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

Hoàng Văn Thông[1], Nguyễn Đức Dư[1], Nguyễn Cát Hồ[2]
[1]Khoa Công nghệ Thông tin, Trường Đại học Giao thông Vận tải
[2]Trung tâm Nghiên cứu và Phát triển, Trường Đại học Duy Tân
E-mail: thonghv@utc.edu.vn, nducdu@utc.edu.vn, ncatho@gmail.com
Tác giả liên hệ: Hoàng Văn Thông
Ngày nhận: 12/07/2017, ngày sửa chữa: 16/10/2017, ngày duyệt đăng: 17/10/2017

Tóm tắt: Trong bài báo này, chúng tôi đề xuất một phương pháp thiết kế ngữ nghĩa dạng tập mờ bằng hàm S và đại số gia tử mở rộng cho các từ ngôn ngữ sử dụng trong các hệ dựa trên các luật mờ (FRBS). Áp dụng phương pháp thiết kế này, phương pháp sinh luật từ dữ liệu của C. H. Nguyen và cộng sự và thuật toán tiến hóa (2+2)M-PAES của Knowles và Corne, chúng tôi phát triển một thuật toán xây dựng các FRBS giải bài toán hồi qui. Kết quả thử nghiệm cho thấy thuật toán sinh ra các FRBS có độ chính xác cao hơn các thuật toán được đối sánh.

Từ khóa: *Hàm thuộc, hệ dựa trên các luật mờ, đại số gia tử, bài toán hồi qui.*

Title: **A Method to Design Semantic of Linguistics based on the Enlarged Hedge Algebra and Applied to Building FRBS for Solving Regression**

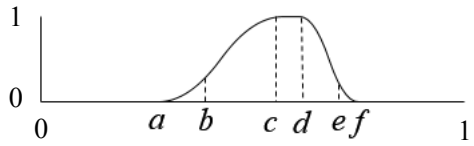
Abstract: This paper proposes a method to design semantic of linguistics based on fuzzy set using S-function and the enlarged hedge algebra for words included in fuzzy rule based systems (FRBSs). By using this method, the method generating rules from data of C. H. Nguyen et al., and the evolution algorithm M-PAES(2+2) of Knowles and Corne, we develop an algorithm to build FRBSs for solving regressions. Experimental results showed that the proposed algorithm generated FRBSs that are more accurate than those by other algorithms.

Keywords: *Membership function, fuzzy rule based systems, hedge algebras, regression problems.*

I. MỞ ĐẦU

Hệ dựa trên luật mờ (FRBS: Fuzzy Rule Based System) đã được ứng dụng thành công trong một số lĩnh vực như điều khiển [1], hay trong một số phương pháp như phân lớp [2–6], và hồi qui [7–11]. Nền tảng của FRBS dựa trên lý thuyết tập mờ (FS: Fuzzy Set). Vào năm 1965, trong [12], lần đầu tiên Zadeh đề xuất một phương pháp biểu diễn ngữ nghĩa tính toán của từ ngôn ngữ bằng FS. Mỗi FS được biểu diễn bằng một hàm toán học có giá trị biến thiên liên tục trong đoạn $[0, 1]$, được gọi là hàm thuộc (membership function). Chúng ta có thể sử dụng nhiều dạng hàm toán học khác nhau để biểu diễn các FS. Việc sử dụng các FS có dạng khác nhau trong các FRBS sẽ ảnh hưởng đến độ chính xác (accuracy) và tính giải nghĩa được (interpretability) của FRBS trong các bài toán cụ thể. Hầu hết các ứng dụng FRBS đều sử dụng hàm thuộc có dạng tam giác như trong [1, 2, 4, 6–10] do tính dễ giải nghĩa được của nó. Một số sử dụng hàm thuộc hình chuông (bell-function) [11] hay hàm hình thang [5]. Việc thiết kế các FS theo hướng tiếp cận dựa trên lý thuyết FS chủ yếu dựa trên kinh nghiệm của các chuyên gia, tuy nhiên thiếu một phương pháp hình thức đầy đủ để xác định FS từ tập từ ngôn ngữ.

Một hướng tiếp cận khác xác định ngữ nghĩa tính toán của từ ngôn ngữ được C. H. Nguyen và W. Wechler đề xuất trong [13] vào năm 1990. Trong đó, C. H. Nguyen và cộng sự xây dựng một đại số, gọi là đại số gia tử (ĐSGT) (hedge algebras), bao gồm xây dựng một ánh xạ từ tập từ ngôn ngữ vào đoạn $[0, 1]$. Ánh xạ này được gọi là ánh xạ ngữ nghĩa và giá trị của nó là giá trị định lượng của từ (ngữ nghĩa toán học của từ). Dựa trên các giá trị định lượng của từ được xác định bằng ánh xạ định lượng ngữ nghĩa, C. H. Nguyen và cộng sự đã đề xuất phương pháp xác định ngữ nghĩa FS



Hình 1. Hình dạng tập mờ biểu diễn ngữ nghĩa từ.

tam giác bằng phân hoạch đơn thể hoặc đa thể hạt [4, 6]. Theo hướng tiếp cận này, ngữ nghĩa dạng FS được xác định bằng một phương pháp hình thức đầy đủ từ tập từ sử dụng của biến ngôn ngữ chỉ bằng một số tham số mờ ngôn ngữ. Phương pháp thiết kế này đã được áp dụng để giải các bài toán phân lớp và hồi qui, được công bố trong các công trình [4, 6, 10], cho kết quả tốt hơn các phương pháp đối sánh có hướng tiếp cận theo lý thuyết FS. Tuy nhiên, việc sử dụng FS tam giác biểu diễn ngữ nghĩa của từ dường như không thực sự phù hợp với ngữ nghĩa thực của từ, vì FS tam giác chỉ có một điểm có giá trị hàm thuộc bằng 1. Trong thực tế, chúng ta thấy rằng hầu hết các từ ngôn ngữ đều có thể có nhiều giá trị chắc chắn thuộc vào nó, ví dụ giá trị *old* của biến tuổi *Age*, với những người có tuổi lớn hơn 90 thì đều chắc chắn thuộc vào *old*, hay nói cách khác là hàm thuộc của các giá trị này phải là 1. Trong [14], các tác giả đưa ra khái niệm hạt thông tin đã khẳng định điều này. Như vậy, một FS phù hợp với ngữ nghĩa của từ phải có nhiều hơn một điểm có giá trị hàm thuộc bằng 1, ví dụ như FS hình thang.

Để xây dựng các FS hình thang, trong [5], C. H. Nguyen và cộng sự đã phát triển một ĐSGT gọi là ĐSGT mở rộng. Trong đó, các tác giả xây dựng một ánh xạ ngữ nghĩa định lượng khoảng, giá trị ngữ nghĩa này được gọi là ngữ nghĩa lõi. Dựa trên ngữ nghĩa lõi của từ, các tác giả xây dựng FS hình thang, áp dụng giải bài toán phân lớp. Sử dụng FS hình thang biểu diễn ngữ nghĩa toán học của từ ngôn ngữ phù hợp hơn so với FS tam giác. Tuy nhiên, theo chúng tôi, khi sử dụng tam giác hay hình thang thì sự biến thiên của nó cũng chưa thực sự mềm dẻo như ngữ nghĩa vốn có của từ, do các cạnh là các hàm tuyến tính. Để FS biểu diễn phù hợp hơn ngữ nghĩa vốn có của từ thì sự biến thiên phải là phi tuyến. Trong bài báo này, chúng tôi đề xuất một phương pháp thiết kế ngữ nghĩa FS dựa trên ĐSGT mở rộng và hàm S (S-function). Dựa trên đó chúng tôi phát triển một thuật toán, viết tắt là EnHA-PAES-SF. Cũng trong thuật toán này, chúng tôi cải tiến toán tử đột biến thêm luật của thuật toán HA-PAES-MG-K_{max} trong [6]. Kết quả thử nghiệm cho thấy, FRBS được sinh ra với phương pháp được đề xuất có độ chính xác cao hơn các phương pháp của Alcalá và cộng sự trong [7], hay phương pháp của C. H. Nguyen và cộng sự trong [6, 10].

Bài báo được bố cục như sau: Mục II đề xuất phương pháp thiết kế ngữ nghĩa của từ ngôn ngữ, mục III trình bày bài toán hồi qui và hệ luật mờ Mamdani, mục IV trình bày thuật toán tiến hóa, mục V trình bày kết quả thử nghiệm, và mục VI kết luận bài báo.

II. PHƯƠNG PHÁP THIẾT KẾ NGỮ NGHĨA FS VÀ PHÂN HOẠCH MỜ DỰA TRÊN ĐSGT MỞ RỘNG

Trong mục này, chúng tôi đề xuất một phương pháp thiết kế ngữ nghĩa FS dựa trên ĐSGT mở rộng và hàm S.

Với ĐSGT mở rộng $\mathcal{AX}^{en} = (\mathbf{X}^{en}, \mathbf{C}, \mathbf{G}, \mathbf{H}^{en}, \leq)$, trong đó $\mathbf{C} = \{0, 1, w\}$ là tập các giá trị hằng với ý nghĩa là phần tử bé nhất, phần tử lớn nhất và phần tử trung hòa (neutral) trong $\mathbf{X}$, $\mathbf{G} = \{c^-, c^+\}$ là tập các phần tử sinh, $\mathbf{H}^{en}$ là tập các gia tử, $\mathbf{H}^{en} = \{h_0\} \cup \{Little, Very\}$, trong đó *Little* (L) là gia tử âm, và *Very* (V) là gia tử dương. Tập các từ ngôn ngữ được sinh ra bằng $\mathcal{AX}^{en}$, trong đó ngữ nghĩa định lượng khoảng của từ được xác định bằng các định lý 5.2 và 5.3 trong [5].

Áp dụng phương pháp thiết kế phân hoạch dạng đa thể hạt đề xuất trong [6], chúng tôi thực hiện thiết kế ngữ nghĩa FS của từ như sau. Giả sử ta có các từ ngôn ngữ A_{j-1}, A_j, $A_{j+1} \in X_k$ (tập từ có độ dài đúng bằng k), các giá trị cận trái lần lượt là $Lft(v_{fm}(A_{j-1}))$, $Lft(v_{fm}(A_j))$, $Lft(v_{fm}(A_{j+1}))$ và độ dài khoảng ngữ nghĩa lần lượt là $v_{fm}(A_{j-1})$, $v_{fm}(A_j)$, $v_{fm}(A_{j+1})$. Trong đó độ dài khoảng ngữ nghĩa của từ x được xác định theo công thức: $v_{fm}(x) = \mathcal{J}(h_0 x)$, với $\mathcal{J}(h_0 x)$ là khoảng tính mờ của từ x và giá trị cận trái của $v_{fm}(x)$ được xác định theo định lý 5.3 [5]:

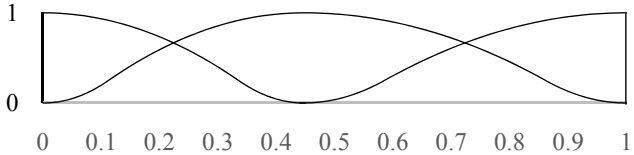
Cho ĐSGT mở rộng $\mathcal{AX}^{en}$ với $H^+ = \{h_j : 1 \leq j \leq p\}$ và $H^- = \{h_j : -1 \leq j \leq -q\}$, đặt $\beta = \sum_{j=1}^{p} \mu(h_j)$ và $\alpha = \sum_{j=-1}^{-q} \mu(h_j)$, ta có $\alpha + \beta + \mu(h_0) = 1$.

1) Với từ có độ dài 1 thì $Lft(v_{fm}(0)) = 0$; $Lft(v_{fm}(c^-)) = fm(0) + \beta fm(c^-)$; $Lft(v(W)) = fm(0) + fm(c^-)$; $Lft(v_{fm}(c^+)) = Lft(v_{fm}(W)) + fm(W) + \alpha fm(c^+)$ và $Lft(v_{fm}(1)) = 1 - fm(1)$.
2) Với từ có độ dài lớn hơn 1 thì

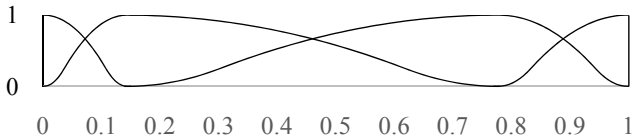
$$\begin{aligned} Lft(v_{fm}(h_j x)) &= v(x) \\ &+ \operatorname{sign}(h_j x) fm(x) \Big\{ 1/2(1 + \operatorname{sign}(h_j x))\mu(h_0) \\ &+ \sum_{i=\operatorname{sign}(j)}^{j} \mu(h_i) - \sigma(x)\mu(h_j) \Big\}. \end{aligned}$$

Khi đó chúng ta xây dựng FS biểu diễn ngữ nghĩa của từ A_j có dạng như Hình 1. Đỉnh FS là đoạn $[c, d] = v_{fm}(A_j)$. Nửa bên trái là hàm S, được xác định bởi

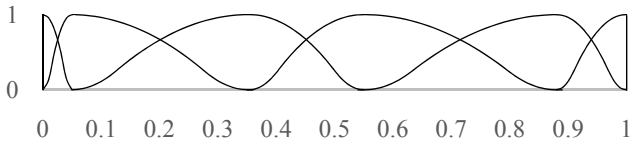
$$S_{\text{left}}(x, a, b, c) = \begin{cases} 0, & 0 \leq x \leq a, \\ \dfrac{(x-a)^2}{(b-a)(c-a)}, & a \leq x \leq b, \\ 1 - \dfrac{(x-c)^2}{(c-b)(c-a)}, & b \leq x \leq c, \\ 1, & x \geq c, \end{cases} \quad (1)$$



Phân hoạch mức-0, tập mờ tương ứng với từ $\mathbf{0}_0$, w, $\mathbf{1}_0$



Phân hoạch mức-1, tập mờ tương ứng với từ $\mathbf{0}_1$, c^-, c^+, $\mathbf{1}_1$



Phân hoạch mức-2, tập mờ tương ứng với từ $\mathbf{0}_2$, Vc^-, Lc^-, Lc^+, Vc^+, $\mathbf{1}_2$

Hình 2. Một ví dụ phân hoạch đa thể hạt với tập mờ S-function được đề xuất trong bài báo.

trong đó $a = Lft(v_{fm}(A_{j-1})) + v_{fm}(A_{j-1})$, $c = Lft(v_{fm}(A_j))$ và $b = a + (c - a)/4$. Nửa bên phải là hàm S, được xác định bởi

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

trong đó $d = c + v_{fm}(A_j)$, $f = Lft(v_{fm}(A_{j+1}))$ và $e = d + (f - d)/4$.

Ví dụ, Hình 2 minh họa thiết kế phân hoạch FS dựa trên ĐSGT mở rộng và hàm S, với các tham số mờ của ĐSGT $fm(0) = 0,006352412$, $fm(c_j^-) = 0,4404566$, $fm(w_j) = 0,0004077147$, $fm(c_j^+) = 0,4801931$, $fm(1) = 0,07259017$, $\mu L = 0,6891874$, $\mu h0 = 0,001883381$ và $k = 2$.

III. BÀI TOÁN HỒI QUI VÀ HỆ LUẬT MỜ MAMDANI

1. Bài toán hồi qui

Cho một tập mẫu dữ liệu $\mathbf{D} = \{(d_i, y_i), i = 1, \ldots, N\}$, trong đó d_i là một véc tơ F chiều có dạng $(d_{i1}, d_{i2}, \ldots, d_{iF})$, $d_{ij} \in U_j \subset \mathbb{R}$ (tập số thực) là miền xác định của các biến độc lập $\mathfrak{X}_j$ (thuộc tính đầu vào) của bài toán, với $j = 1, \ldots, F$, $y_i \in U_{F+1} \subset \mathbb{R}$ là miền xác định của biến phụ thuộc (thuộc tính đầu ra) $\mathfrak{X}_{F+1}$, N là số mẫu dữ liệu. Từ tập dữ liệu mẫu $\mathbf{D}$, xây dựng một hệ mờ cho phép tính giá trị $\hat{y} \in U_{F+1}$ ứng với mỗi giá trị đầu vào $d \in U = U_1 \times \cdots \times U_F$.

2. Hệ luật mờ Mamdani

Hệ luật mờ Mamdani với từ ngôn ngữ được sinh ra bằng ĐSGT mở rộng là một tập luật, gồm M luật, có dạng

$$\begin{aligned} R_m\text{: if } X_1 \text{ is } A_{1,j_m} \text{ and} \ldots \text{and } X_F \text{ is } A_{F,j_m} \\ \text{then } Y \text{ is } A_{F+1,j_m}, m = 1, \ldots, M \end{aligned} \tag{3}$$

trong đó

$$A_{f,j_m} \in L_f = \Big\{\{A_{f,0}\} \cup X_{k_f} = \{A_{f,1}, \ldots, A_{f,|X_{k_f}|}\}\Big\}, \quad f = 1, \ldots, F+1,$$

X_{k_f} là tập các từ ngôn ngữ có độ dài không quá k_f được sinh ra bằng ĐSGT mở rộng $\mathcal{AX}^{en}$, dùng để xây dựng phân hoạch thuộc tính thứ f, ví dụ trong Hình 2, $A_{f,0}$ kí hiệu giá trị *Don't care* với hàm thuộc đồng nhất bằng 1. Lưu ý rằng, L_{F+1} không chứa giá trị *Don't care*.

Giải bài toán hồi qui bằng hệ luật mờ Mamdani là xây dựng một hệ luật mờ với các luật dạng (3) để dự đoán giá trị đầu ra $\hat{y}$ ứng với giá trị đầu vào x có F chiều. Đánh giá độ chính xác của hệ luật dựa trên giá trị trung bình phương sai (MSE: Mean Squared Error) [6–8, 11], theo công thức

$$\text{MSE} = \frac{1}{2N} \sum_{i=1}^{N} (\hat{y}_i - y_i)^2, \tag{4}$$

trong đó $\hat{y}_i$ là giá trị suy diễn từ hệ luật với giá trị đầu vào x_i. Tương tự như các đề xuất [7, 8], chúng tôi sử dụng phương pháp trung bình trọng số để suy diễn, khi đó

$$\hat{y}_i = \frac{\sum_{m=1}^{M} \mu_m(x_i) \bar{A}_{F+1,j_m}}{\sum_{m=1}^{M} \mu_m(x_i)}, \quad i = 1, \ldots, N, \tag{5}$$

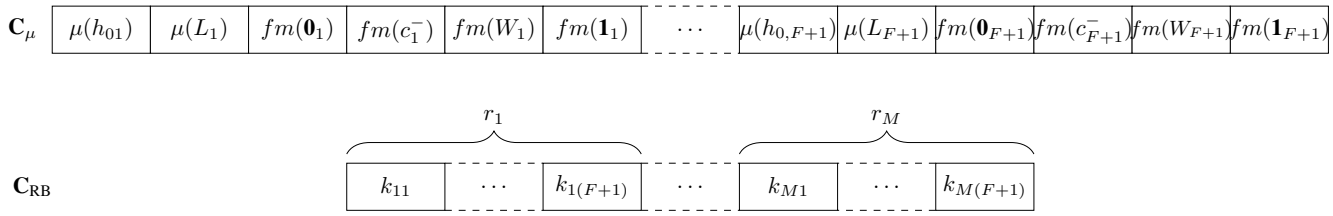
trong đó

$$\mu_m(x_i) = \prod_{f=1}^{F} \mu_{A_{F+1,j_m}}(x_{if}), \tag{6}$$

là độ đốt cháy luật thứ m của mẫu dữ liệu x_i, $\bar{A}_{F+1,j_m}$ là giá trị định lượng của hạng từ A_{F+1,j_m} và $\mu_{A_{F+1,j_m}}(\cdot)$ là hàm thuộc của từ ngôn ngữ A_{f,j_m}. Trong (5), nếu $\sum_{m=1}^{M} \mu_m(x_i) = 0$, có nghĩa là hệ luật *không phủ* mẫu dữ liệu x_i, khi đó $\hat{y}_i$ được suy diễn theo phương pháp đề xuất trong [7].

IV. THUẬT TOÁN TIẾN HÓA ĐA MỤC TIÊU

Trong mục này, chúng tôi trình bày tóm tắt thuật toán tiến hóa được áp dụng để sinh FRBS với phân hoạch mờ được thiết kế trong mục II. Chúng tôi gọi thuật toán đề xuất là EnHA-PAES-SF, được tóm tắt trong Thuật toán 1.



Hình 3. Cấu trúc mã hóa một cá thể.

1. Mã hóa cá thể

Chúng tôi sử dụng một gia tử âm L_j (*Little*) và một gia tử dương V_j (*Very*) trên mỗi biến $\mathfrak{X}_j$ và mục tiêu cần tối ưu là tham số tính mờ của ĐSGT và RB. Do đó mỗi cá thể của quần thể được mã hóa gồm hai phần ($\mathbf{C}_\mu, \mathbf{C}_{RB}$), như trong Hình 3, trong đó:

- $\mathbf{C}_\mu$: Biểu diễn các tham số tính mờ của các ĐSGT mở rộng $\mathcal{AX}^{en}$ tương ứng với các biến, là một véc tơ $\pi = (\pi_1, \ldots, \pi_{n+1})$, trong đó $\pi_j = \{\mu h0_j, \mu L_j, fm(0_j), fm(c_j^-), fm(w_j), fm(1_j)\}$, $j = 1, \ldots, F+1$. Như vậy $\mathbf{C}_\mu$ gồm $6 \times (F+1)$ gen các số thực.
- $\mathbf{C}_{RB}$: Biểu diễn RB, mỗi luật được mã hóa bằng một véc tơ gồm $F+1$ số nguyên. Giả thiết tập các từ được sử dụng trên mỗi biến là $\mathfrak{I}_j \cup \{Don't\ care\}$ được đánh chỉ số như sau: $\mathfrak{I}_j \cup \{Don't\ care\} = \{w_{jk}\colon k = 0, 1, \ldots, T\}$, $j = 1, \ldots, F+1$, trong đó w_{j0} = "*Don't care*", có nghĩa là chỉ số của từ "*Don't care*" là "0". Khi đó, một luật r_q sẽ được mã hóa như là véc tơ $(k_{q1}, \ldots, k_{qF}, k_{q(F+1)})$. Các luật của RB được sinh ra bằng hàm sinh luật GenRule$(d, \pi, l_{\max})$ [6], trong đó d là mẫu dữ liệu được chọn ngẫu nhiên trong tập dữ liệu huấn luyện, π là bộ tham số tính mờ trong $\mathbf{C}_\mu$ và $l_{\max}$ là chiều dài tối đa của luật do người dùng xác định trước.

Để giảm không gian tìm kiếm và tăng tính giải nghĩa của RB sinh, ta giới hạn số luật của mỗi RB nằm trong khoảng $[M_{\min}, M_{\max}]$.

Gắn với mỗi cá thể là một véc tơ hàm mục tiêu gồm hai thành phần (MSE, Comp), trong đó MSE biểu thị độ chính xác của hệ luật được xác định theo (4) và Comp là tổng độ dài của các luật trong RB.

2. Các toán tử di truyền

Tương tự như các thuật toán đề xuất trong [6, 7, 10], ở đây chúng tôi sử dụng lai ghép BLX-α trên phần tham số $\mathbf{C}_\mu$ và lai ghép one-point crossover trên phần $\mathbf{C}_{RB}$.

Với hai cá thể cha mẹ bất kỳ $\mathfrak{p}_1$ và $\mathfrak{p}_2$, thực hiện lai ghép để tạo ra 2 các thể con $\mathfrak{v}_1$, $\mathfrak{v}_2$:

- **Trên $\mathbf{C}_\mu$**: giả sử gen thứ i trên phần $\mathbf{C}_\mu$ của hai các thể con tương ứng là x_i và y_i, với $x_i, y_i \in [a_i, b_i) \subset [0, 1]$, $i = 1, \ldots, 6(F+1)$, để sinh ra các cá thể con $\mathfrak{v}_1$ và $\mathfrak{v}_2$ thì gen z_i trong phần $\mathbf{C}_\mu$ của chúng được chọn ngẫu nhiên trong đoạn $[l_i, u_i]$, trong đó $l_i = \max\{a_i, c_{\min} - I\}$ và $u_i = \min\{b_i, c_{\max} + I\}$, với $c_{\min} = \min\{x_i, y_i\}$, $c_{\max} = \max\{x_i, y_i\}$ và $I = (c_{\max} - c_{\min})\alpha$.
- **Trên $\mathbf{C}_{RB}$**: thực hiện lai ghép one-point crossover, điểm lai ghép được chọn ngẫu nhiên trong đoạn $[1, \rho_{\min} - 1]$, trong đó $\rho_{\min}$ là số luật ít nhất của hai cơ sở luật trong $\mathfrak{p}_1$ và $\mathfrak{p}_2$.
  Lưu ý: Nếu trên $\mathbf{C}_{RB}$ toán tử lai ghép không được thực hiện thì đột biến luôn xảy ra.

Đối với *toán tử đột biến*, với mỗi cá thể con $\mathfrak{v}_1$, $\mathfrak{v}_2$, thực hiện đột biến theo thứ tự và độc lập trên $\mathbf{C}_\mu$ và $\mathbf{C}_{RB}$.

- **Trên $\mathbf{C}_\mu$**: Lựa chọn ngẫu nhiên một số nguyên trong $[1, 6(F+1)]$, thực hiện thay thế gen tại vị trí được chọn bằng một giá trị được chọn ngẫu nhiên trong đoạn $[\mu h0_{\min}, \mu h0_{\max}]$ nếu gen được chọn biểu diễn giá trị $\mu h0$; chọn ngẫu nhiên trong đoạn $[\mu_{\min}, \mu_{\max}]$, nếu gen được chọn biểu diễn giá trị μL_j, chọn ngẫu nhiên trong đoạn $[fmh0_{\min}, fmh0_{\max}]$ nếu gen được chọn biểu diễn giá trị $fm(0_j)$, chọn ngẫu nhiên trong đoạn $[fmw_{\min}, fmw_{\max}]$ nếu gen được chọn biểu diễn giá trị $fm(w_j)$, chọn ngẫu nhiên trong đoạn $[fm1_{\min}, fm1_{\max}]$ nếu gen được chọn biểu diễn giá trị $fm(1_j)$, chọn ngẫu nhiên trong đoạn $[fmC_{\min}, fmC_{\max}]$ nếu gen được chọn biểu diễn giá trị của $fm(c_j^-)$.
  Lưu ý: Khi lai ghép hoặc đột biến trên phần $\mathbf{C}_\mu$:
  1) Nếu lai ghép/đột biến trên các gen μL_j và $\mu h0_j$ mà $1 - \mu L_j - \mu h0_j < \mu_{\min}$ thì phải thực hiện lại cho đến khi thỏa mãn bất đẳng thức $1 - \mu L_j - \mu h0_j \geq \mu_{\min}$ (điều kiện này nhằm đảm bảo $\mu V_j \in [\mu_{\min}, \mu_{\max}]$);
  2) Nếu lai ghép/đột biến trên các gen $fm(0_j)$, $fm(w_j)$, $fm(1_j)$ và $fm(c_j^-)$ mà $1 - fm(0_j) - fm(w_j) - fm(1_j) - fm(c_j^-) < fmC_{\min}$ thì phải thực hiện lại cho đến khi thỏa mãn bất đẳng thức $1 - fm(0_j) - fm(w_j) - fm(1_j) - fm(c_j^-) \geq fmC_{\min}$ (điều kiện này nhằm đảm bảo $fm(c_j^+) \in [fmC_{\min}, fmC_{\max}]$).

Các khoảng ràng buộc của các tham số ở trên được xác định trước bởi người dùng để bảo toàn ngữ nghĩa của các gia tử, các từ hằng và các phần tử sinh.

- **Trên $\mathbf{C}_{RB}$**: Áp dụng một trong hai toán tử sau đây trên mỗi cá thể, tức là nếu áp dụng toán tử thứ nhất thì không áp dụng toán tử thứ hai và ngược lại:
  1) Toán tử $\mathfrak{v}_{m1}$ thay đổi các giá trị tại δ gen của $\mathbf{C}_{RB}$: Lựa chọn ngẫu nhiên một số nguyên δ trong đoạn $[1, \delta_{\max}]$, trong đó $\delta_{\max}$ được xác định trước, sau đó chọn ngẫu nhiên δ gen của $\mathbf{C}_{RB}$. Thay đổi giá trị tại mỗi gen được chọn bằng một giá trị ngẫu nhiên, hoặc là trong $\{0, 1, \ldots, T\}$ nếu gen tương ứng với biến $j \neq F + 1$, hoặc trong $\{1, \ldots, T\}$ nếu gen tương ứng với biến $j = F + 1$;
  2) Toán tử $\mathfrak{v}_{m2}$ bổ sung γ luật vào RB hiện tại có M luật được biểu diễn bởi $\mathbf{C}_{RB}$: Chọn ngẫu nhiên một số nguyên γ trong $[1, \gamma_{\max}]$, $\gamma_{\max}$ được xác định trước bởi người dùng. Nếu $M + \gamma > M_{\max}$ thì bổ sung $\gamma' = \min\{\gamma, M_{\max} - M\}$ luật được sinh bằng hàm GenRule$(d, \pi, l_{\max})$. Trong đó, d được chọn từ tập mẫu huấn luyện $\mathbf{D}'$ (tập các mẫu dữ liệu không được phủ bởi RB), d được chọn từ tập mẫu huấn luyện $\mathbf{D}$ nếu toán tử lai ghép xảy ra trên $\mathbf{C}_{RB}$ hoặc khi $\mathbf{D}' = \phi$.

Lưu ý: Trong quá trình tiến hóa, nếu một luật trở nên có độ dài bằng 0, tức là phần tiền đề của nó đều là "*Don't care*", thì nó sẽ bị loại bỏ. Nếu RB có các luật trùng nhau thì chỉ giữ lại một. Việc này được thực hiện làm cho RB có cơ hội giảm độ phức tạp.

3. Thuật toán tiến hóa đa mục tiêu

Mặt Pareto, ký hiệu là β, chứa các cá thể cần lữu trữ. Thuật toán được phát triển dựa trên lược đồ tiến hóa (2+2)M-PAES trong [9].

Trong Thuật toán 1, hàm testAdd$(\beta, i_j, size)$ [9] bổ sung cá thể i_j vào mặt Pareto β nếu i_j không bị trội bởi bất kỳ cá thể nào trong β, khi đó những cá thể trong β bị trội bởi i_j sẽ bị loại ra khỏi β. Nếu bổ sung i_j vào làm số cá thể trong β lớn hơn *size* thì loại bỏ ngẫu nhiên một cá thể có trong β.

Thuật toán 1: EnHA-PAES-SF

```
begin
    β ← φ;
    Khởi tạo 2 cá thể con i1, i2
    testAdd(β, i1, size);
    testAdd(β, i2, size);
    for i ← t to maxGen do
        [p1, p2] ← chọn ngẫu nhiên 2 cá thể trong β
        v1 ← p1; v2 ← p2;
        if random() < Pcμ then
            [v1.C_μ, v2.C_μ] ← BLX-α
                                -Crossover(p1.C_μ, p2.C_μ)
        end if
        tmpPmRB ← 1;
        if random() < PcRB then
            [v1.C_RB, v2.C_RB] ← onepoint
                                -Crossover(p1.C_RB, p2.C_RB)
            tmpPmRB ← PmRB
        end if
        loop j = 1, 2
            if random() < Pmμ then
                mutation(vj.C_μ)
            end if
            if random() < tmpPmRB then
                if random() > PmAdd then
                    mutation_vm1(vj.C_RB)
                else
                    mutation_vm2(vj.C_RB)
                end if
            end if
            remove_zero_length(vj.C_RB)
            remove_duplicate_rule(vj.C_RB)
            testAdd(β, v1, size); testAdd(β, v2, size);
        end loop
    end for
    return β
end
```

V. KẾT QUẢ THỬ NGHIỆM

Chúng tôi tiến hành thử nghiệm thuật toán EnHA-PAES-SF trên 12 bài toán hồi qui trong Bảng I, được lấy từ http://archive.ics.uci.edu/ml/datasets.html, với #P là số mẫu, #A là số thuộc tính. Các tham số được thiết lập như trong Bảng II. Phương pháp thử nghiệm là 5-Fold (5-phần), trong đó 4 phần để học, 1 phần để kiểm tra. Mỗi phần thử nghiệm 6 lần, số lần thực nghiệm tổng cộng là $6 \times 5 = 30$ lần.

Kết quả thu được biểu thị bằng các mặt xấp xỉ tối ưu Pareto trung bình theo hai mục tiêu MSE và Comp của các RB ứng với 30 lần thử. Mặt Pareto trung bình được xác định như sau: mỗi lần thử nghiệm tạo ra một mặt xấp xỉ tối ưu Pareto, sắp xếp các điểm trên mặt Pareto theo thứ tự tăng dần của mục tiêu MSE trên tập huấn luyện. Trên mỗi mặt Pareto ta chỉ giữ lại số cá thể bằng số cá thể của mặt Pareto có số cá thể ít nhất trong 30 mặt Pareto, tính trung bình trên 30 mặt Pareto để tạo ra mặt Pareto trung bình.

Ký hiệu $\overline{\text{MSE}}_{\text{Tr}}$, $\overline{\text{MSE}}_{\text{Ts}}$, σ_{Tr}, σ_{Ts}, t_{Tr}, t_{Ts} lần lượt là giá trị MSE trung bình, độ lệch chuẩn, kết quả thống kê trên tập dữ liệu huấn luyện (Tr) và tập dữ liệu kiểm tra (Ts); $\overline{\text{Comp}}$ và $\overline{\#\text{R}}$ lần lượt là trung bình độ phức tạp và trung bình số luật của hệ luật.

Bảng I
CÁC BÀI TOÁN SỬ DỤNG THỬ NGHIỆM

TT	Bài toán	#P	#A
1	Electrical Length 1 (ELE1)	495	2
2	Electrical Maintainance 2 (ELE2)	1056	4
3	Weather Ankara (WA)	1609	9
4	Weather Izmir (WI)	1461	9
5	Treasury (TR)	1049	15
6	Abalone (AB)	4177	8
7	Mortgage (MTG)	1049	15
8	Computer Activity (CA)	8192	21
9	Pole Telecommunication (PT)	15000	26
10	AutoMPG6 (MPG6)	392	5
11	Baseball (BB)	337	16
12	Stock (ST)	950	9

Bảng II
CÁC THAM SỐ THỬ NGHIỆM

$\mu_{\min} = 0{,}3$	$size = 64$; $MaxGen = 300000$
$\mu_{\max} = 0{,}7$	$k_{\max} = 3$; $\gamma_{\max} = 5$; $\delta_{\max} = 5$
$M_{\min} = 5$	$M_{\max} = 30$; $fmw_{\min} = 0$; $fmw_{\max} = 0{,}1$
$fm0_{\min} = 0$	$fm0_{\max} = 0{,}1$; $fmC_{\min} = 0{,}3$; $fmC_{\max} = 0{,}7$
$PcRB = 0{,}3$	Xác suất lai ghép trên $\mathbf{C}_{RB}$
$Pc\mu = 0{,}5$	Xác suất lai ghép trên $\mathbf{C}_{\mu}$
$PmRB = 0{,}1$	Xác suất đột biến trên $\mathbf{C}_{RB}$
$P_{Add} = 0{,}75$	Xác suất đột biến thêm luật trên $\mathbf{C}_{RB}$
$Pm\mu = 0{,}3$	Xác suất đột biến trên $\mathbf{C}_{\mu}$
$\alpha = 0{,}5$	$l_{\max} = 5$

Chúng tôi thực hiện đối sánh kết quả thu được của thuật toán đề xuất với các thuật toán thuật toán HA-PAES-MG-$K_{\max}$ trong [6] với ngữ nghĩa của từ dạng FS tam giác, thuật toán EnHA-0-MG trong [10] với ngữ nghĩa của từ dạng FS hình thang trên 12 bài toán, thuật toán PAES-KB [7] trên 9 bài toán tại các điểm đại diện của mặt Pareto. Điểm ứng với hệ luật có $\overline{\text{MSE}}_{\text{Tr}}$ nhỏ nhất kí hiệu là FIRST, $\overline{\text{MSE}}_{\text{Tr}}$ trung bình kí hiệu là MEDIAN và $\overline{\text{MSE}}_{\text{Tr}}$ lớn nhất ký hiệu là LAST. Bên cạnh đối sánh các giá trị thu được, để khẳng định có sự khác biệt giữa các phương pháp được đối sánh, chúng tôi thực hiện phân tích thống kê bằng phương pháp t-test trên các giá trị $\overline{\text{MSE}}_{\text{Tr}}$ và $\overline{\text{MSE}}_{\text{Ts}}$. Do không gian trình bày của bài báo bị hạn chế nên trong phần này chúng tôi chỉ trình bày kết quả so sánh chi tiết tại điểm FIRST là điểm có độ chính xác cao nhất trong Bảng III, các điểm MEDIAN và LAST cũng cho kết quả tương tự điểm FIRST. Các ký hiệu trong bảng này được sử dụng tương tự như trong [8], "*" thể hiện dòng tương ứng cho kết quả tốt nhất với chữ in đậm, "+" thể hiện kết quả của dòng tương ứng kém hơn kết quả tốt nhất, và "=" thể hiện không có sự khác biệt thống kê của dòng tương ứng với kết quả tốt nhất.

Từ Bảng III ta thấy, về mục tiêu độ chính xác của FRBS, trên tập huấn luyện giá trị $\overline{\text{MSE}}_{\text{Tr}}$ của thuật toán đề xuất tốt hơn trên 8 bài toán (ELE1, TR, AB, MTG, PT, MPG6, BB, ST) và kết quả phân tích thống kê cho thấy có sự khác biệt trên cả 8 bài toán. Trên tập kiểm tra giá trị $\overline{\text{MSE}}_{\text{Ts}}$ của thuật toán được đề xuất tốt hơn trên 8 bài toán (WA, TR, AB, MTG, PT, MPG6, BB, ST), kết quả phân tích thống kê cho thấy có sự khác biệt với tất cả các thuật toán được so sánh trên 2 bài toán PT và BB. Có sự khác biệt trên 4 bài toán so với thuật toán PAES-KB. Mặc dù giá trị $\overline{\text{MSE}}_{\text{Ts}}$ trên tập kiểm tra của thuật toán đề xuất kém hơn thuật toán tốt nhất trên 4 bài toán nhưng sự sai khác không nhiều, cụ thể với bài toán ELE1 là 203525,234 (so với 194028,00), ELE2 là 10986,576 (so với 10460,014), WI là 0,925 (so với 0,846), và CA là 5,120 (so với 4,8662). Kết quả phân tích thống kê so sánh với kết quả tốt nhất không có sự khác biệt. Trong một số bài toán, giá trị $\overline{\text{MSE}}_{\text{Ts}}$ trên tập kiểm tra của thuật toán đề xuất tốt hơn nhiều, chẳng hạn trên bài toán PT là 58,171 (so với 66,57; 73,474 và 89), ST là 0,540 (so với 0,708 và 0,759). Kết quả lập luận trên tập dữ liệu kiểm tra của FRBS có tính ổn định cao hơn các thuật toán được so sánh. Điều này được thể hiện ở giá trị độ lệch chuẩn σ_{Ts}, độ lệch chuẩn của thuật toán đề xuất tốt hơn trên 8 bài toán.

Về mục tiêu tính giải nghĩa được của hệ luật, các FRBS được xây dựng bởi thuật toán được đề xuất trong bài báo có độ phức tạp (Comp) tốt hơn trên 6 bài toán. Các từ sử dụng trong FRBS là các từ ngôn ngữ được sinh ra bằng ĐSGT, nên góp phần làm cho FRBS dễ hiểu hơn với người dùng.

VI. KẾT LUẬN

Trích rút FRBS từ tập dữ liệu giải bài toán hồi qui là bài toán khá phức tạp, đòi hỏi chúng ta phải giải quyết nhiều bài toán: thiết kế ngữ nghĩa của từ ngôn ngữ, sinh luật, tìm kiếm FRBS. Nhiều nghiên cứu đã được đề xuất tiếp cận theo lý thuyết FS hoặc lý thuyết ĐSGT. Trong bài báo này, chúng tôi đề xuất một phương pháp thiết kế ngữ nghĩa dạng FS cho các từ ngôn ngữ sử dụng trong FRBS được sinh ra bằng ĐSGT mở rộng và hàm S. Áp dụng phương pháp thiết kế ngữ FS đề xuất, thuật toán sinh luật từ mẫu dữ liệu trong [6] và lược đồ tiến hóa (2+2)M-PAES trong [9], chúng tôi phát triển thuật toán EnHA-PAES-SF. Từ kết quả thử nghiệm thuật toán đề xuất, chúng ta có thể kết luận rằng phương pháp thiết kế ngữ nghĩa dạng FS của từ ngôn ngữ được đề xuất sinh ra các FRBS giải bài toán hồi qui có độ chính xác cao hơn ngữ nghĩa dạng tam giác, hình thang. Các phương pháp tiếp cận dựa trên ĐSGT sinh ra các FRBS có độ chính xác và tính giải nghĩa được cao hơn tiếp cận dựa trên lý thuyết tập mờ.

Bảng III
SO SÁNH KẾT QUẢ THỬ NGHIỆM THUẬT TOÁN ENHA-PAES-SF (ENHA-SF) VỚI CÁC THUẬT TOÁN HA-PAES-MG-K_{max} (HA-TG), ENHA-0-MG (ENHA-TZ), PAES-KB TẠI ĐIỂM FIRST

Bài toán	Thuật toán	$\overline{\#R}$	$\overline{Comp}$	$\overline{MSE}_{Tr}$	σ_{Tr}	t_{Tr}	$\overline{MSE}_{Ts}$	σ_{Ts}	t_{Ts}
ELE1	PAES-KB	27,00	46,00	145995,000	**3855,000**	+	**194028,000**	**24745,000**	*
	HA-Tg	27,27	46,13	141666,344	5618,006	+	202590,688	35321,180	=
	EnHA-Tz	17,33	**28,03**	146715,328	5401,636	+	201659,063	38816,035	=
	EnHA-SF	**25,80**	43,53	**134296,125**	5393,771	*	203525,234	34480,344	=
ELE2	PAES-KB	30,00	65,00	11043,000	2771,000	=	12606,000	3105,000	+
	HA-Tg	29,93	66,97	8813,093	**1533,687**	=	10685,542	3113,682	=
	EnHA-Tz	**26,60**	60,90	**8476,874**	2077,039	*	**10460,014**	3577,895	*
	EnHA-SF	29,57	**57,80**	8814,062	1484,897	=	10986,576	**2239,502**	=
WA	PAES-KB	28,00	103,00	1,640	**0,040**	+	3,920	9,270	+
	HA-Tg	25,00	60,03	1,032	0,071	+	1,247	0,172	=
	EnHA-Tz	24,83	74,83	**0,964**	0,065	*	1,136	0,154	=
	EnHA-SF	**24,37**	**53,60**	1,018	0,079	+	**1,103**	**0,130**	*
WI	PAES-KB	25,00	91,00	1,300	**0,270**	+	1,490	0,260	+
	HA-Tg	24,87	61,30	0,792	0,043	+	0,964	0,127	+
	EnHA-Tz	**23,17**	72,77	**0,718**	0,045	*	**0,846**	0,136	*
	EnHA-SF	25,00	**51,57**	0,773	0,048	+	0,925	**0,110**	+
TR	PAES-KB	**11,00**	40,00	0,080	0,040	+	0,140	0,150	+
	HA-Tg	15,00	**29,40**	0,031	0,004	+	0,045	0,022	+
	EnHA-Tz	25,23	84,70	0,028	0,006	+	0,044	0,028	=
	EnHA-SF	29,60	70,97	**0,023**	**0,002**	*	**0,034**	**0,010**	*
AB	PAES-KB	29,00	107,00	2,320	0,080	+	2,480	0,180	=
	HA-Tg	19,80	**59,57**	2,314	0,050	+	2,411	0,165	=
	EnHA-Tz	18,10	72,60	2,325	0,072	+	2,451	0,190	=
	EnHA-SF	**28,73**	109,47	**2,215**	**0,045**	*	**2,404**	0,151	*
MTG	PAES-KB	**12,00**	49,00	0,050	0,020	+	0,090	0,100	+
	HA-Tg	15,00	28,13	0,016	0,004	+	0,022	0,012	=
	EnHA-Tz	17,57	37,07	0,013	**0,002**	=	0,021	0,011	=
	EnHA-SF	18,00	**26,70**	**0,012**	0,003	*	**0,018**	**0,005**	*
CA	PAES-KB	**10,00**	**30,00**	11,990	2,990	+	13,430	4,660	+
	HA-Tg	13,80	44,67	4,577	0,369	=	**4,862**	0,630	*
	EnHA-Tz	10,53	35,70	**4,506**	0,337	*	4,905	0,760	=
	EnHA-SF	11,60	41,07	4,867	0,467	+	5,120	**0,545**	=
PT	PAES-KB	14,00	53,00	87,000	26,000	+	89,000	25,000	+
	HA-Tg	**13,33**	**38,30**	71,887	17,422	+	73,474	17,016	+
	EnHA-Tz	15,53	45,70	62,584	13,044	+	66,577	15,747	+
	EnHA-SF	14,13	44,67	**55,510**	**7,792**	*	**58,171**	**9,530**	*
MPG6	PAES-KB	-	-	-	-		-	-	
	HA-Tg	**30,00**	**69,23**	2,119	0,174	+	4,201	**0,844**	=
	EnHA-Tz	48,37	142,53	1,946	0,187	+	4,098	1,136	=
	EnHA-SF	49,40	116,03	**1,702**	**0,144**	*	**4,086**	1,042	*
BB	PAES-KB	-	-	-	-		-	-	
	HA-Tg	24,00	86,50	111254,578	**6668,565**	+	300933,594	**55664,977**	=
	EnHA-Tz	18,90	43,13	108939,016	10566,872	+	290203,406	112402,539	=
	EnHA-SF	**19,93**	**38,70**	**98882,602**	11293,128	*	**289854,875**	76645,250	*
ST	PAES-KB	-	-	-	-		-	-	
	HA-Tg	**23,43**	**83,30**	0,682	0,130	+	0,759	0,164	+
	EnHA-Tz	29,80	112,93	0,504	0,061	+	0,708	0,186	+
	EnHA-SF	48,73	154,23	**0,363**	**0,043**	*	**0,540**	**0,140**	*

LỜI CẢM ƠN

Nghiên cứu này nằm trong khuôn khổ đề tài "Nghiên cứu và phát triển các phương pháp thao tác trực tiếp trên các từ ngôn ngữ dựa trên đại số gia tử để giải quyết một số vấn đề trong các lĩnh vực trích rút tri thức, tăng cường chất lượng ảnh và cơ sở dữ liệu mờ", mã số 102.01-2017.06, được tài trợ bởi Quỹ phát triển khoa học và công nghệ quốc gia (NAFOSTED).

TÀI LIỆU THAM KHẢO

[1] D. Vukadinović, M. Bašić, C. H. Nguyen, N. L. Vu, and T. D. Nguyen, "Hedge-algebra-based voltage controller for a self-excited induction generator," *Control Engineering Practice*, vol. 30, pp. 78–90, 2014.

[2] H. Ishibuchi and Y. Nojima, "Analysis of interpretability-accuracy tradeoff of fuzzy systems by multiobjective fuzzy genetics-based machine learning," *International Journal of Approximate Reasoning*, vol. 44, no. 1, pp. 4–31, 2007.

[3] ——, "Repeated double cross-validation for choosing a single solution in evolutionary multi-objective fuzzy classifier design," *Knowledge-Based Systems*, vol. 54, pp. 22–31, 2013.

[4] 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.

[5] 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.

[6] 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.

[7] R. Alcalá, P. Ducange, F. Herrera, B. Lazzerini, and F. Marcelloni, "A multiobjective evolutionary approach to concurrently learn rule and data bases of linguistic fuzzy-rule-based systems," *IEEE Transactions on Fuzzy Systems*, vol. 17, no. 5, pp. 1106–1122, 2009.

[8] M. Antonelli, P. Ducange, B. Lazzerini, and F. Marcelloni, "Learning concurrently data and rule bases of Mamdani fuzzy rule-based systems by exploiting a novel interpretability index," *Soft Computing*, vol. 15, no. 10, pp. 1981–1998, 2011.

[9] J. D. Knowles and D. W. Corne, "Approximating the non-dominated front using the pareto archived evolution strategy," *Evolutionary Computation*, vol. 8, no. 2, pp. 149–172, 2000.

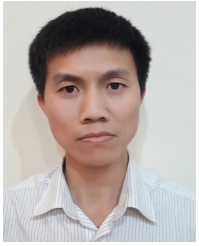
[10] C. H. Nguyen, V. T. Hoang, T. S. Tran, and V. L. Nguyen, "LFoC-Interpretability of Linguistic Rule Based Systems and its Applications To Solve Regression Problems," *International Journal of Computer Technology & Applications*, vol. 8, no. 2, pp. 94–117, 2017.

[11] P. Pulkkinen and H. Koivisto, "A dynamically constrained multiobjective genetic fuzzy system for regression problems," *IEEE Transactions on Fuzzy Systems*, vol. 18, no. 1, pp. 161–177, 2010.

[12] L. Sadeh, "Fuzzy set," *Information and Control*, vol. 8, pp. 338–353, 1965.

[13] N. C. Ho and W. Wechler, "Hedge algebras: an algebraic approach to structure of sets of linguistic truth values," *Fuzzy Sets and Systems*, vol. 35, no. 3, pp. 281–293, 1990.

[14] C. Mencar and A. M. Fanelli, "Interpretability constraints for fuzzy information granulation," *Information Sciences*, vol. 178, no. 24, pp. 4585–4618, 2008.



Hoàng Văn Thông sinh năm 1979 tại Ninh Bình. Ông tốt nghiệp Trường Đại học Khoa học Tự nhiên, Đại học Quốc gia Hà Nội, năm 2001 và nhận bằng Tiến sĩ, năm 2016, tại Học viện Khoa học và công nghệ, Viện Hàn lâm Khoa học và Công nghệ Việt Nam. Hiện nay, ông là giảng viên Khoa Công nghệ Thông tin, Trường Đại học Giao thông Vận tải. Lĩnh vực nghiên cứu hiện nay của ông là khai phá dữ liệu, lô gic mờ, hệ mờ, tính toán mềm, tính toán với từ, học máy.

Nguyễn Đức Dư sinh năm 1979 tại Nam Định. Ông tốt nghiệp Cử nhân năm 2001, Thạc sĩ năm 2004, tại Trường Đại học Khoa học Tự nhiên, Đại học Quốc gia Hà nội. Hiện nay, ông là giảng viên Khoa Công nghệ Thông tin, Trường Đại học Giao thông Vận tải. Lĩnh vực nghiên cứu hiện nay của ông là khai phá dữ liệu, lô gic mờ, tính toán mềm, tính toán với từ.

Nguyễn Cát Hồ sinh năm 1941 tại Hà Nội. Ông tốt nghiệp Trường Đại học Tổng hợp Hà Nội, năm 1960; nhận bằng Tiến sỹ tại Trường Đại học Tổng hợp Vacsava, Ba Lan, năm 1971 và nhận bằng Tiến sĩ Khoa học tại Trường Đại học Tổng hợp Kỹ thuật Dresden, Đức, năm 1987. Hiện nay, ông là cán bộ nghiên cứu, Trung tâm Nghiên cứu và Phát triển, Trường Đại học Duy Tân. Lĩnh vực nghiên cứu của ông là lô gic đại số, lô gic mờ, tính toán mềm, tính toán với từ, các hệ mờ và cơ sở dữ liệu mờ.