A DESIGN METHOD OF SCALABLE FUZZY RULE-BASED SYSTEMS FOR SOLVING REGRESSION PROBLEMS

Nguyen Duc Du[1*], Pham Dinh Phong[1], Hoang Van Thong[1], Nguyen Cat Ho[2]
[1]*University of Transport and Communications*
[2]*Duy Tan University*

ARTICLE INFO

Received: 27/7/2021
Revised: 30/8/2021
Published: 30/8/2021

KEYWORDS

Hedge algebras
Fuzzy rule-based system
Order-based semantics
Scalability
Interpretability

ABSTRACT

This paper proposes an approach for handling linguistic words directly to develop an evolutionary method for designing fuzzy rule-based systems interpretable in Tarski et al.'s sense and scalable to solve dataset regression problems. This interpretability requires that the constructed fuzzy multi-granularity structures representing the currently used word sets of dataset's attributes must be the isomorphic images of their respective semantic word sets' structures. Furthermore, in practice, human domain knowledge are accumulated and grown over time, leading to the requrements of expanding the currently used word sets to solve their encountered problems more effectively. It suggests studying behaviors of fuzzy rule-based systems when allowing the currently used word sets of dataset's attributes to grow while requiring the already constructed fuzzy sets based semantics of the existing linguistic words are reused. Experiments were conducted with 15 regression datasets to show the performance and advantages of the proposed method compared to the existing methods.

MỘT PHƯƠNG PHÁP XÂY DỰNG HỆ DỰA TRÊN LUẬT MỜ CÓ KHẢ NĂNG MỞ RỘNG GIẢI BÀI TOÁN HỒI QUY

Nguyễn Đức Dư[1*], Phạm Đình Phong[1], Hoàng Văn Thông[1], Nguyễn Cát Hồ[2]
[1]*Trường Đại học Giao thông vận tải*
[2]*Trường Đại học Duy Tân*

THÔNG TIN BÀI BÁO

Ngày nhận bài: 27/7/2021
Ngày hoàn thiện: 30/8/2021
Ngày đăng: 30/8/2021

TỪ KHÓA

Đại số gia tử
Hệ dựa trên luật mờ
Thứ tự ngữ nghĩa
Khả năng mở rộng
Tính giải nghĩa được

TÓM TẮT

Bài báo đề xuất tiếp cận tính toán trực tiếp trên từ ngôn ngữ để phát triển phương pháp tiến hóa thiết kế các hệ dựa trên luật mờ có tính giải nghĩa được theo quan điểm của Tarski và có thể mở rộng để giải bài toán hồi quy. Tính giải nghĩa này đòi hỏi rằng các cấu trúc đa thể hạt mờ được xây dựng biểu diễn ngữ nghĩa của tập từ được khai báo của các thuộc tính được sử dụng phải là hình ảnh đẳng cấu của cấu trúc ngữ nghĩa của tập từ tương ứng của chúng. Hơn nữa, trong thực tế, tri thức của con người được tích lũy và gia tăng theo thời gian dẫn đến nhu cầu mở rộng tập từ hiện được sử dụng để giải bài toán ứng dụng trong thực tiễn hiệu quả hơn. Nó gợi ý việc nghiên cứu các hành vi của các hệ dựa trên luật mờ khi cho phép gia tăng tập từ hiện được sử dụng của các thuộc tính trong khi vẫn đảm bảo các tập mờ đã được xây dựng được sử dụng lại. Các thực nghiệm được tiến hành với 15 tập dữ liệu hồi quy cho thấy tính hiệu quả và ưu điểm của phương pháp được đề xuất so với các phương pháp đã được công bố.

DOI: https://doi.org/10.34238/tnu-jst.4811

[*] Corresponding author. *Email: nducdu@utc.edu.vn*

1. Giới thiệu

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. Các hệ dựa trên luật mờ (fuzzy rule-based systems – FRBS) với ngữ nghĩa của các từ ngôn ngữ 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. 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 cho 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 bảo tính giải nghĩa được [1].

Trong bài báo này, chúng tôi nghiên cứu một phương pháp luận tính toán trực tiếp trên các từ ngôn ngữ theo tiếp cận Đại số gia tử [2], [3] để phát triển các thuật toán tiến hóa thiết kế các LRBS có tính giải nghĩa được theo quan điểm của Tarski [4]. Như vậy, khi thiết kế các LRBS cần có một cơ chế hình thức để xác định ngữ nghĩa tính toán của từ ngôn ngữ từ ngữ nghĩa định tính vốn có của nó [4]-[8], tức là các cấu trúc đa thể hạt mờ phải là hình ảnh đẳng cấu của cấu trúc 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 ngữ nghĩa của các từ ngôn ngữ của các biến ngôn ngữ phải giải nghĩa được [8]. Bên cạnh đó, vấn đề về khả năng mở rộng miền từ của biến ngôn ngữ sau khi đã được đưa vào ứng dụng cũng được nghiên cứu nhằm thiết kế các LRBS mới hiệu quả hơn dựa trên các LRBS đã được thiết kế và đang được áp dụng để giải các bài toán ứng dụng thực tế.

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 [4], khái niệm tính giải nghĩa được trong toán học và logic được thể hiện 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 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.

2.2. Cấu trúc ngữ nghĩa đa mức của miền từ ngôn ngữ vô hạn của các thuộc tính

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

Đại số gia tử (ĐSGT) được Nguyễn Cát Hồ và Wechler giới thiệu năm 1990 [2], [3]. Trong [5], các tác giả đã mở rộng ĐSGT truyền thống $\mathcal{A}^A$ thành ĐSGT mở rộng $\mathcal{A}^A_{en}$ bằng việc bổ sung một gia tử nhân tạo h_0 nhằm mô hình hóa lõi ngữ nghĩa của các từ ngôn ngữ.

Miền từ X^A bao gồm hai cấu trúc, cấu trúc ngữ nghĩa dựa trên thứ tự $T^A = (X^A_{en}, \leq)$ và cấu trúc *khái quát - đặc tả* $\boldsymbol{G}^A = (X^A_{en}, g)$. Hai cấu trúc này tạo thành cấu trúc ngữ nghĩa đa mức được biểu thị bằng $S^A = (X^A_{en}, \leq, g)$ và thể hiện dưới dạng bụi đa mức như trong Hình 1 được gọi là bụi ngữ nghĩa $\mathfrak{B}^A$ của S^A. $\mathfrak{B}^A$ là một cấu trúc có 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 $\mathfrak{B}^A_k$. Nó biểu diễn cấu trúc ngữ nghĩa của tập từ $X^A_{en,(k)}$.

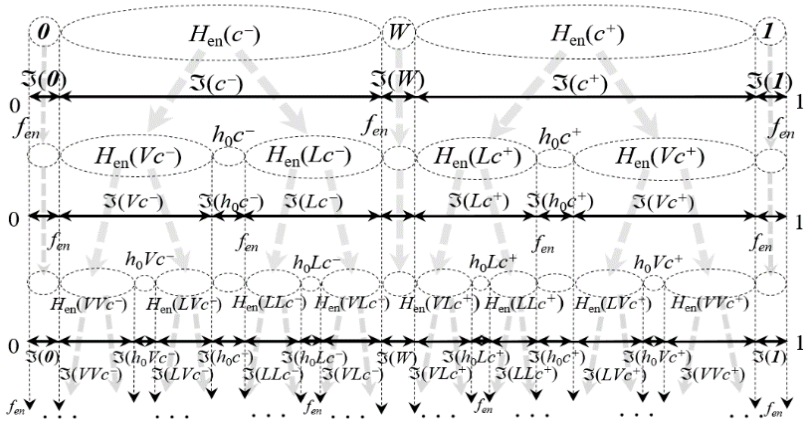
Muốn cấu trúc $T(X^A)$ biểu diễn cấu trúc $\boldsymbol{\mathcal{S}}^A = (X^A, \leq, g)$ bảo toàn cấu trúc của $\boldsymbol{\mathcal{S}}^A$ hay nói cách khác là $T(X^A)$ giải nghĩa được thì đòi hỏi định nghĩa hai quan hệ ký hiệu là $\leq$ và $\subset$ trên $T(X^A)$ vì $\boldsymbol{\mathcal{S}}^A$ có các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

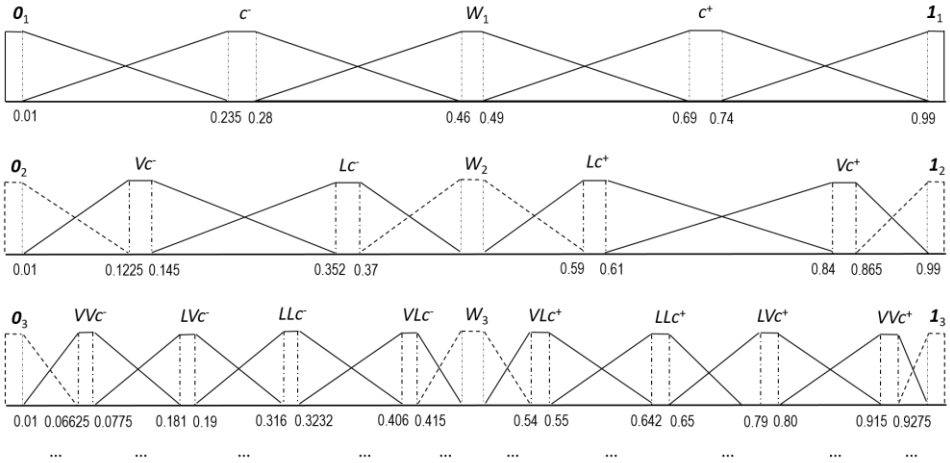
1) Quan hệ thứ tự ≤ trên T(X^A): Hai bộ ba t và t' với $t = (a, \boldsymbol{b}, c)$ và $t' = (a', \boldsymbol{b}', c')$ thỏa mãn $t \leq t'$ nếu và chỉ nếu các lõi của chúng thỏa mãn $\boldsymbol{b} = \boldsymbol{b}'$ hoặc $\boldsymbol{b} < \boldsymbol{b}'$ và thỏa ít nhất một trong các bất đẳng thức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

2) Quan hệ bao hàm ⊂ trên 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 T(X^A) với hai quan hệ ≤ và ⊂ được ký hiệu là M_{Gr}^A = (T(X^A), ≤, ⊂), 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.



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



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*

Trong [8] đã chứng minh được rằng, cấu trúc M_{Gr}^A như Hình 2 là hình ảnh đẳng cấu của cấu trúc ngữ nghĩa $\boldsymbol{S}^A = (X^A, \leq, g)$, tức là $\boldsymbol{S}^A$ có thể giải nghĩa được trong M_{Gr}^A.

2.2.2. Khả năng mở rộng của khung nhận thức ngôn ngữ (*LFoC*) *của biến ngôn ngữ*

Khái niệm Khung nhận thức ngôn ngữ (Lingistic Frame of Cognition - LFoC) được đưa ra trong [6]. Trong nghiên cứu này, LFoC F^A là một tập con hữu hạn của X^A nhằm nhấn mạnh yêu cầu về ngữ nghĩa của F^A phải là một cấu trúc con của toàn bộ cấu trúc ngữ nghĩa $\boldsymbol{S}^A$ của biến A.

Trong thực tiễn ứng dụng, ngữ nghĩa của các từ ngôn ngữ nhìn chung là không thay đổi, trong khi các tri thức đó vẫn gia tăng cùng với sự tồn tại và phát triển của xã hội. Vì vậy, chúng tôi đứng trên quan điểm các từ ngôn ngữ cùng có mặt trong các tập F_k^A, F_l^A và F^A có ngữ nghĩa như

nhau. Nó dẫn đến việc cần nghiên cứu *khả năng mở rộng* (khi mở rộng, ngữ nghĩa của các từ đang sử dụng không bị thay đổi) và dẫn đến vấn đề là liệu các quan hệ cấu trúc giữa các từ cùng có trong các cấu trúc $\boldsymbol{S}_k^A$, $\boldsymbol{S}_l^A$ và $\boldsymbol{S}^A$ cũng giống nhau, tức $\boldsymbol{S}_k^A$ là cấu trúc con của $\boldsymbol{S}_l^A$ và $\boldsymbol{S}_l^A$ là cấu trúc con của $\boldsymbol{S}^A$. Một câu hỏi cần đặt ra là liệu LFoC, F_k^A, có cấu trúc ngữ nghĩa không? Nếu có và được kí hiệu là $\boldsymbol{S}_k^A$ thì liệu nó có phải là cấu trúc con của $\boldsymbol{S}^A$? Ý nghĩa ứng dụng ẩn chứa trong đòi hỏi này được hiểu như sau: Cấu trúc F^A của biến ngôn ngữ A có tiềm năng vô hạn, nhưng tại thời điểm hiện tại của vòng đời của ứng dụng thường chỉ đòi hỏi sử dụng một tập con hữu hạn các từ, dạng F_k^A với mức đặc tả là k. Đứng trên quan điểm ngữ nghĩa định tính của mỗi từ x phải được xác định trong ngữ cảnh toàn miền F^A của biến ngôn ngữ thì về mặt phương pháp luận cần đòi hỏi việc tính toán trên cấu trúc $\boldsymbol{S}_k^A$ toàn cấu trúc $\boldsymbol{S}^A$, nghĩa là $\boldsymbol{S}_k^A$ phải là cấu trúc con của $\boldsymbol{S}^A$. Khi cần thiết, có thể mở rộng $\boldsymbol{S}_k^A$ bằng cách tăng mức đặc tả k.

3. Thiết kế tiến hóa hệ dựa trên luật mờ giải nghĩa được và có khả năng mở rộng

Bài toán hồi quy được phát biểu như sau: Cho tập dữ liệu $D = \{d_p = (a_{p,1}, a_{p,2}, \ldots, a_{p,n}, a_{p,(n+1)}) \in [0, 1]^{n+1} : p = 1,\ldots, N_D\}$ với n biến ngôn ngữ đầu vào A_j, $j = 1, \ldots, n$ và một biến ngôn ngữ đầu ra A_{n+1}, với các tập vũ trụ U_j được chuẩn hóa trong [0, 1]. LRBS giải bài toán hồi quy là một tập các luật mờ dạng **if-then**, mỗi luật mờ có dạng như sau:

$$r_q: \textbf{If } A_{j1} \text{ is } x_{q,j1} \,\&\, \ldots \,\&\, A_{jt} \text{ is } x_{q,jt} \textbf{ Then } A_{n+1} \text{ is } x_{q,n+1} \quad (1)$$

Trong đó, $x_{rq,j}$ là các từ ngôn ngữ trong X^{A_j} (đã bổ sung một giá trị "*Don'tcare*"), j=1, ..., n.

3.1. Mã hóa cá thể

Trong nghiên cứu này, chúng tôi chỉ sử dụng hai gia tử, trong đó có một gia tử âm L_j (*Little*) và một gia tử dương V_j (*Very*) trên mỗi biến ngôn ngữ A^j. Mục tiêu của thuật toán tiến hóa là đi tìm kiếm các bộ tham số tính mờ của ĐSGT mở rộng và LRBS tối ưu cho bài toán hồi quy. Mỗi cá thể của quần thể được mã hóa gồm hai phần C_μ và C_{RB}, trong đó:

- $\boldsymbol{C_\mu}$: Biểu diễn các tham số tính mờ của các ĐSGT mở rộng $\mathcal{A}_{en}^{A_j}$ tương ứng với các biến ngôn ngữ A^j, là một véctơ $\pi = (\pi_1, \ldots, \pi_{n+1})$, trong đó $\pi_j = \{\mu(h_{0j}), \mu(L_j), m(\mathbf{0}_j), m(c_j^-), m(W_j), m(\mathbf{1}_j)\}$, $j = 1, \ldots, n+1$. Như vậy, $\boldsymbol{C_\mu}$ gồm $6\times(n+1)$ gen các số thực.
- $\boldsymbol{C_{RB}}$: Biểu diễn cơ sở luật. Mỗi luật r_q được mã hóa bằng một véctơ gồm $n + 1$ số nguyên. Các luật của LRBS được sinh bằng thủ tục sinh luật *GenerateRule* tương tự thủ tục **Pr** trong [7].

Mỗi cá thể có hàm mục tiêu gồm hai thành phần (*MSE*, *Comp*), trong đó *MSE* là độ chính xác của LRBS được xác định theo (2) và *Comp* là tổng độ dài của các luật trong LRBS.

$$MSE = \frac{1}{2N_D}\sum_{p=1}^{N_D}(\hat{y}_p - y_p)^2 \quad (2)$$

Trong đó, $\hat{y}_p$ là giá trị suy diễn từ LRBS với giá trị đầu vào d_p theo công thức (3).

$$\hat{y}_p = \sum_{q=1}^{M}\mu_{A_q}(d_p)\,\bar{x}_{rq,(n+1)} \Big/ \sum_{q=1}^{M}\mu_{F_q}(d_p) \quad (3)$$

Trong đó, $\mu_{F_q}(d_p) = \prod_{j=1}^{n}\mu_{x_{rq,j}(a_{p,j})}$ là độ đốt cháy luật thứ q đối với mẫu dữ liệu d_p, $\bar{x}_{rq,(n+1)}$ là giá trị giải mờ của tập mờ có nhãn tập mờ $x_{rq,(n+1)}$ và $\mu_{x_{rq,j}}(.)$ là hàm thuộc của tập mờ tương ứng với nhãn ngôn ngữ $x_{rq,j}$. Nếu $\sum_{q=1}^{M}\mu_{F_q}(d_p) = 0$, có nghĩa là điểm dữ liệu d_p không bị phủ bởi luật nào thì $\hat{y}_i$ được xác định theo phương pháp lập luận của Alcalá đề xuất trong [9].

3.2. Các toán tử di truyền

Áp dụng toán tử lai ghép một điểm trên C_μ và C_{RB}. Thực hiện đột biến theo thứ tự và độc lập trên C_μ và C_{RB}. Với toán tử đột biến trên C_{RB}, áp dụng một trong hai toán tử thay đổi gen trên C_{RB} và thêm luật, tức là nếu áp dụng toán tử thứ nhất thì không áp dụng toán tử thứ hai và ngược lại.

Trong quá trình tiến hóa, nếu một luật bị thay đổi và có độ dài bằng 0, tức là phần tiền đề của nó đều là *"Don'tcare"* thì nó sẽ bị loại bỏ; nếu có các luật trùng nhau thì chỉ giữ lại một.

3.3. Thuật toán tiến hóa đa mục tiêu thiết kế LRBS có tính giải nghĩa và có khả năng mở rộng

Thuật toán *IS-LRBMOEA*(*D*, $Sem_{EnHA}(\mathcal{A}(D)$, *paretofile*)

Đầu vào: Tập dữ liệu $D = \{d_p = (a_{i,1}, a_{i,2}, \ldots, a_{i,n}, a_{i,(n+1)}): i = 1 \text{ to } N_D\}$;

– $Sem_{EnHA}(\mathcal{A}(D)$: ngữ nghĩa cú pháp của các biến ứng với các thuộc tính;

– Các xác suất lai ghép: $P_c(C_\mu)$ và $P_c(C_{RB})$, xác suất đột biến: $P_m(C_\mu)$ và $P_m(C_{RB})$, xác suất đột biến thêm luật $P_m\_Add\_RB$;

– *k*: Một mảng chứa mức đặc tả tối đa của các LFoC hiện được khai báo của các biến;

– $\tau_{\max}$: độ dài tối đa của luật, $M_{\min}$ và $M_{\max}$ tương ứng là số luật nhỏ nhất và lớn nhất của LRBS trên mặt Pareto, *MaxGen*: số thế hệ, *Paretofile*: tệp chứa mặt Pareto $\mathbb{P}$ cuối cùng;

Đầu ra: $\mathbb{P}$ – Các phương án tốt nhất trên mặt Pareto.

Begin

Bước 1: Khối khởi tạo: mục đích là xây dựng các LFoC, $\mathbb{P}$ khởi tạo.

If *paretofile == "" ***then***

For *h = 1 to 2* // 2 cá thể

B1.1. Sinh tập từ $X_{(k_j)}^{A_j}$ (LFoC) cho A_j và tập chỉ số $\mathbb{I}_{ex}(X_{(k_j)}^{A_j})$, j =1,.., n+1.

- Sinh ngẫu nhiên các giá trị của π_j = ($\mu(h_{0j})$, $\mu(L_j)$, $fm(\mathbf{0}_j)$, $fm(c_j^-)$, $fm(W_j)$, $fm(\mathbf{1}_j)$).

- Tính toán hệ khoảng tương tự $\mathbb{S}_{(k_j)}^{A_j} = \{\mathbb{S}_{(k_j)}^{A_j}(x): x \in X_{(k_j)}^{A_j}\}, j = 1, .., n + 1$.

B1.2. Xây dựng các cấu trúc đa thể hình thang của các LFoC như Hình 2.

B1.3. Sinh các luật ngôn ngữ từ mỗi mẫu dữ liệu, dựng C_{RB}.

- Sinh ngẫu nhiên một số nguyên $M_k \in [M_{\min}, M_{\max}]$ và gọi M_k lần *GenerateRule*(d_p, $\{\mathbb{S}_{(k_j)}^{Aj}, \mathbb{I}_{ex}(X_{(k_j)}^{Aj}: j \le n + 1\}$, $\tau_{\max}$) để sinh M_k luật với d_p được chọn ngẫu nhiên từ *D*.

B1.4. Tính giá trị MSE và độ phức tạp *Comp* và gán *h* vào $\mathbb{P}$.

End for

Else

B1.1b. *Phục hồi tệp tin "paretofile", tăng các LFoC được khai báo hiện tại đến các mức đặc tả mới và xây dựng các cấu trúc đa thể hình thang bổ sung ở các mức k.*

- Khôi phục $\mathbb{P}$ từ tệp tin "*paretofile*" chứa mặt Pareto của lượt chạy cuối cùng.
- Sinh LFoC $X_{(k_j)}^{Aj}$, j =1,.., n+1 nếu $A_j < k_j$.

B1.2b. Xây dựng các cấu trúc đa thể hình thang của các LFoC.

End if

Bước 2: Khối tiến hóa được lặp với MaxGen lần để lưu trữ một mặt Pareto tối ưu.

B2.1. *Tạo hai cá thể con (offspring)*

- Chọn ngẫu nhiên hai cá thể p_1 và p_2 từ $\mathbb{P}$ và áp dụng toán tử lai ghép để sinh hai cá thể con o_1 và o_2 từ p_1 và p_2 trên C_μ và C_{RB} với xác suất $P_c(C_\mu)$ và $P_c(C_{RB})$ tương ứng.
- Áp dụng toán tử đột biến trên C_μ độc lập đối với o_1 và o_2 với xác suất $P_m(C_\mu)$.
- Áp dụng các toán tử đột biến trên C_{RB} độc lập đối với o_1 và o_2 với xác suất $P_m(C_{RB})$.

If toán tử thêm luật được áp dụng với xác suất $P_m\_Add\_RB$ **then**

– Xây dựng các khoảng tương tự $\mathbb{S}_{(k_j)}^{A_j}$, $j = 1, \ldots, n + 1$.

– Xây dựng các cấu trúc đa thể hình thang của các LFoC như Hình 2.
– Áp dụng toán tử đột biến thêm luật P_m_Add_RB.

Else Áp dụng toán tử thay đổi cơ sở luật;

B2.2. Tính toán giá trị của tất cả các mục tiêu của o_1 và o_2

B2.3. Đưa từng o_1 và o_2 vào $\mathbb{P}$ nếu chúng không bị trội hơn bởi bất kỳ phương án nào trong $\mathbb{P}$. Nếu $\mathbb{P}$ đầy, loại bỏ ngẫu nhiên cá thể thuộc về vùng có mật độ cao nhất.

***Bước 3. Lưu mặt Pareto*:** Ghi mặt Pareto $\mathbb{P}$ vào tập tin có tên là *"paretofile"*.

End.

4. Kết quả và bàn luận

4.1. Cài đặt thực nghiệm

Các tham số thực nghiệm: Các ràng buộc đối với giá trị của các tham số tính mờ của các biến: $0,3 \le fm(c^-)$, $\mu(L) \le 0,7$, $0 < fm(\mathbf{0})$, $fm(W) = fm(\mathbf{1}_j) \le 0,1$ và $0 < \mu(h_0) \le 0,2$. Giá trị của các tham số của thuật toán tiến hóa được đề xuất là giống nhau như trong Bảng 1, riêng các tập dữ liệu với số thuộc tính lớn hơn 10 thay đổi $\tau_{\max} = 5$ hoặc 8. Trong trường hợp thuật toán ***IS-LRBMOEA*** được sử dụng lại để thiết kế các LRBS tối ưu mới khi các LFoC gia tăng tới mức đặc tả cao hơn, số thế hệ tối đa là *MaxGen* = 200000.

Bảng 1. *Các tham số thực nghiệm*

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

- *Phương pháp thực nghiệm*: phương pháp kiểm tra chéo 5-fold được sử dụng. Mỗi fold được thực nghiệm 6 lần và ta có $6 \times 5 = 30$ lần thực nghiệm. Kết quả tổng hợp của 30 lần thực nghiệm được biểu thị bằng một mặt xấp xỉ tối ưu Pareto trung bình theo hai mục tiêu *MSE* và *Comp* của 30 lần thử. Phương pháp kiểm định thống kê Wilcoxon với mức ý nghĩa $\alpha = 0,05$ được sử dụng để kết luận về ý nghĩa so sánh giữa các phương pháp thiết kế LRBS cho bài toán hồi quy.

4.2. Kết quả mô phỏng thực nghiệm và bàn luận

Các tập dữ liệu thực nghiệm được lấy từ [10] bao gồm Electrical Length 1 (ELE1), Electrical Maintainance 2 (ELE2), Weather Ankara (WA), Weather Izmir (WI), Treasury (TR), Abalone (AB), Mortgage (MTG), Computer Activity (CA).

- *Thực nghiệm 1 chứng tỏ tính hiệu quả của biểu diễn đa thể hình thang có tính giải nghĩa và có khả năng mở rộng.*

Các kết quả thực nghiệm của phương pháp thiết kế LRBS với mức đặc tả *kmax* = 3 (độ dài lớn nhất của các từ ngôn ngữ là 3) được đề xuất (được ký hiệu là $\boldsymbol{A}_{Gr3}$) được so sánh với các kết quả thu được của các phương pháp thiết kế HA-PAES-MG-$K_{\max}$ với ngữ nghĩa dựa trên tập mờ tam giác trong [6] (được ký hiệu là HA3_Tg), ngữ nghĩa dựa trên tập mờ hình thang trong [7] (được ký hiệu là HA3_Tz) trên 9 tập dữ liệu đầu tiên trong danh sách trên tại điểm FIRST (điểm có giá trị MSE nhỏ nhất trên tập huấn luyện) trên mặt Pareto. Các phương pháp HA3_Tz và HA3_Tg đều thiết kế các LRBS có mức đặc tả *kmax* = 3.

Các kết quả thực nghiệm và so sánh của các phương pháp thiết kế LRBS này được thể hiện trong Bảng 2, trong đó cột Comp là độ phức tạp của LRBS, MSE_{tr} là giá trị MSE trên tập huấn luyện và MSE_{ts} là giá trị MSE trên tập kiểm tra. Trực quan ta thấy rằng, phương pháp $\boldsymbol{A}_{Gr3}$ có giá trị MSE_{ts} nhỏ hơn so với phương pháp HA3_Tz đối với 8 trên 9 tập dữ liệu được thực nghiệm và nhỏ hơn so với phương pháp HA3_Tg trên tất cả các tập dữ liệu được thực nghiệm.

Bảng 2. *Giá trị MSE trên tập huấn luyện và kiểm tra tại điểm FIRST*

Data set	Comp			MSE_{tr}			MSE_{ts}				
	A_{Gr3}	HA_3_Tz	HA_3_Tg	A_{Gr3}	HA_3_Tz	HA_3_Tg	A_{Gr3}	HA_3_Tz	Diff (%)	HA_3_Tg	Diff(%)
ELE1	47,57	**28,03**	46,13	**138060**	146715	141666	**193388**	201659	-4,10	202591	-4,54
ELE2	61,63	**60,90**	66,97	9065	**8477**	8813	**10337**	10460	-1,18	10686	-3,27
WA	**50,87**	74,83	60,03	0,99	**0,964**	1,03	**1,11**	1,14	-2,63	1,25	-11,20
WI	**56,07**	72,77	61,30	0,769	**0,718**	0,79	0,911	**0,85**	7,18	0,96	-5,10
TR	71,87	84,70	**29,40**	**0,021**	0,028	0,03	**0,035**	0,04	-12,50	0,04	-12,50
AB	107,03	72,60	**59,57**	**2,205**	2,325	2,31	**2,395**	2,45	-2,24	2,41	-0,62
MTG	31,73	37,07	**28,13**	**0,012**	0,013	0,02	**0,017**	0,02	-15,00	0,02	-15,00
CA	78,27	**35,70**	44,67	**4,446**	4,506	4,58	**4,74**	4,91	-3,46	4,86	-2,47
PT	68,00	45,70	**38,30**	**59,7**	62,584	71,89	**62,75**	66,58	-5,75	73,47	-14,59

Bảng 3. *So sánh độ phức tạp của LRBS tại điểm FIRST*

A_{Gr3} vs	R^+	R^-	Exact P-value	Hypoth. (H_0)
HA_3_Tz	16	29	>0,2	Không bị bác bỏ
HA_3_Tg	22	33	>0,2	Không bị bác bỏ

Bảng 4. *So sánh giá trị MSE_{ts} tại điểm FIRST*

A_{Gr3} vs	R^+	R^-	Exact P-value	Hypoth. (H_0)
HA_3_Tz	40	5	0,03906	Bị bác bỏ
A_3_Tg	45	0	0,003906	Bị bác bỏ

Kết quả kiểm định thống kê Wilcoxon với mức ý nghĩa $\alpha = 0,05$ đối với độ phức tạp và giá trị MSE_{ts} của các LRBS được thiết kế bởi các phương pháp $\boldsymbol{A}_{Gr3}$, HA3_Tz và HA3_Tg tương ứng được thể hiện trong Bảng 3 và Bảng 4. Dễ dàng thấy rằng, các giá trị *Exact P-value* trong Bảng 3 đều lớn hơn mức ý nghĩa $\alpha = 0,05$ nên các giả thuyết H_0 không bị bác bỏ. Do đó, độ phức tạp của các LRBS được thiết kế bởi các phương pháp là tương đương nhau. Các giá trị *Exact P-value* trong Bảng 4 đều nhỏ hơn mức ý nghĩa $\alpha = 0,05$ nên các giả thuyết H_0 bị bác bỏ. Do đó, ta có thể kết luận rằng, $\boldsymbol{A}_{Gr3}$ hiệu quả hơn so với HA3_Tz và HA3_Tg.

- *Thực nghiệm 2 chứng tỏ ý nghĩa của việc mở rộng của tập từ được khai báo trong miền giá trị của biến ngôn ngữ.*

Giả sử các LRBS đã được thiết kế với mức đặc tả *kmax* = 2, được ký hiệu là $\boldsymbol{A}_{Gr2}$. Người sử dụng mong muốn mở rộng các LFoC lên mức đặc tả *kmax* = 4 (được ký hiệu là $\boldsymbol{A}_{Gr2\uparrow4}$) nhằm gia tăng tri thức miền với kỳ vọng nâng cao độ chính xác của các LRBS. Do đó, sau khi tăng mức đặc tả lên 4, các LRBS hiện có trên mặt Pareto được tiếp tục tối ưu để thu được các LRBS mới có cơ sở luật đã được bổ sung các từ có độ dài 3 và 4. Việc mở rộng các LFoC được khai báo của các biến không làm phá vỡ cấu trúc phân hoạch đa thể biểu diễn cấu trúc ngữ nghĩa dựa trên tập mờ hình thang hiện tại như đã được trình bày ở trên.

Bảng 5. *Giá trị MSE trên tập huấn luyện và kiểm tra, so sánh giữa $\boldsymbol{A}_{Gr2\uparrow4}$ và $\boldsymbol{A}_{Gr2}$, $\boldsymbol{A}_{Gr3}$, P_{KB}*

Dataset	$\boldsymbol{A}_{Gr2\uparrow4}$		$\boldsymbol{A}_{Gr2}$		Diff(%)	$\boldsymbol{A}_{Gr3}$		Diff(%)	FSMOGFSe+TUNe		Diff(%)
	MSE_{tr}	MSE_{ts}	MSE_{tr}	MSE_{ts}		MSE_{tr}	MSE_{ts}		MSE_{tr}	MSE_{ts}	
ELE1	127041	214067	150665	202248	5,84	138060	193388	10,69	151600	195000	9,78
ELE2	7906	9709	9384	11512	-15,66	9064	10337	-6,08	9665	10548	-7,95
WAN	0,9053	1,1089	1,005	1,189	-6,74	0,99	1,11	-0,1	1,441	1,635	-32,18
WIZ	0,6817	0,8588	0,748	0,871	-1,4	0,769	0,911	-5,73	0,929	1,011	-15,05
TRE	0,0194	0,031	0,023	0,033	-6,06	0,021	0,035	-11,43	0,034	0,044	-29,55
ABA	2,1694	2,3817	2,281	2,429	-1,95	2,205	2,395	-0,56	2,445	2,509	-5,07
MOR	0,0079	0,0121	0,0129	0,0164	-26,22	0,012	0,017	-28,82	0,016	0,019	-36,32
CA	3,975	4,319	4,326	4,592	-5,95	4,446	4,74	-8,88	0,158	5,216	-17,2
POLE	51,08	55,34	60,87	66,1	-16,28	59,7	62,75	-11,81	100,85	102,81	-46,17
PLA	1,105	1,182	1,156	1,217	-2,88	1,142	1,21	-2,31	1,106	1,19	-0,67
FRIE	1,239	1,542	1,3	1,586	-2,77	1,502	1,917	-19,56	2,71	3,13	-50,73
MPG6	1,618	4,238	1,9	4,295	-1,33	1,768	3,994	6,11	2,86	4,56	-7,06
ANA	0,00193	0,00304	0,00236	0,00317	-4,1	0,00189	0,00345	-11,88	0,003	0,003	1,33
CON	17,22	22,6935	18,45	23,8489	-4,84	19,0	25,8277	-12,14	29,901	32,977	-31,18
MV	0,4024	0,4103	0,4978	0,5058	-18,88	0,5183	0,5224	-21,46	0,158	0,158	159,68
Tổng					**-109,22**			**-123,96**			**-108,34**

Các kết quả thực nghiệm được thể hiện trong Bảng 5 và kết quả kiểm định thống kê Wilcoxon với mức ý nghĩa $\alpha = 0,05$ đối với giá trị MSE_{ts} của các LRBS được thể hiện trong Bảng 6. Phân tích số liệu trong Bảng 5 ta thấy rằng, phương pháp $\boldsymbol{A}_{Gr2\uparrow4}$ có giá trị MSE_{ts} nhỏ hơn so với các phương pháp $\boldsymbol{A}_{Gr2}$, $\boldsymbol{A}_{Gr3}$ và FSMOGFSe+TUNe trong [11] tương ứng đối với 14, 13 và 12 trên 15 tập dữ liệu được thực nghiệm. Xét trên tỷ lệ phần trăm giá trị MSE_{ts} giảm, phương pháp $\boldsymbol{A}_{Gr2\uparrow4}$ có

tỷ lệ giảm tương ứng so với các $\boldsymbol{A}_{Gr2}$, $\boldsymbol{A}_{Gr3}$ và FSMOGFSe+TUNe là 109,22%, 123,96% và 108,34%. Như vậy ta thấy rằng, khi các LFoC ứng với các thuộc tính được mở rộng thì giá trị MSE_{ts} giảm về mặt tổng thể, tức là độ chính xác của các LRBS tăng lên. Kết quả kiểm định giả thuyết thống kê trong Bảng 6 cho thấy các giả thuyết H_0 đều bị bác bỏ. Do đó, ta có thể khẳng định rằng, phương pháp $\boldsymbol{A}_{Gr2\uparrow 4}$ tốt hơn so với các phương pháp $\boldsymbol{A}_{Gr2}$, $\boldsymbol{A}_{Gr3}$ và FSMOGFSe+TUNe.

Bảng 6. *So sánh giá trị MSE_{ts} tại các điểm FIRST*

So sánh	R^+	R^-	Exact P-value	Hypoth. (H_0)
$\boldsymbol{A}_{Gr2\uparrow 4}$ vs $\boldsymbol{A}_{Gr2}$	105	15	0,008362	Bị bác bỏ
$\boldsymbol{A}_{Gr2\uparrow 4}$ vs $\boldsymbol{A}_{Gr3}$	96	24	0,04126	Bị bác bỏ
$\boldsymbol{A}_{Gr2\uparrow 4}$ vs FSMOGFSe+TUNe	97	23	0,03534	Bị bác bỏ

5. Kết luận

Bài báo tập trung nghiên cứu vấn đề giải nghĩa được và khả năng mở rộng của các LRBS được trích rút từ dữ liệu số cho bài toán hồi quy. Tính giải nghĩa được của các LRBS được hiểu theo định nghĩa của Taski trong [4] và các thuật toán thiết kế LRBS phải có khả năng thao tác trực tiếp trên các từ ngôn ngữ. Trên cơ sở đó, bài báo đã chứng tỏ rằng phương pháp biểu diễn cấu trúc tập mờ đa thể hình thang của các LFoC được xây dựng dựa trên ĐSGT mở rộng là giải nghĩa được theo định nghĩa của Taski. Bài báo cũng đề xuất một thuật toán tiến hóa trích rút các LRBS giải nghĩa được và có thể mở rộng theo yêu cầu của người quản trị ứng dụng, chẳng hạn, mở rộng khung nhận thức ngôn ngữ LFoC. Các kết quả thực nghiệm đã chứng tỏ rằng, phương pháp thiết kế được đề xuất trong bài báo cho kết quả tốt hơn so với các phương pháp tiếp cận theo lý thuyết tập mờ và phương pháp tiếp cận ĐSGT đã được đề xuất trước đây.

TÀI LIỆU THAM KHẢO/ REFERENCES

[1] C. Mencar and A. M. Fanelli, "Interpretability constraints for fuzzy information granulation," *Information Sciences*, vol. 178, pp. 4585-4618, 2008.

[2] N. C. Ho 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.

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

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

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

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

[7] N. C. Ho, T. T. Son, H. V. Thong, and N. V. Long, "LFoC-Interpretability of Linguistic Rule Based Systems and its Applications To Solve Regression Problems," *International Journal of Computer Technology & Applications*, no. 2, pp. 94-117, 2017.

[8] N. C. Ho, P. T. Lan, N. N. Tu, H. C. Ha, and N. T. Anh, "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.

[9] 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 Transaction on Fuzzy Systems*, vol. 17, no. 5, pp. 1106-1122, 2009.

[10] F. Alcalá et al., "KEEL Data-Mining Software Tool: Data Set Repository, Integration of Algorithms and Experimental Analysis Framework," *Journal of Multiple-Valued Logic and Soft Computing*, vol. 17, no. 2, pp. 255-287, 2011.

[11] R. Alcalá, M. J. Gacto, and F. Herrera, "A fast and scalable multiobjective genetic fuzzy system for linguistic fuzzy modeling in high-dimensional regression problems," *IEEE Transaction Fuzzy Systems*, vol. 19, no. 4, pp. 666-681, 2011.