ chế hình thức để xác định ngữ nghĩa tính toán từ ngữ nghĩa định tính vốn có của các từ ngôn ngữ [15], tức là các cấu trúc đa thể hạt mờ phải là hình ảnh đẳng cấu của cấu trúc đa ngữ nghĩa của tập từ tương ứng của các thuộc tính. Để đáp ứng đòi hỏi này thì các cấu trúc phân hoạch mờ biểu diễn cấu trúc đa ngữ nghĩa của các từ ngôn ngữ của các biến ngôn ngữ phải giải nghĩa được [6, 15]. Ngữ nghĩa tính toán dựa trên tập mờ trong nghiên cứu [6, 7, 16] có hàm thuộc dạng hình thang có ưu điểm so với hình tam giác [17] 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. Ngữ nghĩa tính toán dựa trên tập mờ dạng hàm *S* lần đầu được đề xuất trong [18] và được ứng dụng cho bài toán hồi quy và bài toán phân lớp [19]. 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. Tuy nhiên, trong các nghiên cứu [18, 19], cấu trúc phân hoạch mờ đa thể hạt với tập mờ dạng hàm *S* chưa đảm bảo tính giải nghĩa theo quan điểm của Tarski [14, 15]. Phương pháp thiết kế cấu trúc đa ngữ nghĩa với ngữ nghĩa tính toán dựa trên tập mờ dạng hàm *S* cho các FRBC đảm bảo tính giải nghĩa theo quan điểm của Tarski được trình bày trong bài báo này. Các thực nghiệm đối với phương pháp thiết kế FRBC được đề xuất được tiến hành với 15 tập dữ liệu phân lớp được lấy từ nguồn KEEL-Dataset cho thấy tính hiệu quả của phương pháp thiết kế FRBC được đề xuất so với các phương pháp đã được công bố.

2. CẤU TRÚC NGỮ NGHĨA DỰA TRÊN TẬP MỜ CỦA CÁC TỪ NGÔN NGỮ

2.1 Khái niệm tính giải nghĩa được

Theo Tarski và các cộng sự [14], khái niệm tính giải nghĩa được trong toán học và logic được phát biểu như sau:

Lý thuyết S được gọi là có thể giải nghĩa được trong lý thuyết T nếu tồn tại một bản dịch $\mathcal{I}$ từ ngôn ngữ hình thức $\mathcal{L}(S)$ của S sang ngôn ngữ hình thức $\mathcal{L}(T)$ của T thỏa mãn điều kiện,

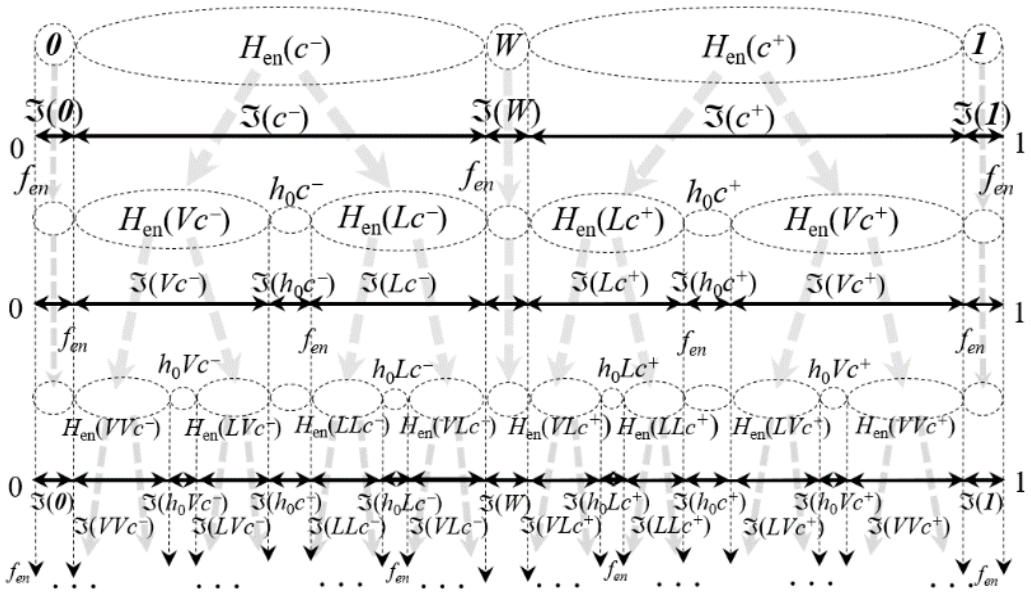
với mọi mệnh đề $p \in \mathcal{L}(S)$ thì p có thể chứng minh được trong S khi và chỉ khi $\mathcal{I}(p) \in \mathcal{L}(T)$ có thể chứng minh được trong T.

Khái niệm này cho biết rằng, thay vì giải một bài toán đã cho P_s trong lý thuyết S người ta có thể giải nó trong một lý thuyết T khác bằng cách biến đổi P_S sang T bằng phép biến đổi $\mathcal{J}$ khi và chỉ khi S có thể giải nghĩa được trong T bằng phép biến đổi $\mathcal{J}$. Như vậy theo khái niệm này, nếu lý thuyết T thỏa mãn điều kiện này thì T được gọi là có thể giải nghĩa được đối với S. Ký hiệu $\boldsymbol{S}^A = (X^A, \leq, g)$ là cấu trúc ngữ nghĩa miền từ X^A của thuộc tính A với các quan hệ thứ tự $\leq$ và quan hệ khái quát – đặc tả g. Như vậy, khi tính toán với từ thông qua các các tập mờ tương ứng của chúng là giải nghĩa được theo khái niệm của Tarski khi và chỉ khi các tập mờ tạo thành một cấu trúc là ảnh đẳng cấu của cấu trúc ngữ nghĩa $\boldsymbol{S}^A = (X^A, \leq, g)$. Khi đó cấu trúc tập mờ này có thể giải nghĩa được cho $\boldsymbol{S}^A$.

2.2 Cấu trúc đa mức của ngữ nghĩa định tính và định lượng của miền từ vô hạn của các thuộc tính

2.2.1. Đại số gia tử mở rộng biểu diễn lõi ngữ nghĩa của từ ngôn ngữ

Đại số gia tử (ĐSGT) mở rộng được cải tiến từ ĐSGT bằng việc bổ sung một gia tử nhân tạo h_0 nhằm mô hình hóa lõi ngữ nghĩa của các từ ngôn ngữ [16].



Hình 1. *Cấu trúc bụi ngữ nghĩa $\mathfrak{B}^A$ và các quan hệ của chúng* [6].

Cho một ĐSGT tuyến tính $\boldsymbol{\mathcal{A}}^A = (X^A, G, C, H, \leq)$ của một biến ngôn ngữ A. Một gia tử nhân tạo $h_0 \notin H$ được bổ sung để sinh lõi ngữ nghĩa của mỗi từ $x \in X^A$. Về mặt cú pháp, $h_0x \notin X^A$ và đặt $X^A_{en} = X^A \cup \{h_0x: x \in X^A\}$. Ta có ĐSGT mở rộng của $\boldsymbol{\mathcal{A}}^A$ là $\boldsymbol{\mathcal{A}}^A_{en} = (X^A_{en}, G, C, H_{en}, \leq)$, trong đó $H_{en} = H \cup \{h_0\}$, $X^A_{en} = C \cup H_{en}(G) = C \cup \{h_n \ldots h_1c : c \in G, h_j \in H_{en}$, với $j = 1$ to $n\}$. Vì vậy, $X^A = C \cup H(G) \subseteq X^A_{en} = C \cup H_{en}(G)$.

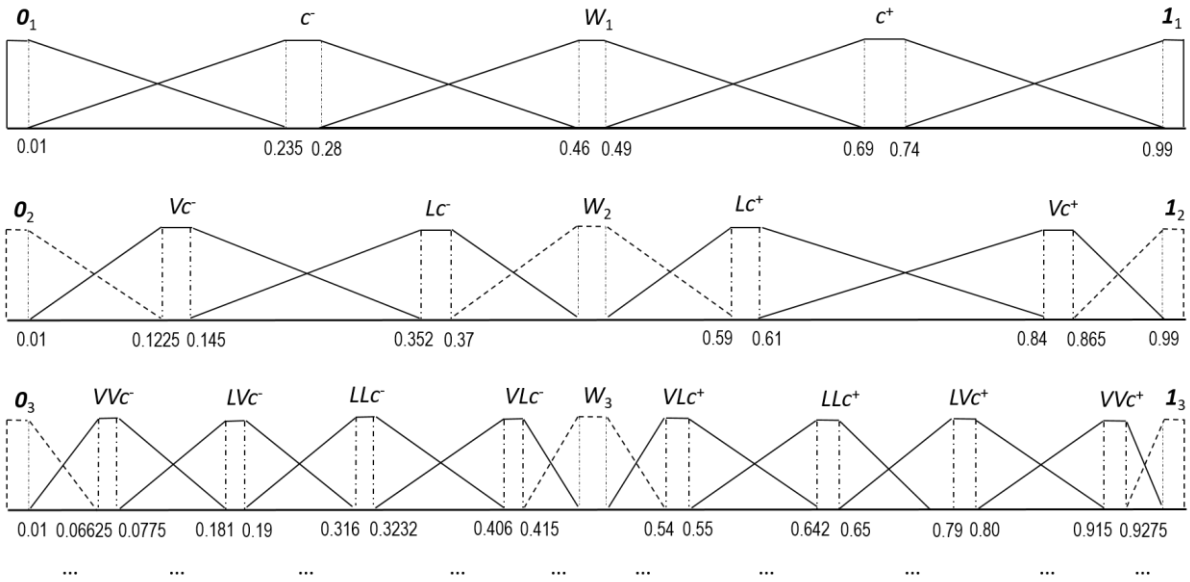
Đặt $X^A_{en,k} = \{x \in X^A_{en}: |x| = k\}$ với $|x|$ là ký hiệu độ dài của từ x, tập các từ có độ đặc tả mức k (k-specificity) và $X^A_{en,(k)} = \{x \in X^A_{en}: |x| \leq k\} = \bigcup_{1 \leq i \leq k} X^A_{en,i}$ – tập của các từ có mức đặc tả không lớn hơn k. Khi đó, $X^A_{en,1} = G \cup C$ và với mọi $k \geq 2$ thì $X^A_{en,k} = X^A_k \cup \{h_0u: u \in$

$X^{A}_{en,(k-1)}$}, tức là với $k > 0$, $X^{A}_{en,(k)}$ bao gồm tất cả các từ có mức đặc tả k, lõi ngữ nghĩa của chúng và tất cả các từ có mức đặc tả thấp hơn k.

Cấu trúc khái quát-đặc tả của biến ngôn ngữ A, $G^A = (X^A, g)$: Miền từ của X^A cũng bao hàm những cấu trúc ngữ nghĩa khác, quan hệ *khái quát-đặt tả*, tức là một từ x có tính khái quát hơn từ y và được ký hiệu bởi $g(x, y)$ và ngược lại, y được gọi là có tính đặc tả hơn x. Do đó, quan hệ g được gọi là quan hệ *khái quát-đặc tả* (*generality-specificity*).

Miền từ X^A bao gồm hai cấu trúc, cấu trúc ngữ nghĩa dựa trên thứ tự $\boldsymbol{S}^A = (X^A_{en}, \leq)$ và cấu trúc *khái quát-đặc tả*, $\boldsymbol{G}^A = (X^A_{en}, g)$, tức là một biến A không chỉ có một cấu trúc ngữ nghĩa mà còn nhiều vấn đề phức tạp ở trong nó. $\boldsymbol{G}^A = (X^A_{en}, g)$ được gọi là cấu trúc ngữ nghĩa đa mức và được biểu thị bằng $\mathcal{S}^A = (X^A_{en}, \leq, g)$. Để mô tả rõ ràng cấu trúc này, nghiên cứu này trình bày dưới dạng bụi đa mức như trong Hình 1 và nó được gọi là bụi ngữ nghĩa $\mathfrak{B}^A$ của $\boldsymbol{S}^A$. $\mathfrak{B}^A$ có thể được xây dựng như là một cấu trúc tiềm năng vô hạn. Mỗi nút của nó biểu diễn tính mờ của một từ ở mức đặc tả k. Gọi cấu trúc bao gồm tất cả các mức $l = 1$ đến k là k-section của bụi ngữ nghĩa $\mathfrak{B}^A$, ký hiệu là $\mathfrak{B}^A_\kappa$. Nó biểu diễn cấu trúc ngữ nghĩa của tập từ $X^A_{en,(k)}$.

2.2.2. Biểu diễn cấu trúc ngữ nghĩa dựa trên tập mờ của miền từ theo tiếp cận ĐSGT



Hình 2. Cấu trúc phân hoạch đa thể hình thang biểu diễn cấu trúc ngữ nghĩa $\mathcal{S}^A = (X^A, \leq, g)$ của biến A [6].

Muốn cấu trúc $\mathfrak{I}(X^A)$ biểu diễn cấu trúc $\boldsymbol{S}^A = (X^A, \leq, g)$ bảo toàn cấu trúc của $\boldsymbol{S}^A$ hay nói cách khác là $\mathfrak{I}(X^A)$ giải nghĩa được thì đòi hỏi định nghĩa hai quan hệ ký hiệu là $\leq$ và $\subset$ trên $\mathfrak{I}(X^A)$ vì $\boldsymbol{S}^A$ có các *quan hệ thứ tự* $\leq$ và *khái quát-đặc tả* g. Ký hiệu mỗi tập mờ hình thang là bộ ba $(a, \boldsymbol{b}, c)$, trong đó $a, c \in [0, 1]$, $\boldsymbol{b}$ là một khoảng con của $[0, 1]$ đóng vai trò là lõi của bộ ba và $a < \boldsymbol{b} < c$.

Định nghĩa 1. Với mọi tập mờ hình thang được xây dựng $\mathfrak{I}(X^A)$, định nghĩa:

1) Quan hệ thứ tự ≤ trên $\mathcal{I}(X^A)$: hai bộ ba t và t' với $t = (a, \boldsymbol{b}, c)$ và $t' = (a', \boldsymbol{b}', c')$ thỏa mãn $t \leq t'$ nếu và chỉ nếu các lõi của chúng thỏa mãn $\boldsymbol{b} = \boldsymbol{b}'$ hoặc $\boldsymbol{b} < \boldsymbol{b}'$ và thỏa ít nhất một trong các bất đẳng thức $a \leq a'$ và $c \leq c'$.

2) Quan hệ bao hàm ⊂ trên $\mathcal{I}(X^A)$: hai bộ ba t và t' ở trên được gọi là thỏa mãn $t \subset t'$ nếu và chỉ nếu đáy lớn của t được bao hàm trong đáy lớn của t', tức là $(a, c) \subset (a', c')$.

Tập $\mathcal{I}(X^A)$ với hai quan hệ ≤ và ⊂ được ký hiệu là $M_{Gr}^{A} = (\mathcal{I}(X^A), \leq, \subset)$, được gọi là cấu trúc đa thể hình thang của A. Trong thực tế ứng dụng, miền từ sử dụng trên mỗi biến thường được giới hạn với một mức đặc tả tối đa là k nào đó.

Định nghĩa 2. Với mọi số nguyên $k > 1$, *k-section* $\mathfrak{B}_k^A$ của cấu trúc ngữ nghĩa $\boldsymbol{\mathcal{S}}^A = (X^A, \leq, g)$ là cấu trúc con $\boldsymbol{S}_{\kappa}^{A} = (X_{(k)}^{A}, \leq_k, g_k)$ thỏa mãn các điều kiện sau:

(i) $X_{(k)}^{A} = \{x \in X^A: |x| \leq k\}$, tập hợp các từ có mức độ đặc tả không lớn hơn k;

(ii) Các quan hệ $\leq_k$ và g_k lần lượt là các quan hệ ≤ và g bị giới hạn trên tập từ $X_{(k)}^{A}$.

Định nghĩa 3. Với mọi số nguyên $k > 1$, một *k-section* của cấu trúc đa thể hình thang $M_{Gr}^{A} = (\mathcal{I}(X^A), \leq, \subset)$ của A là cấu trúc $M_{Gr,k}^{A} = (\mathcal{I}(X_{(k)}^{A}), \leq_k, \subset_k)$, được gọi là một cấu trúc con đa thể hình thang mức k thỏa mãn các điều kiện sau:

(i) $\mathcal{I}(X_{(k)}^{A})$, trong đó $X_{(k)}^{A}$ được định nghĩa như trong Định nghĩa 2 là tập các tập mờ hình thang của các từ $X_{(k)}^{A}$ được xây dựng theo mức từ $l = 1$ đến k;

(ii) Các quan hệ $\leq_k$ và $\subset_k$ lần lượt là các quan hệ ≤ và ⊂ bị giới hạn trên $\mathcal{I}(X_{(k)}^{A})$.

Trong [6] đã chứng minh được rằng, cấu trúc M_{Gr}^{A} như Hình 2 là hình ảnh đẳng cấu của cấu trúc ngữ nghĩa $\boldsymbol{\mathcal{S}}^A = (X^A, \leq, g)$, tức là $\boldsymbol{\mathcal{S}}^A$ có thể giải nghĩa được trong M_{Gr}^{A}.

3. ĐỀ XUẤT MÔ HÌNH THIẾT KẾ FRBC VỚI NGỮ NGHĨA TÍNH TOÁN DỰA TRÊN TẬP MỜ DẠNG HÀM *S* ĐẢM BẢO TÍNH GIẢI NGHĨA

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

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

trong đó $\mathcal{X} = \{\mathcal{X}_j, j = 1, .., n\}$ 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:

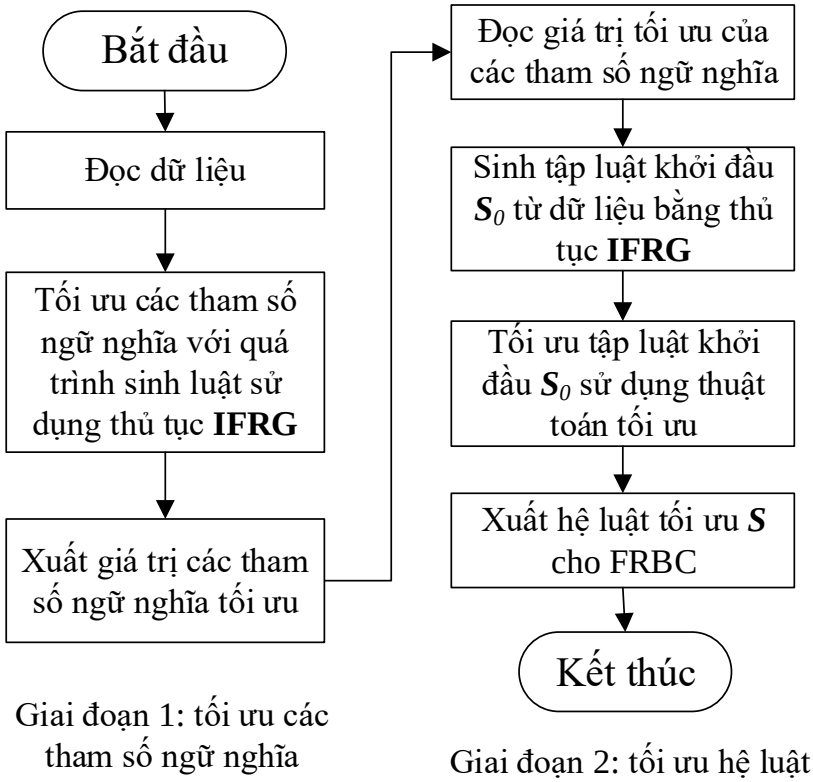
$$\boldsymbol{A}_q \Rightarrow C_q \text{ with } CF_q, \text{ với } q=1, \ldots, N \quad (2)$$

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

Giải bài toán **P** là trích xuất một tập luật ***S*** có dạng (1) từ tập dữ liệu ***P*** với điều kiện rằng nó cần phải 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 như sau (Hình 3):

(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 3. Phương pháp hai bước thiết kế hệ phân lớp dựa trên luật mờ theo tiếp cận ĐSGT [17].

Đại số gia tử mở rộng 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. Tương tự [18], trong bài báo này chúng tôi sử dụng đại số gia tử mở rộng 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 $\mathcal{AX}^{en}_j$ được liên kết với một thuộc tính thứ *j* của tập dữ liệu cảm sinh các từ ngôn ngữ $X_{j,(kj)}$ 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:

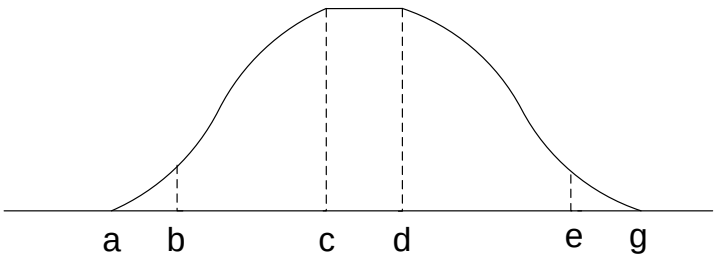
$$S_{left} = \begin{cases} 0, & 0 \le v \le a \\ \dfrac{(v-a)^2}{(b-a)(c-a)}, & a \le v \le b \\ 1 - \dfrac{(v-c)^2}{(c-b)(c-a)}, & b \le v \le c \\ 1, & v \ge c \end{cases}$$

và hàm biểu diễn độ thuộc của v vào nửa phải của hàm S như sau:

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

Tập mờ dạng hàm S được biểu diễn như Hình 4 [18].

Các ưu điểm của tập mờ dạng hàm phi tuyến (hàm S) đã được khẳng định trong các nghiên cứu trước đây [18, 19]. Vì vậy, trong bài báo này chúng tôi tiếp tục sử dụng tập mờ dạng hàm S để xây dựng cấu trúc phân hoạch đa ngữ nghĩa trên miền giá trị thuộc tính của tập dữ liệu như được thể hiện trong Hình 4.

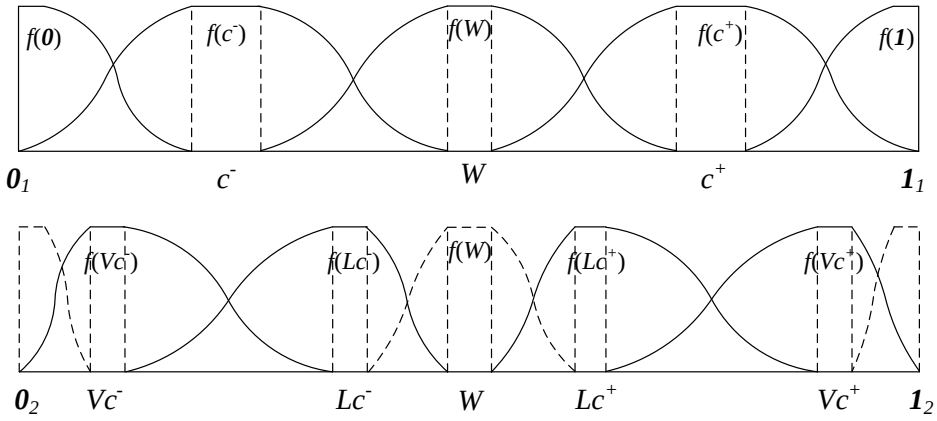


Hình 4. Biểu diễn tập mờ dạng hàm S [*18*].

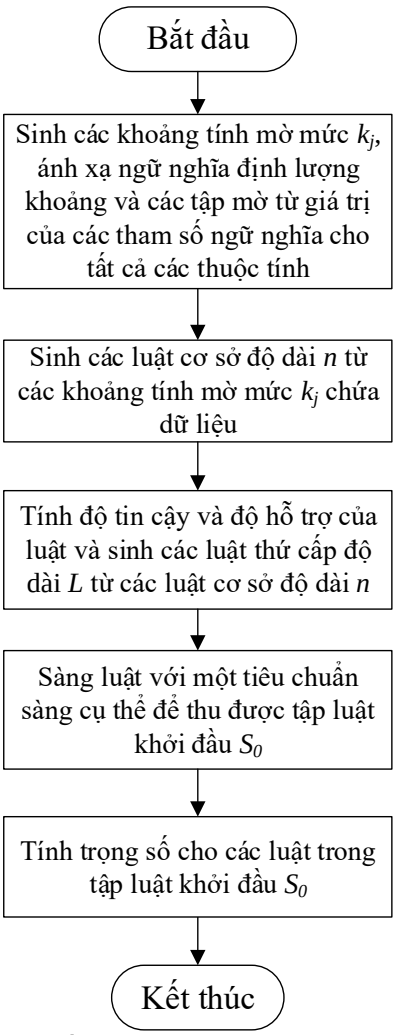
Với các từ ngôn ngữ không phải là các hằng tử ***0*** và ***1***, giá trị của a là giá trị đầu mút phải của giá trị định lượng khoảng của từ gần nhất bên trái có cùng độ dài và giá trị g là đầu mút trái của giá trị định lượng khoảng của từ gần nhất bên phải có cùng độ dài. Ví dụ, Hình 5 biểu diễn cấu trúc đa ngữ nghĩa 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 $\Im_{kj}(x_{j,i})$ 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 $\Im_{kj}(x_{j,i})$ chứa điểm dữ liệu $d_{p,j}$

của mẫu dữ liệu d_p. Tất cả các khoảng tính mờ mức k_j chứa $d_{p,j}$ ($0 \leq j \leq n$) tạo thành một siêu hộp $\mathbf{H}_p$ và chỉ sinh các luật mờ từ các siêu hộp loại này. Luật mờ cơ sở có độ dài n được sinh từ $\mathbf{H}_p$ với nhãn lớp C_p của mẫu dữ liệu d_p và các luật mờ thứ cấp có độ dài $L \leq n$ thu được bằng cách bỏ bớt n - L thuộc tính.



Hình 5. Cấu trúc phân hoạch đa ngữ nghĩa với tập mờ dạng hàm S và k_j = 2.



Hình 6. Lưu đồ thủ tục sinh tập luật khởi đầu [17].

Quá trình sinh luật thông qua thủ tục sinh tập luật khởi đầu **IFRG**($\mathcal{J}$, $\boldsymbol{P}$, NR_0, L) [17], trong đó $\mathcal{J}$ 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 6. 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 $\boldsymbol{D}$ và đã được chứng minh trong [17].

Mỗi loại dữ liệu có sự phân bố khác nhau, yêu cầu các bộ tham số ngữ nghĩa phù hợp để tối ưu hóa hiệu suất phân lớp. Vì vậy, chúng tôi áp dụng một thuật toán tối ưu để tìm ra bộ tham số ngữ nghĩa tối ưu, sau đó sử dụng chúng để tạo ra tập luật ban đầu cho quá trình lựa chọn tập luật nhỏ gọn và dễ hiểu cho hệ phân lớp, đồng thời đảm bảo sự cân bằng giữa độ chính xác và độ phức tạp của FRBC.

4. THỰC NGHIỆM

4.1. Cài đặt thực nghiệm

Để công bằng trong so sánh, đánh giá, các thực nghiệm được tiến hành giống như trong [15, 16, 19]. Cụ thể, các thực nghiệm cũng 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 [19] được sử dụng để kiểm tra giả thuyết H_0 (null hypothesis) có độ tin cậy là 95% (α = 0,05) 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 và để đảm bảo tính cân đối giữa các miền ngữ nghĩa, 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); $1 \le k_j \le 3$; $0{,}2 \le \left\{ fm\left(c_j^-\right), fm\left(c_j^+\right)\right\} \le 0{,}7$; $0{,}00001 \le \{fm(\boldsymbol{0}_j),\ fm(\boldsymbol{1}_j)\} \le 0{,}01$; $0{,}0001 \le fm(W_j) \le 0{,}2$; $fm\left(0_j\right) + fm\left(c_j^-\right) + fm\left(W_j\right) + fm\left(c_j^+\right) + fm\left(1_j\right) = 1$; $0{,}2 \le \{\mu(L_j),\ \mu(V_j)\} \le 0{,}7$; $0{,}01 \le \mu(h_{0,j}) \le 0{,}5$; and $\mu(L_j) + \mu(V_j) + \mu(h_{0,j}) = 1$.

Thuật toán tối ưu bầy đàn đa mục tiêu (PSO) [20] 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 $|\boldsymbol{S}_0| = 300 \times$ *số lớp*; độ dài tối đa của luật là 3.

4.2. Kết quả thực nghiệm và thảo luận

Để chứng tỏ tính hiệu quả của hệ phân lớp với cấu trúc đa ngữ nghĩa dựa trên tập mờ dạng hàm S được được đề xuất (được ký hiệu là **FRBC_MS**), kết quả thực nghiệm của **FRBC_MS** được so sánh với hệ phân lớp với tập mờ dạng hình thang đảm bảo tính giải nghĩa được đề xuất trong [15] (**FRBC_TRA_MS**) và hệ phân lớp với tập mờ dạng hình thang không đảm bảo tính giải nghĩa được đề xuất trong [16] (**FRBC_TRA**). Kết quả thực nghiệm và so sánh các hệ phân lớp nêu trên được thể hiện trong Bảng 1. Trong Bảng 1, cột P_{te} là độ chính xác phân lớp trung bình trên tập kiểm tra và $\#R \times C$ là độ phức tạp của hệ phân lớp được tính bằng tích của số luật trung bình và số điều kiện trung bình của các luật.

Với số liệu trong Bảng 1 ta thấy, hệ phân lớp **FRBC_MS** được đề xuất có độ chính xác phân lớp trung bình trên tập kiểm tra đối với 15 tập dữ liệu thực nghiệm là 85,07, lớn hơn tương ứng so với các hệ phân lớp **FRBC_MS** và **FRBC_TRA** là 0,36% và 0,67. Xét theo độ phức tạp của hệ phân lớp, hệ phân lớp **FRBC_MS** được đề xuất có độ phức trung bình tạp tương đương hai hệ phân lớp được so sánh.

Bảng 1. Kết quả thực nghiệm của các hệ phân lớp **FRBC_MS**, **FRBC_TRA_MS và FRBC_TRA.**

STT	Tập dữ liệu	FRBC_MS		FRBC_TRA_MS		$\neq P_{te}$	$\neq R\times C$	FRBC_TRA		$\neq P_{te}$	$\neq R\times C$
		$\#R\times C$	P_{te}	$\#R\times C$	P_{te}			$\#R\times C$	P_{te}		
1	Appendicitis	19,72	88,73	18,35	88,52	0,21	1,37	16,77	88,15	0,58	2,95
2	Australian	49,00	87,73	48,34	87,54	0,19	0,66	46,50	87,15	0,58	2,50
3	Glass	428,07	73,03	443,60	72,33	0,70	-15,53	474,29	72,24	0,79	-46,22
4	Haberman	20,00	77,05	9,60	76,77	0,28	10,40	10,80	77,40	-0,35	9,20
5	Hayes-roth	121,4	85,42	110,06	85,00	0,42	11,09	114,66	84,17	1,25	6,48
6	Heart	93,22	84,57	87,53	84,57	0,00	5,70	123,29	84,57	0,00	-30,07
7	Ionosphere	91,87	92,13	85,90	91,84	0,29	5,97	88,03	91,56	0,57	3,84
8	Iris	21,68	98,67	16,00	98,00	0,67	5,68	30,37	97,33	1,34	-8,69
9	Mammogr.	72,90	83,91	78,55	84,33	-0,42	-5,65	73,84	84,20	-0,29	-0,94
10	Newthyroid	41,61	96,94	52,20	96,46	0,48	-10,59	39,82	95,67	1,27	1,79
11	Pima	46,73	77,18	55,49	76,95	0,23	-8,76	56,12	77,01	0,17	-9,40
12	Saheart	78,71	70,92	64,88	70,49	0,43	13,83	59,28	70,05	0,87	19,43
13	Tae	170,45	63,57	142,71	62,07	1,50	27,73	210,70	61,00	2,57	-40,26
14	Wine	39,10	98,87	34,98	98,70	0,17	4,11	40,39	98,49	0,38	-1,29
15	Wisconsin	78,76	97,28	75,29	97,09	0,19	3,48	69,81	96,95	0,33	8,95
	Trung bình	**91,53**	**85,07**	**88,23**	**84,71**			**96,98**	**84,40**		

Bảng 2. So sánh độ chính xác của hệ phân lớp **FRBC_MS** so với **FRBC_TRA_MS** và **FRBC_TRA** bằng phương pháp kiểm định Wilcoxon với α = 0,05.

So sánh (α = 0,05)	R^+	R^-	Exact *P*-value	Giả thuyết H_0
FRBC_MS vs **FRBC_TRA_MS**	96,5	8,5	0,00354	Bị bác bỏ
FRBC_MS vs **FRBC_TRA**	99,0	6,0	0,001709	Bị bác bỏ

Để khẳng định tỷ lệ tốt hơn về độ chính xác phân lớp của hệ phân lớp được đề xuất so với hai hệ phân lớp được so sánh là có ý nghĩa, phương pháp kiểm định thống kê Wilcoxon được sử dụng để kiểm tra giả thuyết H_0 (null hypothesis) với độ tin cậy là 95% (α = 0,05) và giả định rằng độ chính xác phân lớp của các phương pháp phân lớp được so sánh là tương đương nhau. Kết quả kiểm tra được thể hiện trong Bảng 2 cho thấy rằng, các giả thuyết H_0 bị bác bỏ.

Điều này cho thấy rằng, hệ phân lớp được đề xuất **FRBC_MS** thực sự tốt hơn hai hệ phân lớp được so sánh về độ chính xác phân lớp.

Tiếp tục sử dụng phương pháp kiểm định thống kê Wilcoxon đối với độ phức tạp của hệ luật phân lớp với giả thuyết độ phức tạp của các hệ phân lớp là tương đương nhau. Kết quả kiểm tra được thể hiện trong Bảng 3 cho thấy rằng, các giả thuyết H_0 không bị bác bỏ. Do đó, hệ phân lớp được đề xuất **FRBC_MS** có độ phức tạp tương đương hai hệ phân lớp còn lại.

Với các kết quả so sánh và kiểm định thống kê Wilcoxon nếu trên, ta có thể khẳng định rằng hệ phân lớp được đề xuất **FRBC_MS** có độ chính xác phân lớp tốt hơn so với các hệ phân lớp được so sánh nhưng không làm tang độ phức tạp của hệ luật.

Bảng 3. So sánh độ phức tạp của hệ phân lớp **FRBC_MS** so với **FRBC_TRA_MS** và **FRBC_TRA** bằng phương pháp kiểm định Wilcoxon với $\alpha = 0,05$.

So sánh ($\alpha = 0,05$)	R^+	R^-	Exact *P*-value	Giả thuyết H_0
FRBC_MS vs **FRBC_TRA_MS**	39,0	81,0	$\geq 0,2$	Không bị bác bỏ
FRBC_MS vs **FRBC_TRA**	64,0	56,0	$\geq 0,2$	Không bị bác bỏ

5. KẾT LUẬN

Ngữ nghĩa tính toán dựa trên tập mờ của các từ ngôn ngữ trong cơ sở luật của các hệ dựa trên luật mờ nói chung và của hệ phân lớp dựa trên luật mờ nói riêng đóng vai trò quan trọng, là một trong các yếu tố nâng cao độ chính xác của hệ phân lớp. Bài báo này trình bày phương pháp biểu diễn ngữ nghĩa tính toán dựa trên tập mờ dạng hàm *S* với cấu trúc đa ngữ nghĩa đảm bảo tính giải nghĩa của hệ phân lớp và được xây dựng bởi đại số gia tử mở rộng. 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.

TÀI LIỆU THAM KHẢO

[1]. R. Alcala, 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, 15 (2011) 2303–2318. https://doi.org/10.1007/s00500-010-0671-2

[2]. N. N. Quynh, N. P. Dong, N. L. Giang, H. V. Long, A new Takagi-Sugeno fuzzy system approach for fuzzy state feedback controller design and its application to malware propagation on heterogeneous complex network, Journal of Science and Technology on Information Security, 3 (2023) 43-53. https://doi.org/10.54654/isj.v3i20.988

[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, 141 (2004) 59-88. https://doi.org/10.1016/S0165-0114(03)00114-3

[4]. H. Ishibuchi, T. Yamamoto, Rule weight specification in fuzzy rule-based classification systems, IEEE Transactions on Fuzzy Systems, 13 (2005) 428–435. https://doi.org/10.1109/TFUZZ.2004.841738

[5]. C. H. Nguyen, V. T. Hoang, V. L. Nguyen, A discussion on interpretability of linguistic rule based systems and its application to solve regression problems, Knowledge-Based Systems, 88 (2015) 107–133. https://doi.org/10.1016/j.knosys.2015.08.002
[6]. V. T. Hoang, C. H. Nguyen, D. D. Nguyen, D. P. Pham, V. L. Nguyen, The interpretability and scalability of linguistic-rule-based systems for solving regression problems, International Journal of Approximate Reasoning, 149 (2022) 131-160. https://doi.org/10.1016/j.ijar.2022.07.007
[7]. D. D. Nguyen, D. P. Pham, V. T. Hoang, C. H. Nguyen, Một phương pháp xây dựng hệ dựa trên luật mờ có khả năng mở rộng giải bài toán hồi quy, Tạp chí Khoa học và Công nghệ - Đại học Thái Nguyên, 226 (2021) 341-348. https://doi.org/10.34238/tnu-jst.4811
[8]. 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, 127 (2017) 67–84. https://doi.org/10.1016/j.knosys.2016.12.028
[9]. M. Pota, M. Esposito, G. D. Pietro, Designing rule-based fuzzy systems for classification in medicine, Knowledge-Based Systems, 124 (2017) 105–132. https://doi.org/10.1016/j.knosys.2017.03.006
[10].C. Mencar, A.M. Fanelli, Interpretability constraints for fuzzy information granulation, Information Sciences, 178 (2008) 4585–4618. https://doi.org/10.1016/j.ins.2008.08.015
[11].C. H. Nguyen, W. Wechler, Hedge algebras: an algebraic approach to structures of sets of linguistic domains of linguistic truth variables, Fuzzy Sets and Systems, 35 (1990) 281-293. https://doi.org/10.1016/0165-0114(90)90002-N
[12].C. H. Nguyen, W. Wechler, Extended hedge algebras and their application to fuzzy logic, Fuzzy Sets and Systems, 52 (1992) 259–281. https://doi.org/10.1016/0165-0114(92)90237-X
[13].C. H. Nguyen, V. L. Nguyen, Fuzziness measure on complete hedges algebras and quantifying semantics of terms in linear hedge algebras, Fuzzy Sets and Systems, 158 (2007) 452-471. https://doi.org/10.1016/j.fss.2006.10.023
[14].A. Tarski, A. Mostowski, and R. Robinson, Undecidable Theories. North-Holland, 1953.
[15].D. P. Pham, V. T. Hoang, D. D. Nguyen, Biểu diễn ngữ nghĩa tính toán đảm bảo tính giải nghĩa của hệ phân lớp dựa trên luật mờ, Tạp chí Khoa học và Công nghệ - Đại học Thái Nguyên, 227 (2022) 107 - 114. https://doi.org/10.1016/j.knosys.2014.04.047
[16].C. H. Nguyen, T. S. Tran, D. P. Pham, Modeling of a semantics core of linguistic terms based on an extension of hedge algebra semantics and its application, Knowledge-Based Systems, 67 (2014) 244–262. https://doi.org/10.1016/j.knosys.2014.04.047
[17].C. H. Nguyen, W. Pedrycz, T. L. Duong, T. S. Tran, A genetic design of linguistic terms for fuzzy rule based classifiers, International Journal of Approximate Reasoning, 54 (2013) 1-21. https://doi.org/10.1016/j.ijar.2012.07.007
[18].V. T. Hoang, D. D. Nguyen, C. H. Nguyen, Một phương pháp thiết kế ngữ nghĩa dạng tập mờ của từ ngôn ngữ dựa trên đại số gia tử mở rộng và ứng dụng xây dựng FRBS giải bài toán hồi qui, Chuyên san Các công trình nghiên cứu, phát triển và ứng dụng Công nghệ Thông tin và Truyền thông, 38 (2017) 51-57. https://doi.org/10.32913/rd-ict.vol2.no38.527
[19].D. D. Nguyen, D. P. Pham, D. V. Pham, D. T. Nguyen, Một phương pháp thiết kế ngữ nghĩa tính toán của các từ ngôn ngữ giải bài toán phân lớp dựa trên luật mờ, Chuyên san Các công trình nghiên cứu, phát triển và ứng dụng Công nghệ Thông tin và Truyền thông, 1 (2020) 9-18. https://doi.org/10.32913/mic-ict-research-vn.v2020.n1.914
[20].J. Demˇsar, Statistical Comparisons of Classifiers over Multiple Data Sets, Journal of Machine Learning Research, 7 (2006) 1–30.
[21]. D. P. Pham, C. H. Nguyen, T. T. Nguyen, 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 2013, 12–17. https://doi.org/10.1109/RIVF.2013.6719858