COMPUTATIONAL SEMANTIC REPRESENTATION GUARANTEES INTERPRETABILITY OF FUZZY RULE BASED CLASSIFIER

Pham Dinh Phong, Hoang Van Thong*, Nguyen Duc Du
University of Transport and Communications

ARTICLE INFO

Received: 26/9/2022
Revised: 19/10/2022
Published: 20/10/2022

KEYWORDS

Hedge algebras
Order-based semantics
Classifier
Interpretability
Fuzzy rule-based systems

ABSTRACT

The fuzzy rule-based classifier design methods have been widely studied by the research community due to many practical applications in the real life. The quality of a classifier clearly depends on the semantic representations of linguistic words in the rule bases. Hedge algebra allows to the creation of a formal formalism for designing the fuzzy sets-based computational semantics of linguistic words from their inherent semantics. However, the existing design methods of fuzzy sets-based computational semantics of linguistic words do not guarantee the interpretability of the fuzzy rule-based classifiers. Specifically, the designed multi-granularity representation does not retain the generality-specificity relation of linguistic terms. This paper presents a fuzzy sets-based computational semantic representation that guarantees the interpretability of the fuzzy rule-based classifier. Experimental results on 23 real-world datasets have shown that the proposed method gives better classification accuracy while not increasing the complexity of the fuzzy rule-based systems in comparison with the existing methods.

BIỂU DIỄN NGỮ NGHĨA TÍNH TOÁN ĐẢM BẢO TÍNH GIẢI NGHĨA CỦA HỆ PHÂN LỚP DỰA TRÊN LUẬT MỜ

Phạm Đình Phong, Hoàng Văn Thông*, Nguyễn Đức Dư
Trường Đại học Giao thông vận tải

THÔNG TIN BÀI BÁO

Ngày nhận bài: 26/9/2022
Ngày hoàn thiện: 19/10/2022
Ngày đăng: 20/10/2022

TỪ KHÓA

Đại số gia tử
Thứ tự ngữ nghĩa
Hệ phân lớp
Tính giải nghĩa được
Hệ dựa trên luật mờ

TÓM TẮT

Phương pháp thiết kế hệ phân lớp dựa trên luật mờ đã và đang được nghiên cứu rộng rãi do có nhiều ứng dụng trong thực tiễn. Chất lượng của một hệ phân lớp phụ thuộc vào các 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. Tuy nhiên, các phương pháp thiết kế ngữ nghĩa tính toán dựa trên tập mờ chưa đảm bảo tính giải nghĩa của hệ phân lớp dựa trên luật mờ. Cụ thể, biểu diễn đa thể hạt của khung nhận thức ngôn ngữ chưa đảm bảo tính chung - riêng của các từ ngôn ngữ. Bài báo này trình bày một phương pháp biểu diễn ngữ nghĩa tính toán dựa trên tập mờ đảm bảo tính giải nghĩa được của hệ phân lớp. Kết quả thực nghiệm với 23 tập dữ liệu chuẩn 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.

DOI: https://doi.org/10.34238/tnu-jst.6566

* Corresponding author. *Email: thonghv@utc.edu.vn*

1. Giới thiệu

Phương pháp thiết kế hệ phân lớp dựa trên luật mờ (Fuzzy rule based classifiers – FRBCs) được nghiên cứu rộng rãi do có nhiều ứng dụng thực tế trong lĩnh vực khai phá dữ liệu [1] – [5]. Phương pháp thiết kế FRBC theo tiếp cận lý thuyết tập mờ [1] – [5] sử dụng các tập mờ để phân hoạch miền giá trị của các thuộc tính dựa trên tri thức của các chuyên gia. Do đó, số tập mờ được sử dụng bị giới hạn là 7±2 và các tập mờ thường được biểu diễn dưới dạng phân hoạch đều. Bên cạnh đó, do không có cầu nối hình thức giữa ngữ nghĩa của các từ ngôn ngữ với các tập mờ nên các tập mờ thu được sau quá trình tối ưu không phản ánh đúng ngữ nghĩa thực của các từ ngôn ngữ theo ý đồ thiết kế của các chuyên gia và làm giảm tính giải nghĩa của hệ luật phân lớp.

Đại số gia tử (ĐSGT) [6] – [8] được Nguyễn Cát Hồ và cộng sự giới thiệu vào năm 1990 đã được ứng dụng hiệu quả trong nhiều lĩnh vực khác nhau như khai phá dữ liệu [9] – [14], điều khiển mờ [15], xử lý ảnh [16],... ĐSGT cung cấp một cơ sở 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 khai thác tính thứ tự về ngữ nghĩa của các từ trong miền giá trị ngôn ngữ, cho phép tạo ra một cơ sở hình thức sinh ngữ nghĩa tính toán dựa trên tập mờ từ ngữ nghĩa định tính vốn có của các từ ngôn ngữ. Dựa trên cơ sở hình thức này, Nguyễn Cát Hồ và các cộng sự lần đầu tiên đã ứng dụng ĐSGT để thiết kế tối ưu các từ ngôn ngữ cùng với ngữ nghĩa tính toán dựa trên tập mờ cho FRBC một cách hiệu quả [9], [10], trong đó ngữ nghĩa dựa trên tập mờ hình thang được chứng minh là hiệu quả hơn ngữ nghĩa tính toán dựa trên tập mờ tam giác.

Dựa trên ĐSGT, một phương pháp luận tính toán trực tiếp trên các từ ngôn ngữ để thiết kế các hệ dựa trên luật mờ có tính giải nghĩa theo quan điểm của Tarski đã được đề xuất và được áp dụng hiệu quả đối với bài toán hồi quy [13], [17] và tóm tắt ngôn ngữ từ dữ liệu [14]. Phương pháp luận này đảm bảo 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. Trong bài báo này, chúng tôi áp dụng phương pháp luận nói trên giải bài toán phân lớp dựa trên luật mờ (FRBC).

2. Phương pháp nghiên cứu

2.1. Cấu trúc đa ngữ nghĩa của miền hạng từ

2.1.1. Tính giải nghĩa được

Tính giải nghĩa được Tarski và các cộng sự [18] định nghĩa trong toán học và logic như sau:

Lý thuyết S được gọi là có thể giải nghĩa được trong lý thuyết T nếu tồn tại một bản dịch T từ ngôn ngữ hình thức L(S) của S sang ngôn ngữ hình thức L(T) của T thỏa mãn điều kiện, với mọi mệnh đề p $\in L(S)$ *thì p có thể chứng minh được trong S khi và chỉ khi T(p)* $\in L(T)$ *có thể chứng minh được trong T.*

Theo khái niệm này, 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 T khi và chỉ khi S có thể giải nghĩa được trong T bằng phép biến đổi T. Như vậy, nếu lý thuyết T thỏa mãn điều kiện này thì T được gọi là có thể giải nghĩa được đối với S.

Trong các mục tiếp theo sẽ khẳng định rằng, cấu trúc đa ngữ nghĩa $\mathbf{S}^A = (X^A, \leq, g)$ của miền từ X^A của thuộc tính A với các quan hệ thứ tự ngữ nghĩa $\leq$ và quan hệ khái quát-đặc tả g thì khi tính toán với từ ngôn ngữ thông qua 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 đa ngữ nghĩa $\mathbf{S}^A = (X^A, \leq, g)$. Khi đó cấu trúc tập mờ này có thể giải nghĩa được cho S^A.

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

ĐSGT mở rộng được Nguyễn Cát Hồ và các cộng sự giới thiệu trong [10] là một mở rộng của ĐSGT truyền thống bằng việc bổ sung một gia tử nhân tạo h_0 nhằm mô hình hóa lõi ngữ nghĩa

của các từ ngôn ngữ. Nhờ đó, ĐSGT mở rộng đã đáp ứng được các yêu cầu đa dạng trong biểu diễn cấu trúc đa ngữ nghĩa của các ứng dụng trong thực tiễn.

Cho một ĐSGT tuyến tính $\mathcal{A}^A = (X^A, G, C, H, \leq)$ 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 $X_{en}^A = X^A \cup \{h_0x: x \in X^A\}$. Ta có ĐSGT mở rộng của $\mathcal{A}^A$ là $\mathcal{A}_{en}^A = (X_{en}^A, G, C, H_{en}, \leq)$, trong đó, $H_{en} = H \cup \{h_0\}$, $X_{en}^A = C \cup H_{en}(G) = C \cup \{h_n \dots h_1 c: c \in G, h_j \in H_{en}$, với $j = 1, \dots, n\}$. Do đó, ta có, $X^A = C \cup H(G) \subseteq X_{en}^A = C \cup H_{en}(G)$.

Đặt $X_{en,k}^A = \{x \in X_{en}^A: |x| = k\}$, trong đó $|x|$ là độ dài của x, là tập các từ có độ đặc tả k (k-specificity) và $X_{en,(k)}^A = \{x \in X_{en}^A: |x| \leq k\} = \bigcup_{1 \leq i \leq k} X_{en,i}^A$ là tập các từ có độ đặc tả không lớn hơn k. Khi đó, $X_{en,1}^A = G \cup C$ và với mọi $k \geq 2$ thì $X_{en,k}^A = X_k^A \cup \{h_0u: u \in X_{en,(k-1)}^A\}$, tức là với $k > 0$ thì $X_{en,(k)}^A$ 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.

Ngoài cấu trúc ngữ nghĩa dựa trên thứ tự, được ký hiệu là $\mathbb{S}^A = (X_{en}^A, \leq)$, miền từ của X^A bao hàm một cấu trúc ngữ nghĩa khác được thể hiện thông qua quan hệ *khái quát-đặc tả* (*generality-specificity*), 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. Cấu trúc này được gọi là cấu trúc *khái quát-đặc tả* và được ký hiệu là $\boldsymbol{G}^A = (X_{en}^A, g)$.

Như vậy, miền từ của X^A bao gồm hai cấu trúc: $\mathbb{S}^A = (X_{en}^A, \leq)$ và $\boldsymbol{G}^A = (X_{en}^A, g)$, tức là biến ngôn ngữ A không chỉ có một cấu trúc ngữ nghĩa theo thứ tự như quan niệm trước đây mà còn nhiều vấn đề phức tạp ở trong nó. Kết hợp cấu trúc thứ tự $\mathbb{S}^A$ và cấu trúc *khái quát-đặc tả* $\boldsymbol{G}^A$ ta có cấu trúc ngữ nghĩa đa mức hay cấu trúc đa ngữ nghĩa và được biểu thị bằng $\mathrm{S}^A = (X_{en}^A, \leq, g)$.

2.1.3. Biểu diễn cấu trúc đa ngữ nghĩa của miền từ dựa trên ĐSGT

Muốn cấu trúc $\mathrm{T}(X^A)$ biểu diễn cấu trúc $\boldsymbol{\mathcal{S}}^A = (X^A, \leq, g)$ bảo toàn cấu trúc của $\boldsymbol{\mathcal{S}}^A$ hay nói cách khác là $\mathrm{T}(X^A)$ giải nghĩa được thì cần định nghĩa hai quan hệ ký hiệu là $\leq$ và $\subset$ trên $\mathrm{T}(X^A)$ vì $\boldsymbol{\mathcal{S}}^A$ 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 $\mathrm{T}(X^A)$, định nghĩa:

1) Quan hệ thứ tự $\leq$ trên $\mathrm{T}(X^A)$: 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 $\subset$ trên $\mathrm{T}(X^A)$: 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 $\mathrm{T}(X^A)$ với hai quan hệ $\leq$ và $\subset$ được ký hiệu là $M_{Gr}^A = (\mathrm{T}(X^A), \leq, \subset)$, 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;

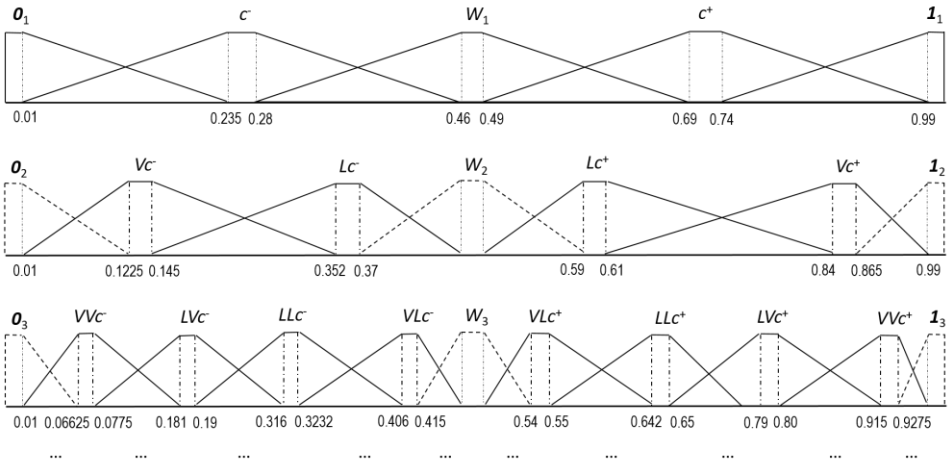
(ii) Các quan hệ $\leq_k$ và g_k lần lượt là các quan hệ $\leq$ và g bị giới hạn trên tập từ $X_{(k)}^A$.

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

(i) $\mathrm{T}(X_{(k)}^A)$, 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 $\leq$ và $\subset$ bị giới hạn trên $\mathrm{T}(X_{(k)}^A)$.

Trong [14], cấu trúc M_{Gr}^{A} như Hình 1 đã được chứng minh là hình ảnh đẳng cấu của cấu trúc ngữ nghĩa $\mathcal{S}^A = (X^A, \leq, g)$, tức là $\mathcal{S}^A$ có thể giải nghĩa được trong M_{Gr}^{A}.



Hình 1. *Cấu trúc phân hoạch đa thể hình thang biểu diễn cấu trúc ngữ nghĩa $\mathcal{S}^A = (X^A, \leq g)$ của biến A*

2.2. Thiết kế hệ phân lớp dựa trên luật mờ trên cơ sở cấu trúc đa ngữ 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:

Luật R_q: **If** X_1 is $A_{q,1}$ **and** ... **and** X_n is $A_{q,n}$ **then** C_q with CF_q, for $q=1, ..., N$ (1)

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

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

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

Giải bài toán **P** là trích xuất từ tập dữ liệu $\boldsymbol{P}$ một tập luật $\boldsymbol{S}$ có dạng (1) nhỏ gọn, dễ hiểu với người dùng và có độ chính xác phân lớp cao.

Cấu trúc đa thể hình thang M_{Gr}^{A} biểu diễn cấu trúc đa ngữ nghĩa của miền hạng từ được sinh ra như trong Hình 1. Như đã được đề cập ở trên, đây là cấu trúc đảm bảo tính giải nghĩa, do đó hệ phân lớp dựa trên hệ luật mờ với ngữ nghĩa tính toán của các từ ngôn ngữ được biểu diễn bởi cấu trúc này sẽ đảm bảo tính giải nghĩa của hệ phân lớp đó. Thủ tục sinh luật trong [9] được sử dụng sinh tập luật mờ từ dữ liệu. 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.

3. Kết quả thực nghiệm và thảo luận

Mục này trình bày các kết quả thực nghiệm của phương pháp biểu diễn ngữ nghĩa tính toán của các từ ngôn ngữ dựa trên tập mờ hình thang theo cấu trúc đa ngữ nghĩa mới được sinh bởi ĐSGT $\boldsymbol{AX}^{en}$, đảm bảo tính giải nghĩa của hệ phân lớp và chứng minh tính hiệu quả hệ phân lớp mới này so với cấu trúc đa thể hạt cũ và tiếp cận lý thuyết tập mờ.

3.1. Cài đặt thực nghiệm

Các thực nghiệm được cài đặt bằng ngôn ngữ C# chạy trên Windows 10 với cấu hình máy Intel Core i5-8250U 1,8GHz, 8GB RAM. Các tập dữ liệu dùng trong các 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 (*ten-folds cross-validation*) được áp dụng để huấn luyện và kiểm tra. Phương pháp kiểm định giả thuyết thống kê Wilcoxon [19] được sử dụng để kết luận về ý nghĩa so sánh giữa các phương pháp.

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*); $0 \le k_j \le 3$; $0{,}2 \le \{\text{fm}(c_j^-), fm(c_j^+)\} \le 0{,}7$; $0{,}00001 \le \{fm(0_j),\ fm(1_j)\} \le 0{,}1$; $0{,}0001 \le fm(W_j) \le 0{,}2$; $fm(0_j) + fm(c_j^-) + fm(W_j) + fm(c_j^+) + fm(1_j)$; $0{,}2 \le \{\mu(L_j), \mu(V_j)\} \le 0{,}7$; $0{,}01 \le \mu(h_{0,j}) \le 0{,}5$; và $\mu(L_j) + \mu(V_j) + \mu(h_{0,j}) = 1$.

Để tối ưu các tham số ngữ nghĩa và lựa chọn hệ luật tối ưu cho hệ phân lớp, thuật toán tối ưu bầy đàn đa mục tiêu (PSO) [20] được sử dụng. Trong tối ưu các tham số ngữ nghĩa, giá trị của các tham số của thuật toán: 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, giá trị của các tham số của thuật toán: số thế hệ là 1000; số luật khởi tạo là $|\boldsymbol{S}_0| = 300 \times$ *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 tích của độ tin cậy và độ hỗ trợ tương ứng theo công thức (4) và (5) trong [4] và trọng số luật được tính toán theo công thức (10) trong [4].

3.2. Kết quả thực nghiệm

Bảng 1. *Kết quả thực nghiệm của hệ phân lớp* ***FRBC_GS*** *và* ***FRBC_AX^{en}***

STT	Tập dữ liệu	FRBC_GS				FRBC_AX^{en}				$\neq P_{te}$	$\neq R \times C$
		$\#R$	$\#R \times C$	P_{tr}	P_{te}	$\#R$	$\#R \times C$	P_{tr}	P_{te}		
1	Appendicitis	3,93	18,35	92,52	**88,52**	3,67	**16,77**	92,38	88,15	0,37	1,58
2	Australian	4,80	48,34	88,53	**87,54**	5,00	**46,50**	88,56	87,15	0,39	1,84
3	Bands	6,00	**56,22**	76,44	**74,32**	6,00	58,20	78,19	73,46	0,86	-1,98
4	Bupa	9,77	186,31	77,28	**72,44**	8,97	**181,19**	79,78	72,38	0,06	5,12
5	Cleveland	13,93	**410,52**	68,97	62,17	14,57	468,13	66,64	**62,39**	-0,22	-57,62
6	Dermatology	12,20	259,86	96,88	**95,62**	10,43	**182,84**	96,37	94,40	1,22	77,02
7	Glass	14,53	**443,60**	78,91	**72,33**	14,23	474,29	78,78	72,24	0,09	-30,68
8	Haberman	3,00	**9,60**	77,20	76,77	3,00	10,80	77,60	**77,40**	-0,63	-1,20
9	Hayes-roth	9,57	**110,06**	88,31	**85,00**	9,80	114,66	89,40	84,17	0,83	-4,60
10	Heart	7,50	**87,53**	88,46	84,57	8,37	123,29	89,19	84,57	0,00	-35,77
11	Hepatitis	4,00	**19,48**	93,51	**89,93**	3,70	25,53	93,68	89,28	0,65	-6,05
12	Ionosphere	8,53	**85,90**	95,17	**91,84**	8,63	88,03	94,69	91,56	0,28	-2,13
13	Iris	4,00	**16,00**	98,00	**98,00**	5,30	30,37	98,25	97,33	0,67	-14,37
14	Mammogr.	6,97	78,55	85,64	**84,33**	7,10	**73,84**	85,49	84,2	0,13	4,71
15	Newthyroid	6,00	52,20	97,57	**96,46**	5,33	**39,82**	96,76	95,67	0,79	12,38
16	Pima	6,40	**55,49**	78,23	76,95	5,97	56,12	78,69	**77,01**	-0,06	-0,63
17	Saheart	6,60	64,88	75,86	**70,49**	5,63	**59,28**	75,51	70,05	0,44	5,59
18	Sonar	6,00	**48,42**	88,16	**79,75**	5,87	49,31	87,59	78,61	1,14	-0,89
19	Tae	9,47	**142,71**	69,91	**62,07**	10,90	210,70	68,97	61,00	1,07	-67,98
20	Vehicle	11,17	195,81	70,13	**68,52**	11,23	**195,07**	70,74	68,20	0,32	0,74
21	Wdbc	4,73	40,82	97,42	96,78	4,00	**25,04**	97,08	96,78	0,00	15,78
22	Wine	5,67	**34,98**	99,73	**98,70**	5,77	40,39	99,60	98,49	0,21	-5,41
23	Wisconsin	7,90	75,29	97,92	**97,09**	7,87	**69,81**	97,78	96,95	0,14	5,48
Trung bình			**110,47**	**86,12**	**83,05**		114,78	86,16	82,67		

Ký hiệu hệ phân lớp được đề xuất là **FRBC_GS** và ký hiệu hệ phân lớp sử dụng cấu trúc đa thể hạt cũ không đảm bảo tính giải nghĩa được của hệ phân lớp [10] là **FRBC_*AX*^{en}**. Bảng 1 thể hiện các kết quả thực nghiệm và so sánh giữa hai hệ phân lớp **FRBC_GS** và **FRBC_*AX*^{en}**, trong đó, ký hiệu $\#R$ là số luật trung bình, $\#R \times C$ là độ phức tạp của hệ phân lớp được tính bằng tích của số luật trung bình $\#R$ và số điều kiện luật trung bình C, P_{tr} và P_{te} lần lượt là độ chính xác phân lớp trung bình trên tập huấn luyện và tập kiểm tra, $\neq P_{te}$ và $\neq R \times C$ tương ứng là chênh lệch của độ chính xác trên tập kiểm tra và độ phức tạp của hai hệ phân lớp được so sánh.

Trực giác quan sát các kết quả thực nghiệm trong Bảng 1 cho thấy, hệ phân lớp **FRBC_GS** có độ chính xác phân lớp trên tập kiểm tra cao hơn so với hệ phân lớp **FRBC_*AX*^{en}** đối với 20 trong số 23 tập dữ liệu được thực nghiệm. Xét trên độ chính xác phân lớp trung bình của của 23 tập dữ liệu, hệ phân lớp **FRBC_GS** có độ chính xác phân lớp trung bình là 83,05%, tốt hơn so với hệ phân lớp **FRBC_*AX*^{en}** có độ chính xác phân lớp trung bình là 82,67%, trong khi có độ phức tạp trung bình thấp hơn một chút (110,47 so với 114,78).

Bảng 2. *So sánh độ chính xác giữa hai hệ phân lớp **FRBC_GS** và **FRBC_AX**^{en} bằng Wilcoxon Signed Rank test với $\alpha = 0,05$*

So sánh (α = 0,05)	R^+	R^-	Exact *P*-value	Hypothesis
FRBC_GS vs **FRBC_*AX*^{en}**	247,0	29,0	4,08E-4	Rejected

Bảng 3. *So sánh độ phức tạp của hai hệ phân lớp **FRBC_GS** và **FRBC_AX**^{en} bằng Wilcoxon Signed Rank test với $\alpha = 0,05$*

So sánh (α = 0,05)	R^+	R^-	Exact *P*-value	Hypothesis
FRBC_GS vs **FRBC_*AX*^{en}**	158,0	118,0	$\geq 0,2$	Not Rejected

Thực hiện các kiểm định giả thuyết thống kê Wilcoxon [19] với độ tin cậy 95% (α = 0,05) sử dụng dữ liệu trong Bảng 1 với giả thiết độ chính xác phân lớp và độ phức tạp tương ứng của hai hệ phân lớp là tương đương nhau. Trong Bảng 2, ta thấy giá trị *Exact p-value* < 0,05 nên giả thuyết tương đương về độ chính xác phân lớp của hai hệ phân lớp **FRBC_GS** và **FRBC_*AX*^{en}** bị bác bỏ. Trong Bảng 3, giá trị *Exact p-value* > 0,05 nên giả thuyết tương đương về độ phức tạp của hai hệ phân lớp không bị bác bỏ. Với các kết quả kiểm định này, ta có thể khẳng định rằng phương pháp thiết kế ngữ nghĩa tính toán dựa trên cấu trúc đa ngữ nghĩa mới không những có biểu diễn các phân hoạch mờ đảm bảo tính giải nghĩa của FRBC mà còn có độ chính xác phân lớp cao hơn so với phương pháp biểu diễn đa thể hạt cũ. Hơn nữa, các kết quả thực nghiệm so sánh trên cũng cho thấy việc đảm bảo tính giải nghĩa của FRBC đóng vai trò quan trọng đảm bảo ngữ nghĩa tính toán phản ánh đúng tính mờ của thông tin và làm tăng hiệu suất của hệ phân lớp.

Để chỉ ra tính hiệu quả của hệ phân lớp được đề xuất, các kết quả thực nghiệm của hệ phân lớp **FRBC_CS** được so sánh với các kết quả của hệ phân lớp theo tiếp cận lý thuyết tập mờ được đề xuất trong [1] và [2] tương ứng là **Product-1-ALL TUN** và **PAES-RCS**. Kết quả thực nghiệm trong Bảng 4 cho thấy, hệ phân lớp **FRBC_CS** cho độ chính xác phân lớp cao hơn hai hệ phân lớp **Product-1-ALL TUN** và **PAES-RCS** đối với 22 trên 23 tập dữ liệu được thử nghiệm. Xét trên giá trị trung bình của độ chính xác phân lớp, hệ phân lớp **FRBC_CS** có giá trị trung bình là 83,05%, cao hơn 2,48% và 2,39% tương ứng so với hệ phân lớp **Product-1-ALL TUN** và **PAES-RCS**. Xét trên độ phức tạp của hệ phân lớp, hệ phân lớp **FRBC_CS** có độ phức tạp phân lớp thấp hơn nhiều so với hai hệ phân lớp còn lại, tương ứng là 110,47 so với 163,40 và 355,23.

Kết quả kiểm định giả thuyết thống kê Wilcoxon với độ tin cậy 95% (α = 0,05) sử dụng dữ liệu trong Bảng 4 đối với độ chính xác phân lớp và độ phức tạp của hệ luật tương ứng được thể hiện trong Bảng 5 và Bảng 6. Do các giá trị *Exact p-value* đều nhỏ 0,05 nên giả thuyết tương đương về độ chính xác phân lớp và độ phức tạp của hệ phân lớp của **FRBC_CS** so với **Product-1-ALL TUN** và **PAES-RCS** bị bác bỏ. Do đó, ta có thể khẳng định rằng hệ phân lớp **FRBC_CS** 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.

Bảng 4. *Kết quả thực nghiệm của các hệ phân lớp **FRBC_CS**, **Product-1-ALL TUN** và **PAES-RCS***

STT	Tập dữ liệu	FRBC_CS		PAES-RCS		$\neq P_{te}$	$\neq R \times C$	*Product-1-ALL TUN*		$\neq P_{te}$	$\neq R \times C$
		$\#R \times C$	P_{te}	$\#R \times C$	P_{te}			$\#R \times C$	P_{te}		
1	Appendicitis	18,35	88,52	35,28	85,09	3,43	-16,93	20,89	87,30	1,22	-2,54
2	Australian	48,34	87,54	329,64	85,80	1,74	-281,30	62,43	85,65	1,89	-14,10
3	Bands	56,22	74,32	756,00	67,56	6,76	-699,78	104,09	65,80	8,52	-47,87
4	Bupa	186,31	72,44	256,20	68,67	3,77	-69,89	210,91	67,19	5,25	-24,60
5	Cleveland	410,52	62,17	1140,00	59,06	3,11	-729,48	1020,66	58,80	3,37	-610,14
6	Dermatology	259,86	95,62	389,40	95,43	0,19	-129,54	185,28	94,48	1,14	74,58
7	Glass	443,60	72,33	487,90	72,13	0,20	-44,30	534,88	71,28	1,05	-91,28
8	Haberman	9,60	76,77	202,41	72,65	4,12	-192,81	21,13	71,88	4,89	-11,53
9	Hayes-roth	110,06	85,00	120,00	84,03	0,97	-9,94	158,52	78,88	6,12	-48,47
10	Heart	87,53	84,57	300,30	83,21	1,36	-212,78	164,61	82,84	1,73	-77,09
11	Hepatitis	19,48	89,93	300,30	83,21	6,72	-280,82	20,29	88,53	1,40	-0,81
12	Ionosphere	85,90	91,84	670,63	90,40	1,44	-584,73	86,75	90,79	1,05	-0,85
13	Iris	16,00	98,00	69,84	95,33	2,67	-53,84	18,54	97,33	0,67	-2,54
14	Mammogr.	78,55	84,33	132,54	83,37	0,96	-53,99	106,74	80,49	3,84	-28,18
15	Newthyroid	52,20	96,46	97,75	95,35	1,11	-45,55	56,47	94,60	1,86	-4,27
16	Pima	55,49	76,95	270,64	74,66	2,29	-215,15	57,20	77,05	-0,10	-1,71
17	Saheart	64,88	70,49	525,21	70,92	-0,43	-460,33	110,84	70,13	0,36	-45,96
18	Sonar	48,42	79,75	524,60	77,00	2,75	-476,18	47,59	78,90	0,85	0,83
19	Tae	142,71	62,07	323,14	60,81	1,26	-180,43	215,92	60,78	1,29	-73,20
20	Vehicle	195,81	68,52	555,77	64,89	3,63	-359,96	382,12	66,16	2,36	-186,31
21	Wdbc	40,82	96,78	183,70	95,14	1,64	-142,88	44,27	94,90	1,88	-3,45
22	Wine	34,98	98,70	170,94	93,98	4,72	-135,96	58,99	93,03	5,67	-24,00
23	Wisconsin	75,29	97,09	328,02	96,46	0,63	-252,73	69,11	96,35	0,74	6,18
	Trung bình	**110,47**	**83,05**	**355,23**	**80,66**			**163,40**	**80,57**		

Bảng 5. *So sánh độ chính xác của hệ phân lớp **FRBC_CS** so với **Product-1-ALL TUN** và **PAES-RCS** bằng kiểm định Wilcoxon với $\alpha = 0,05$*

So sánh (α = 0,05)	R^+	R^-	Exact *P*-value	Hypothesis
FRBC_CS vs **PAES-RCS**	273,0	3,0	1.192E-6	Rejected
FRBC_CS vs **Product-1-ALL TUN**	275,0	1,0	4.768E-7	Rejected

Bảng 6. *So sánh độ phức tạp của hệ phân lớp **FRBC_CS** so với **Product-1-ALL TUN** và **PAES-RCS** bằng kiểm định Wilcoxon với $\alpha = 0,05$*

So sánh (α = 0,05)	R^+	R^-	Exact *P*-value	Hypothesis
FRBC_CS vs **PAES-RCS**	276,0	0,0	2.384E-7	Rejected
FRBC_CS vs **Product-1-ALL TUN**	246,0	30,0	4.752E-4	Rejected

4. Kết luận

Đảm bảo tính giải nghĩa được của hệ dựa trên luật mờ nói chung và của hệ phân lớp dựa trên luật mờ nói riêng đóng vai trò quan trọng đảm bảo ngữ nghĩa tính toán phản ánh đúng tính mờ của thông tin và quá trình xử lý thông tin được nhất quán. Bài báo trình bày một phương pháp thiết kế ngữ nghĩa tính toán dựa trên tập mờ của các từ ngôn ngữ đảm bảo tính giải nghĩa của hệ phân lớp dựa trên luật mờ. Bằng các kết quả thực nghiệm và kết luận so sánh bằng phương pháp kiểm định giả thuyết thống kê Wilcoxon cho thấy tính hiệu quả của phương pháp biểu diễn này khi áp dụng thiết kế hệ phân lớp dựa trên luật mờ.

Lời cám ơn

Nghiên cứu này được tài trợ bởi Trường Đại học Giao thông vận tải trong đề tài mã số T2022-CN-001TĐ.

TÀI LIỆU THAM KHẢO/ REFERENCES

[1] R. Alcalá, Y. Nojima, F. Herrera, and 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, and 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 and 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, 2004.

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

[8] C. H. Nguyen and V. L. Nguyen, "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] C. H. Nguyen, W. Pedrycz, T. L. Duong, and T. S. Tran, "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] C. H. Nguyen, T. S. Tran, and 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*, vol. 67, pp. 244–262, 2014.

[11] C. H. Nguyen, V. T. Hoang, and V. L. Nguyen, "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. S. Tran and T. A. Nguyen, "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] V. T. Hoang, C. H. Nguyen, D. D. Nguyen, D. P. Pham, and V. L. Nguyen, "The interpretability and scalability of linguistic-rule-based systems for solving regression problems," *International Journal of Approximate Reasoning*, vol. 149, pp. 131-160, 2022.

[14] C. H. Nguyen, T. L. Pham, N. T. Nguyen, C. H. Ho, and T. A. Nguyen, "The linguistic summarization and the interpretability, scalability of fuzzy representations of multilevel semantic structures of word-domains," *Microprocessors and Microsystems*, vol. 81, 2021, Art. no. 103641.

[15] H. L. Bui, M. N. Pham, and T. T. H. Nguyen, "Swing-up control of an inverted pendulum cart system using the approach of Hedge-algebras theory," *Soft Computing*, vol. 26, pp. 4613–4627, 2022.

[16] H. H. Ngo, C. H. Nguyen, and V. Q. Nguyen, "Multichannel image contrast enhancement based on linguistic rule-based intensificators," *Applied Soft Computing Journal*, vol. 76, pp. 744–762, 2019.

[17] D. D. Nguyen, D. P. Pham, V. T. Hoang, and C. H. Nguyen, "A design method of scalable fuzzy rule-based systems for solving regression problems," *TNU Journal of Science and Technology*, vol. 226, no. 11, pp. 341-348, 2021.

[18] A. Tarski, A. Mostowski, and R. Robinson, *Undecidable Theories*. North-Holland, 1953.

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

[20] D. P. Pham, C. H. Nguyen, and 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, pp. 12–17.