Journal of Computer Science and Cybernetics, V.30, N.3 (2014), 227–238
DOI:10.15625/1813-9663/30/3/3236

MỘT PHƯƠNG PHÁP SINH HỆ LUẬT MỜ MAMDANI CHO BÀI TOÁN HỒI QUI VỚI NGỮ NGHĨA ĐẠI SỐ GIA TỬ[1]

NGUYỄN CÁT HỒ[1], HOÀNG VĂN THÔNG[2,†], NGUYỄN VĂN LONG[2,‡]

[1] *Viện Công nghệ thông tin, Viện Khoa học và Công nghệ Việt Nam*
ncatho@gmail.com
[2] *Trường Đại học Giao thông Vận tải*
[†]thonghoangvan@yahoo.com; [‡]nvlongdt@yahoo.com.vn

Tóm tắt. Trong bài báo này, chúng tôi đề xuất một thuật toán tiến hóa HA-(2+2)M-PAES sinh các hệ luật mờ Mamdani (MFRBS) đạt được độ thỏa hiệp khác nhau giữa hai mục tiêu độ phức tạp và độ chính xác. Thuật toán được phát triển dựa trên lược đồ tiến hóa (2+2)M-PAES đề xuất trong [6]. Điểm mới của thuật toán là thực hiện học đồng thời cơ sở luật, phân hoạch mờ và hạng từ ngôn ngữ cùng với tập mờ của chúng dựa trên phương pháp luận Đại số gia tử (ĐSGT). Thuật toán cho phép sinh các luật từ mẫu dữ liệu sử dụng thông tin mới nhất của các phân hoạch và các tập mờ trong cùng cá thể. Thêm vào đó, chúng tôi đề xuất một phương pháp mã hóa cá thể mới theo hướng tiếp cận Đại số gia tử để giải quyết bài toán toán này. Thuật toán được thử nghiệm trên sáu bài toán hồi qui mẫu lấy từ [10] được cộng đồng nghiên cứu chấp nhận, kết quả cho thấy thuật toán sinh ra các MFRBS tốt hơn so với thuật toán sử dụng cùng lược đồ tiên hóa trong [8] trên cả hai mục tiêu độ phức tạp và độ chính xác.

Từ khóa. Hệ luật mờ Mamdani, hồi qui, đại số gia tử, tính dễ hiểu.

Abstract. In this paper, we propose an evolutionary algorithm to generate Mamdani Fuzzy Rule-based Systems (MFRBS) with different trade-offs between complexity and accuracy. The algorithm was developed by taking the idea of the schema evolution (2+2)M-PAES proposed in [6]. The main novelty of the algorithm is to learn concurrently rule bases, fuzzy partitions and linguistic terms along with their fuzzy sets by using hedge algebra (HA) based methodology. The algorithm allows to generate generating rules from pattern data utilizing new information of partitions and fuzzy sets in the same individual. In addition, we propose a new method for encoding individuals that can be realized in the hedge algebra approach to solving regression problems. The computer simulation is carried out with six standard regression problems in [10], accepted by the research community and the obtained results show that the MFRBSs generated by the proposed algorithm are better than those examined in [8] with respect to two objectives, the complexity and the accuracy.

Keywords. Mamdani Fuzzy Rule-based system, regression, hedge algebra, interpretability.

[1]This research is funded by Vietnam National Foundation for Science and Technology Development (NAFOSTED) under grant number 102.05-2013.34

1. MỞ ĐẦU

Hệ luật mờ (FRBS: Fuzzy Rule-Based System) đã có những ứng dụng thành công trong nhiều lĩnh vực khác nhau như: điều khiển [9], phân lớp [1, 2, 3] và hồi qui [5, 6, 7, 8]. Nhiều kiểu hệ mờ khác nhau đã được đề xuất, tuy nhiên hệ luật mờ dạng Mamdani (MFRBS) có vai trò trội hơn các dạng khác nhờ MFRBS được định nghĩa bằng các mệnh đề if-then tương tự trong ngôn ngữ tự nhiên [8]. Khi xây dựng FRBS, hai mục tiêu cần đạt được của hệ luật là tính dễ hiểu và độ chính xác. Đây là bài toán tối ưu đa mục tiêu với các mục tiêu xung đột nhau, đòi hỏi phải có giải pháp thỏa hiệp giữa hai mục tiêu này. Với FRBS cho bài toán hồi qui, độ chính xác thường được đo bằng giá trị trung bình phương sai (MSE: Mean Squared Error). Tính dễ hiểu của FRBS rất khó hình thức hóa, vì vậy các nhà nghiên cứu thường tập trung vào một số đặc trưng của khái niệm này và đưa ra các ràng buộc để thỏa mãn những đặc trưng đó. Trong [11] các tác giả đưa ra một số đặc trưng: 1) sự rõ ràng của phân hoạch (số tập mờ, khả năng phân biệt giữa các tập mờ, phân hoạch có phủ toàn bộ vũ trụ); 2) độ phức tạp của hệ luật (số luật, chiều dài của luật).

Yếu tố 1) dễ dàng đạt được nếu sử dụng phân hoạch mờ đều với các tập mờ tam giác biểu thị ngữ nghĩa của các nhãn ngôn ngữ được gán với chúng [3,6]. Tuy nhiên sử dụng phân hoạch đều thường làm giảm độ chính xác của hệ luật. Một số nghiên cứu thực hiện điều chỉnh tham số tập mờ để nâng cao độ chính xác, khi đó làm gia tăng không gian tìm kiếm và có thể làm giảm tính dễ hiểu của hệ luật. Để đạt được yếu tố 2), hệ luật phải có ít luật và độ dài của luật phải ngắn. Điều này dẫn đến các luật phải có tính khái quát cao và vì vậy chúng làm giảm độ chính xác của hệ luật. Để cân bằng giữa độ chính xác và độ phức tạp, một số nghiên cứu phát triển các thuật toán tiến hóa đa mục tiêu thực hiện học đồng thời cơ sở luật, điều chỉnh tập mờ và lựa chọn số tập mờ để phân hoạch các thuộc tính trong quá trình xây dựng FRBS như trong [8].

Trong bài báo này, chúng tôi đề xuất thuật toán HA-(2+2)M-PAES xây dựng MFRBS dựa trên phương pháp luận của ĐSGT và lược đồ tiến hóa (2+2)M-PAES ((2+2)Modify-Pareto Archive Evolution Strategy) đề xuất trong [6] giải bài toán hồi qui đạt được sự cân bằng giữa độ chính xác và các yếu tố 1) và 2). Để thỏa mãn yếu tố 1) chúng tôi sử dụng phân hoạch mờ được xây dựng dựa trên tập từ ngôn ngữ được sinh ra bằng ĐSGT. Thực hiện điều chỉnh tập mờ dựa vào điều chỉnh ngữ nghĩa của các từ ngôn ngữ thông qua điều chỉnh tham số mờ của ĐSGT. Với cách làm này, phân hoạch luôn đảm bảo phủ toàn bộ vũ trụ. Để thỏa yếu tố 2), chúng tôi thực hiện chọn phân hoạch cho từng thuộc tính bằng cách chọn chiều dài tối đa của từ, nhằm đạt được sự cân bằng giữa tính khái quát (generality) và tính riêng (specificity) của hệ luật. Bên cạnh đó, chúng tôi đề xuất phương pháp mã hóa cá thể mới và phương pháp sinh luật từ mẫu dữ liệu sử dụng thông tin mới nhất của các phân hoạch trong các cá thể. Thuật toán được thử nghiệm trên sáu bài toán hồi qui mẫu trong [10]. Kết quả thử nghiệm được đối sánh với các kết quả của các thuật toán được phát triển dựa trên lược đồ tiến hóa (2+2)M-PAES trong [8] là (2+2)M-PAES(C) và (2+2)M-PAES(I). Mặt Pareto đạt được trội hơn, trong khi độ phức tạp của hệ luật tương đương nhưng độ chính xác cao hơn. Các luật có tính khái quát cao hơn do có độ dài ngắn vì vậy làm tăng tính dễ hiểu của hệ luật, đồng thời dễ hiểu hơn với người dùng do sử dụng các từ ngôn ngữ có ngữ nghĩa tự nhiên.

Phần tiếp theo bài báo được tổ chức như sau: trong phần 2 chúng tôi mô tả tóm tắt MFRBS với ngữ nghĩa ĐSGT cho bài toán hồi qui; phần 3 mô tả phương pháp thiết kế phân hoạch; phần 4 mô tả chi tiết phương pháp mã hóa cá thể, các toán tử di truyền và thuật toán

tiến hóa dựa trên ĐSGT; phần 5 trình bầy kết quả thử nghiệm và phân tích đánh giá; phần 6 rút ra một số kết luận.

2. BÀI TOÁN HỒI QUI VÀ HỆ LUẬT MỜ MAMDANI VỚI NGỮ NGHĨA ĐSGT

Bài toán hồi qui: cho tập mẫu dữ liệu $D = \{(x_i, y_i), i = 1, \ldots, N\}$, trong đó $x_i \in U = U_1 \times U_2 \times \ldots \times U_F$ là tích Đề-các của các miền tương ứng của F biến (thuộc tính) độc lập $X_1, ..., X_F$, $y_i \in U_{F+1}$ là biến phụ thuộc, N là số mẫu dữ liệu và thông thường U_i với $i = 1, .., F + 1$ là tập số thực. Từ tập mẫu D xây dựng một mô hình cho phép dự đoán giá trị y ứng với giá trị x.

Giải bài toán hồi qui bằng hệ luật mờ dạng Mamdani với ngữ nghĩa ĐSGT là đi xây dựng hệ luật mờ Mamdani từ tập mẫu dữ liệu D. Với các luật mờ có dạng như sau:

$$\mathrm{R_m}\text{: If } \mathrm{X_1} \text{ is } \mathrm{A_{1,j_m}} \text{ and ... and } \mathrm{X_F} \text{ is } \mathrm{A_{F,j_m}} \text{ then Y is } \mathrm{A_{F+1,}}_{j_m} \tag{1}$$

trong đó:

- $\mathrm{A}_{f,j_m} \in \{\{\mathrm{A}_{f,0} \cup X_{(k_f)} = \{\mathrm{A}_{f,0}, \mathrm{A}_{f,1}, \ldots, \mathrm{A}_{f,|X_{(k_f)}|}\}\}, f = 1, \ldots, F$ là tập các hạng từ có độ dài không quá k_f được sinh ra bằng ĐSGT dùng để phân hoạch thuộc tính thứ $f, A_{f,0}$ kí hiệu giá trị *Don'tcare* với giá trị hàm thuộc đồng nhất bằng 1.

- $\mathrm{A_{F+1,j_m}} \in \mathrm{X_{(k_{F+1})}} = \{\mathrm{A_{F+1,1}}, .., \mathrm{A_{F+1,|X_{(k_{F+1})}|}}\}$, $\mathrm{X_{(k_{F+1})}}$ là tập các hạng từ có độ dài không quá k_{F+1} của ĐSGT dùng để phân hoạch biến phụ thuộc Y.

- $m = 1, \ldots, M$ với M là số luật.

Như đã trình bầy trong phần 1, mục tiêu xây dựng MFRBS cho bài toán hồi qui là hệ luật phải dễ hiểu và có độ chính xác cao. Độ phức tạp (complexity) của hệ luật được xem là yếu tố quan trọng thể hiện tính dễ hiểu và được xác định bằng tổng độ dài của các luật trong hệ luật. Độ chính xác của hệ luật được đo bằng giá trị trung bình phương sai theo công thức:

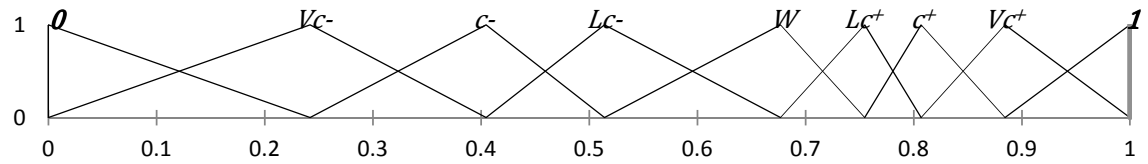
$$MSE = \frac{1}{2N} \sum_{i=1}^{N} (\hat{y}_i - y_i)^2 \tag{2}$$

trong đó $\hat{y}_i$ là giá trị suy diễn từ hệ luật của điểm dữ liệu (x_i, y_i) theo phương pháp trung bình trọng số, được tính như sau: $\hat{y}_i = \frac{\sum_{m=1}^{M} \mu_m(x_i) \overline{A}_{F+1,j_m}}{\sum_{m=1}^{M} \mu_m(x_i)}$ $\quad i = 1..N$, với $\mu_m(x_i) = \prod_{f=1}^{F} \mu_{A_{f,j_m}}(x_{if})$ 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 từ ngôn ngữ A_{F+1,j_m} và $\mu_{\mathrm{A_{f,j_m}}}(.)$ là hàm thuộc của từ ngôn ngữ A_{f,j_m}.

Lưu ý: nếu $\sum_{m=1}^{M} \mu_m(x_i) = 0$, có nghĩa là mẫu dữ liệu x_i không đốt cháy luật nào thì $\hat{y}_i$ sẽ được xác định theo phương pháp đề xuất trong [5].

3. THIẾT KẾ PHÂN HOẠCH MỜ

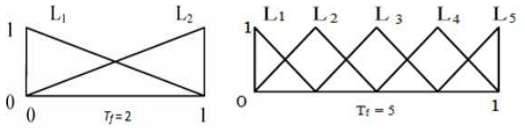
Trong nghiên cứu này chúng tôi sử dụng các từ ngôn ngữ được sinh ra bằng ĐSGT để xây dựng các phân hoạch, ngữ nghĩa của từ là tập mờ dạng tam giác (xem hình 1) được định nghĩa bằng bộ ba giá trị định lượng $(\nu(A_{f,j-1}), \nu(A_{f,j}), \nu(A_{f,j+1}))$, trong đó $A_{f,j-1}$ và $A_{f,j+1}$ lần lượt là từ bên trái và bên phải của từ $A_{f,j}$ trong $X_{(k_f)}$. Để điều chỉnh ngữ nghĩa của từ ngôn ngữ ta chỉ cần điều chỉnh bộ tham số μL, μc^-, số lượng tham số không phụ thuộc vào số lượng tập mờ được sử dụng trong phân hoạch. Như vậy, theo tiếp cận ĐSGT không gian tìm kiếm cho việc điều chỉnh phân hoạch của bài toán có F chiều là $2 * (F + 1)$ chiều. Tiếp cận theo tập mờ trong [8], việc điều chỉnh phân hoạch thông qua điều chỉnh đỉnh các tam giác, như vậy số lượng tham số phụ thuộc vào số từ ngôn ngữ sử dụng. Giả sử số từ ngôn ngữ sử dụng cho mỗi phân hoạch là T_{max} (với $5 \leq T_{max} \leq 9$) thì không gian tìm kiếm là $(T_{max} - 2) * (F + 1)$ chiều. Như vậy, theo tiếp cận ĐSGT thì không gian tìm kiếm giảm đi do $T_{max} - 2 > 2$.



Hình 1: Một thiết kế phân hoạch tập mờ dạng tam giác với tham số $k = 2$, $\mu L = 0.4020657$, $\mu c^- = 0.6768686$

4. THUẬT TOÁN TIẾN HÓA DỰA TRÊN ĐSGT

Khi thiết kế các thuật toán tiến hóa, mã hóa cá thể là công việc quan trọng. Dựa trên cấu trúc mã hóa chúng ta thiết kế các toán tử lai ghép, đột biến nhằm tìm kiếm lời giải tốt hơn sau mỗi thế hệ. Trong [8] phát triển thuật toán (2+2)M-PAES(I) và (2+2)M-PAES(C) dựa trện lược đồ tiến hóa (2+2)M-PAES đề xuất trong [6]. Để thực hiện học đồng thời cơ sở luật, phân hoạch mờ và điều chỉnh ngữ nghĩa của nhãn ngôn ngữ, các tác giả thực hiện mã hóa cá thể gồm 3 phần: cơ sở luật, phân hoạch mờ, hàm tuyến tính từng khúc. Mỗi luật được mã hóa bằng 1 véc tơ $F + 1$ chiều với các phần tử là chỉ số của nhãn ngôn ngữ trong phân hoạch. Cơ sở luật được mã hóa không phải là cơ sở luật thực sự cần xây dựng mà chỉ là cơ sở luật được xây dựng trên các phân hoạch có số tập mờ đồng nhất bằng T_{max}. Cơ sở luật này được gọi là *cơ sở luật ảo* và các phân hoạch như vậy được gọi là *phân hoạch ảo*. Các tác giả trong [8] phải làm như vậy nhằm duy trì được ngữ nghĩa của các nhãn ngôn ngữ trong cơ sở luật của cá thể cha mẹ ở trong các cá thể. Nếu mã hóa cơ sở luật thực thay vì cơ sở luật ảo thì sau khi thực hiện lai ghép, đột biến nó có thể làm mất đi ngữ nghĩa của nhãn ngôn ngữ trong cá thể con. Ví dụ: giả sử một cá thể cha mẹ có véc tơ luật $R = (1, 2, 2, 5)$ và thuộc tính thứ 3 được phân hoạch bằng 2 tập mờ (L_1, L_2 – Hình 2), như vậy tiền điều kiện thứ 3 của R



Hình 2: Phân hoạch với 2 tập mờ và 5 tập mờ

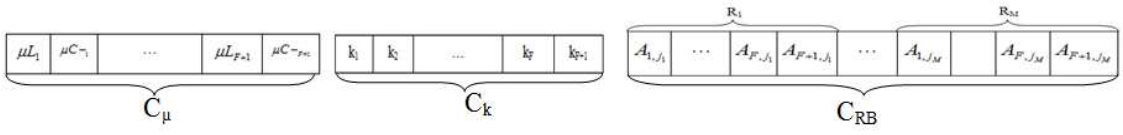
là nhãn ngôn ngữ L_2, ở đây L_2 nằm ở tận cùng phía phải của phân hoạch. Sau khi lai ghép, cá thể con có véc tơ luật $R = (1, 2, 2, 5)$ và thuộc tính thứ 3 được phân hoạch bằng 5 tập mờ $(L_1, L_2, L_3, L_4, L_5$ – Hình 2), như vậy tiền điều kiện thứ 3 của R vẫn là nhãn ngôn ngữ L_2 nhưng lúc này L_2 lại nằm gần sát phía trái của phân hoạch (*tức là ngữ nghĩa của nhãn ngôn ngữ thay đổi hoàn toàn*).

Với cách mã hóa dựa trên cơ sở luật ảo, để tính toán giá trị hàm mục tiêu, các tác giả phải thực hiện chuyển đổi cơ sở luật ảo thành cơ sở luật thực. *Quá trình này cũng làm mất mát ngữ nghĩa của nhãn ngôn ngữ và làm tăng thời gian tính toán.*

Chúng tôi tiến hành mã hóa các thể gồm 3 phần: các tham số mờ gia tử, chiều dài tối đa của hạng từ, cơ sở luật. Mỗi luật mã hóa bằng một véc tơ, mỗi phần tử là một từ ngôn ngữ được sinh ra bằng ĐSGT hoặc giá trị *Don'tcare*. Với phương pháp mã hóa này sau quá trình lai ghép, đột biến, nếu phân hoạch của thuộc tính bị thay đổi thì không làm mất đi ngữ nghĩa cốt lõi của từ sử dụng trong hệ luật. Thật vậy, giả sử thuộc tính thứ f trước khi lai ghép, đột biến được phân hoạch bằng tập từ có độ dài không quá k_f. Sau khi lai ghép, đột biến được phân hoạch bằng tập từ có độ dài không quá k'_f. Nếu $k'_f > k_f$ thì $X_{(k_f)} \subset X_{(k'_f)}$ vì vậy ngữ nghĩa của từ trong các luật của cá thể con ít thay đổi. Nếu $k'_f < k_f$ thì $X_{(k_f)} \subset X_{(k'_f)}$, khi đó ta chỉ phải biến đổi những từ có độ dài k_f có trong các luật thành từ có độ dài k'_f bằng cách cắt bỏ những gia tử bên trái của từ để thu được từ có độ dài bằng k'_f. Do tính kế thừa ngữ nghĩa của từ được sinh ra từ gia tử, từ mới thu được sau khi biến đổi vẫn giữ được ngữ nghĩa lõi của từ gốc. Ví dụ: nếu $k_f = 3$, $k'_f = 2$, từ "*Little Very True*" sẽ được biến đổi thành "*Very True*". *Với phương pháp mã hóa này, quá trình tính giá trị hàm mục tiêu không phải chuyển đổi cơ sở luật, vì vậy làm giảm thời gian tính toán so với phương pháp đề xuất trong [8].*

4.1. Mã hóa cá thể dựa trên ĐSGT

Mỗi cá thể p của quần thể được mã hóa gồm ba phần $C\mu$, C_k, C_{RB} (xem Hình 3), trong đó: $C\mu$ là dãy số mã hóa các tham số mờ của ĐSGT bao gồm $F + 1$ véc tơ, mỗi véc tơ gồm 2 phần tử thực mã hóa tham số mờ của ĐSGT μL_f và μC_f^- (ở đây chúng tôi sử dụng ĐSGT có 2 gia tử). C_k là một véc tơ $F + 1$ chiều, phần tử thứ f là một số tự nhiên kf xác định độ dài tối đa các hạng từ sử dụng để phân hoạch thuộc tính thứ f. C_{RB} mã hóa cơ sở luật gồm M_p luật (M_p có thể khác nhau giữa các cá thể), với mỗi luật là một véc tơ có $F + 1$ phần tử, mỗi phần tử gồm một từ ngôn ngữ và tập mờ tương ứng trong $X_{(k_f)}$. Như vậy C_{RB} gồm $M_p * (F + 1)$ phần tử.



Hình 3: Cấu trúc mã hóa một cá thể

Chúng ta giới hạn số luật trong mỗi cơ sở luật nằm trong khoảng $[M_{min}, M_{max}]$ nhằm đảm bảo hệ luật sinh ra đạt được sự cân bằng giữa tính dễ hiểu và độ chính xác đồng thời giới hạn không gian tìm kiếm các hệ luật. Hàm mục tiêu của mỗi cá thể gồm hai thành phần $(MSE, Comp)$, trong đó MSE được xác định theo (2) và $Comp$ là tổng độ dài của các luật

trong cơ sở luật.

4.2. Các toán tử di truyền

4.2.1. Toán tử lai ghép

Với hai cá thể bố mẹ p_1, p_2 sử dụng phương pháp lai ghép một điểm (one-point crossover) độc lập trên $C\mu$, C_k và C_{RB}. Thực hiện lai ghép trên C_μ với xác suất $Pc\mu$, điểm lai ghép được chọn ngẫu nhiên trong đoạn $[1, (F+1)*2-1]$. C_k được lai ghép với xác suất Pck, điểm lại ghép được chọn ngẫu nhiên trong đoạn $[1, F+1]$. C_{RB} được lai ghép với xác suất $PcRB$, điểm lai ghép được chọn ngẫu nhiên trong đoạn $[1, \rho_{min}-1]$, trong đó ρ_{min} là số luật ít nhất của 2 cơ sở luật trong p_1 và p_2. Chú ý nếu trên C_{RB} toán tử lai ghép không được thực hiện thì toán tử đột biến trên C_{RB} luôn xảy ra. Sau khi lai ghép có thể tạo ra các luật trùng nhau, thực hiện loại bỏ các luật trùng nhau chỉ giữ lại một luật.

4.2.2. Toán tử đột biến

Với cá thể con s thực hiện đột biến độc lập trên C_μ, C_k và C_{RB}. Thực hiện đột biến trên C_μ với xác suất $Pm\mu$, chọn ngẫu nhiên một gen thứ $f \in [1, (F+1)*2-1]$, thay đổi giá trị tại f của C_μ bằng một giá trị ngẫu nhiên trong đoạn $[min, max] \subset [0,1]$ do người dùng xác định trước. Trên C_k thực hiện đột biến với xác suất Pmk, chọn ngẫu nhiên một gen thứ $f \in [1, (F+1)]$, thay đổi giá trị của gen này bằng cách ngẫu nhiên tăng hoặc giảm một đơn vị. Nếu giá trị mới nằm ngoài đoạn $[1, kmax]$ thì đột biến sẽ bị bỏ qua. Trên C_{RB} thực hiện đột biến với xác suất $PmRB$, nếu đột biến xảy ra thì chọn thực hiện một trong hai toán tử đột biến dưới đây. Toán tử đột biến thứ nhất xảy ra với xác suất $PmAdd$, nếu không toán tử đột biến thứ hai được áp dụng.

- *Toán tử đột biến thứ nhất:* thêm γ luật vào C_{RB} (các luật được sinh ra theo thuật toán trong mục 3.2.3), với γ được chọn ngẫu nhiên trong đoạn $[1, \gamma_{max}]$, nếu $\gamma + M > M_{max}$ thì $\gamma = M_{max} - M$.

- *Toán tử đột biến thứ hai:* thay đổi ngẫu nhiên δ giá trị của C_{RB}, với δ được chọn ngẫu nhiên trong đoạn $[1, \delta_{max}]$. Chọn ngẫu nhiên một luật R, tiếp theo chọn ngẫu nhiên một phần tử, giả sử là f. Nếu $f \leq F$ thì chọn ngẫu nhiên một từ ngôn ngữ w trong $\{Don't care\} \cup X_{(k_f)}$, nếu $f = F+1$ thì chọn ngẫu nhiên một từ ngôn ngữ w trong $X_{(k_{F+1})}$, đặt phần tử thứ f của luật R bằng w. Nếu luật sau khi đột biến có độ dài lớn hơn ℓ_{max} (chiều dài tối đa của luật do người dùng chọn trước) thì đột biến sẽ bị bỏ qua.

Sau khi đột biến hệ luật có thể có các luật có độ dài bằng 0 và các luật trùng nhau. Thực hiện loại bỏ các luật này, với luật trùng nhau thì giữ lại một luật.

4.2.3. Thuật toán sinh luật ngẫu nhiên từ dữ liệu

Chọn ngẫu nhiên một mẫu dữ liệu (x_i, y_i) trong tập dữ liệu huấn luyện. Thực hiện sinh một luật R có độ dài ℓ dựa trên các hệ khoảng mờ tương tự $\mathcal{S}(k_f)$ được xác định từ bộ tham số mờ của ĐSGT trong $C\mu$ và C_k trong cùng cá thể như sau:

Bước 1: Xác định các tập từ ngôn ngữ $\mathrm{X}_{(kf)}$ và các hệ khoảng mờ tương tự $\mathcal{S}_{(k_f)}$ của các thuộc tính tương ứng với các bộ tham số mờ của DSGT trong $C\mu$ và chiều dài tối đa của từ trong C_k.
Bước 2: Sinh luật R_i có chiều dài F, với mẫu dữ liệu $x_i = (x_{i1}, x_{i2}, \ldots, x_{iF}, y_i) \in D$, thực hiện:

- Xác định tập các điều kiện tiền đề A_i của luật được sinh bởi mẫu dữ liệu (x_i, y_i)

$$A_i = \{(A_{1,j_i}, A_{2,j_i}, \ldots, A_{F,j_i}) | A_{f,j_i} \in \mathrm{X}_{(k_f)}, x_{if} \in \mathcal{T}(A_{f,j_i}),\ f = 1, 2, \ldots, F, j_i \in [1, |\mathrm{X}_{(k_f)}|]\}$$

- Sinh luật gồm vế trái là A_i và kết luận là A_{F+1,j_i} tương ứng với mẫu dữ liệu (x_i, y_i)

$$R_i = (A_i \in A_{F+1,j_i}) \text{ với } A_{F+1,j_i} \in \mathrm{X}_{(k_{F+1})}, y_i \in \mathcal{T}(A_{F+1,j_i}),\ j_i \in [1, |\mathrm{X}_{(k_{F+1})}|]$$

  Trong đó $\mathcal{T}(A_{f,j_i})$ là khoảng mờ tương tự của hạng từ A_{f,j_i} trong $\mathcal{S}_{(k_f)}$.

Bước 3: Sinh luật R có chiều dài ℓ từ luật R_i. Chọn ngẫu nhiên $\ell \in [1, \ell_{max}]$, thực hiện sinh ngẫu nhiên ℓ giá trị nguyên $f_1, \ldots, f_l$ khác nhau trong đoạn $[1, F]$, thay đổi các điều kiện tiền đề có thứ tự khác với $f_1, \ldots, f_l$ của luật R_i thành *Don'tcare* khi đó chúng ta được luật R có chiều dài ℓ.

4.2.4. Thuật toán tiến hóa HA-(2+2)M-PAES

Trong phần này trình bầy những bước cơ bản của thuật toán HA-(2+2)M-PAES được phát triển dựa trên lược đồ tiên hóa (2+2)M-PAES đề xuất trong [6]. Mục tiêu của thuật toán là xây dựng một mặt xấp xỉ tối ưu Pareto với mỗi điểm là một cá thể với 2 mục tiêu MSE và $Comp$. Kí hiệu P_A là mặt xấp xỉ tối ưu Pareto của thế hệ hiện tại. Thuật toán gồm các bước sau:
Bước 1: Sinh ngẫu nhiên 2 cá thể o_1, o_2. Các gen trong $C\mu$ là các số thực được sinh ngẫu nhiên trong đoạn $[min, max]$. Các gen trong C_k là các số tự nhiên được sinh ngẫu nhiên trong đoạn $[1, kmax]$. Các gen trong C_{RB} gồm M luật được sinh theo thuật toán trong mục 3.2.3 với M được chọn ngẫu nhiên trong đoạn $[M_{min}, M_{max}]$.
Bước 2: Bổ sung o_1, o_2 vào P_A
Bước 3: Lặp $i = 1, \ldots, MaxGen$ (số thế hệ tối đa)
Bước 3.1. Chọn ngẫu nhiên hai cá thể cha mẹ p_1, p_2 trong P_A (p_1, p_2 có thể trùng nhau)
Bước 3.2. Thực hiện lai ghép hai cá thể cha mẹ p_1, p_2 để sinh ra hai cá thể con s_1, s_2
Bước 3.3. Thực hiện đột biến lần lượt trên s_1, s_2
Bước 3.4. Tính giá trị mục tiêu (MSE, $Comp$) của s_1, s_2
Bước 3.5. Lần lượt thực hiện bổ sung s_1, s_2 vào P_A nếu có thể
Bước 3.6. Lặp lại bước 3.1 cho đến khi $i > MaxGen$
Một cá thể con s nếu không bị trội bởi bất kỳ phần tử nào có trong P_A thì s được bổ sung vào P_A đồng thời loại bỏ tất cả các phần tử trong P_A bị trội bởi s. Sau đó kiểm tra nếu số cá thể trong P_A lớn hơn số lượng tối đa (archiveSize) được phép lưu trữ trong P_A thì loại bỏ ngẫu nhiên một cá thể trong vùng có mật độ cao nhất ra khỏi P_A. Xác định vùng có mật độ cao nhất theo thuật toán trong [4].

5. NGHIÊN CỨU MÔ PHỎNG MÁY TÍNH

Chúng tôi tiến hành thử nghiệm thuật toán HA-(2+2)M-PAES trên 6 bài toán hồi qui cho trong Bảng 1, #Pat là số mẫu, #Att là số thuộc tính. Phương pháp thử nghiệm Five-Fold, mỗi dataset được chia thành 5 phần (đã được chia và lưu trữ trong [10]), thử nghiệm mỗi Fold 6 lần ($6 \times 5 = 30$ lần thử). Mỗi lần thử nghiệm tạo ra một mặt Pareto xấp xỉ tối ưu, sắp xếp các cá thể trong mặt Pareto theo thứ tự tăng dần của giá trị MSE trên tập huấn luyện. Trên mỗi mặt Pareto đã được sắp chỉ giữ lại tối đa 20 hoặc bằng với số lượng cá thể của mặt Pareto có số cá thể ít nhất trong 30 mặt Pareto. Thực hiện tính trung bình giá trị MSE, Comp trên 30 mặt Pareto. Như vậy với mỗi bài toán ta thu được một mặt Pareto xấp xỉ tối ưu trung bình theo giá trị MSE và Comp. Các mặt Pareto xấp xỉ tối ưu trung bình của 6 bài toán được thể hiện trong hình 4. Bên cạnh thể hiện trực quan mặt xấp xỉ tối ưu Pareto, chúng tôi thực hiện phân tích thống kê theo phương pháp t-test với độ tin cậy 95% trên giá trị $\overline{MSE}_{Tr}$ và $\overline{MSE}_{Ts}$ để xác định có sự khác biệt giữa thuật toán của chúng tôi và các thuật toán trong [8] hay không?.

Bài toán	#Pat	#Att
Electrical Maintainance (ELE)	1,056	4
(*Dự báo bảo trì lưới điện*)		
Weather Ankara (WA)	1,609	9
(*Dự báo nhiệt độ trung bình ở Ankara*)		
Weather Izmir (WI)	1,461	9
(*Dự báo nhiệt độ trung bình ở Izmir*)		
Auto-MPG (MPG6)	398	5
(*Dự báo tiêu thụ nhiên liệu của ô tô*)		
Stock (STP-Dự báo giá cổ phiếu)	950	9
Treasury (TR)	1,049	15
(*Dự báo lãi xuất huy động của ngân hàng*)		

Bảng 1: Các bài toán sử dụng thử nghiệm

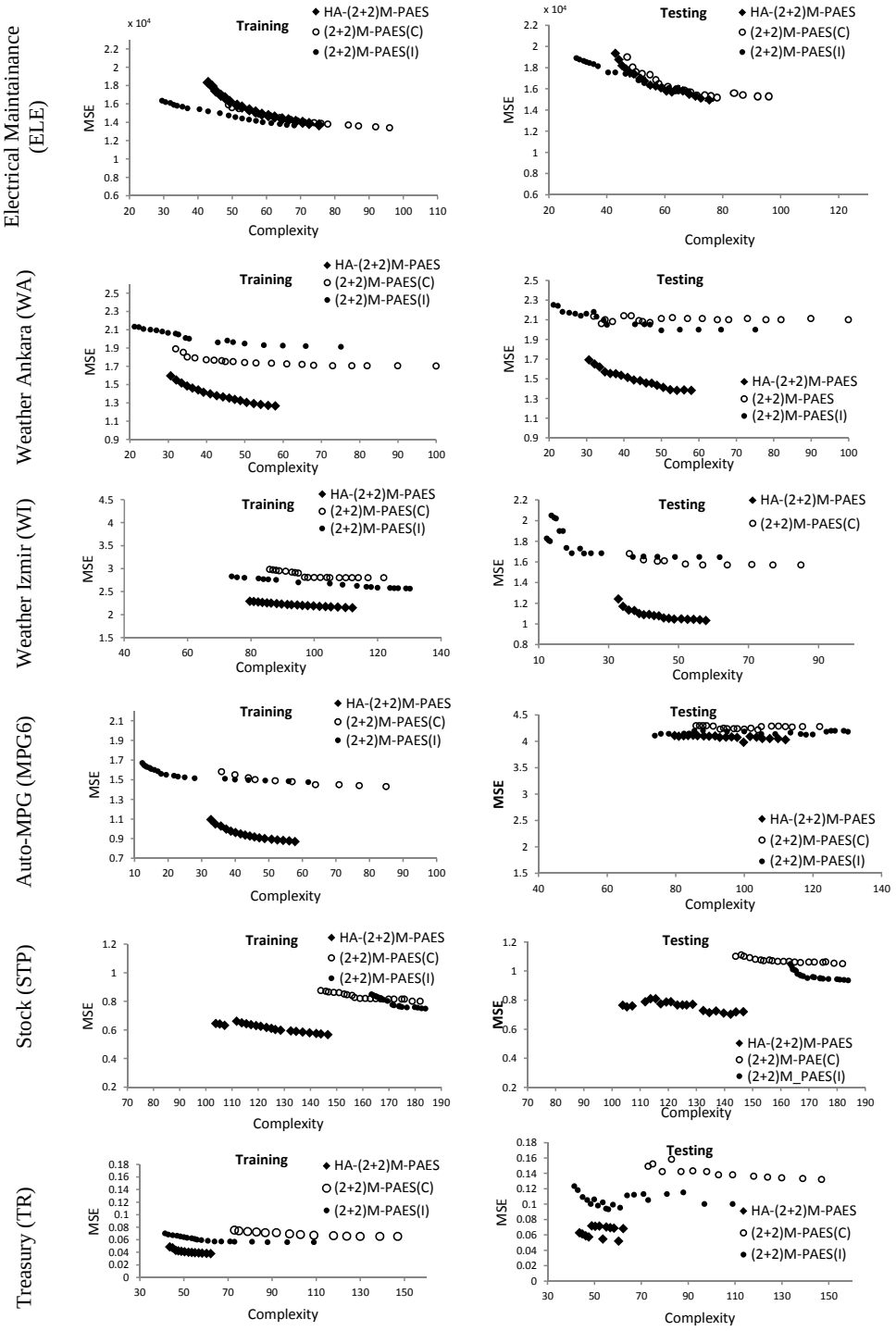
min	0.3	$Pc\mu$	0.75
max	0.7	Pck	0.3
$kmax$	3	$PcRB$	0.3
$archiveSize$	64	$Pm\mu$	0.3
$MaxGen$	3,000,000	Pmk	0.3
M_{min}	5	$PmRB$	0.1
M_{max}	50	$PmAdd$	0.75
γ_{max}	5	ℓ_{max}	5
δmax	5		

Bảng 2: Các tham số thử nghiệm

Chúng tôi áp dụng kiểm tra trên 2 điểm có độ chính xác và độ phức tạp trái ngược nhau của mặt Pareto: điểm ứng với lời giải có $\overline{MSE}_{Tr}$ nhỏ nhất, $\overline{Comp}$ lớn nhất (kí hiệu là: FIRST) và $\overline{MSE}_{Tr}$ lớn nhất, $\overline{Comp}$ nhỏ nhất (ký hiệu là: LAST). Các kết quả thống kê được trình bầy trong các Bảng 3,4. Các ký hiệu được sử dụng tương tự như trong [8], "*" thể hiện 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. Ký hiệu $\overline{MSE}_{Tr}$, $\overline{MSE}_{Ts}$, 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 và tập dữ liệu kiểm tra và $\overline{Comp}$, $\overline{\#R}$ lần lượt là trung bình độ phức tạp và trung bình số luật của hệ luật.

Quan sát Hình 4 chúng ta thấy rằng, thuật toán HA-(2+2)M-PAES tạo ra mặt xấp xỉ tối ưu Pareto trội hơn trên 5 bài toán, trừ bài toán ELE. Hệ luật được sinh ra có độ phức tạp nhỏ hơn nhiều trên các bài toán STP, TR, MPG6. Mặt Pareto trên tập kiểm tra không khác nhiều

so với mặt Pareto trên tập huấn luyện trên 5 bài toán trừ bài toán MPG6. Từ đây chúng ta có thể kết luận rằng thuật toán tạo ra MFRBS có tính ổn định trong quá trình suy diễn.



Hình 4: Mặt xấp xỉ tối ưu Pareto trung bình theo độ chính xác MSE và độ phức tạp $Comp$

Bài toán	Thuật toán	$\overline{\#R}$	$\overline{Comp}$	$\overline{MSE}_{Tr}$	σ_{tr}	t_{tr}	$\overline{MSE}_{Ts}$	σ_{ts}	t_{ts}
ELE	(2+2)M-PAES(I)	34.480	**68.210**	13660.200	**1851.500**	=	15768.600	**3239.900**	=
	(2+2)M-PAES(C)	**24.240**	96.480	**13539.800**	3764.700	*	15278.800	4129.000	=
	HA-(2+2)M-PAES	34.966	75.414	13732.337	2499.690	=	**14969.681**	4010.176	*
WA	(2+2)M-PAES(I)	20.200	75.160	1.911	0.381	+	1.997	0.298	+
	(2+2)M-PAES(C)	**15.270**	98.650	1.694	0.489	+	2.094	0.973	+
	HA-(2+2)M-PAES	24.100	**58.000**	**1.265**	**0.175**	*	**1.383**	**0.229**	*
WI	(2+2)M-PAES(I)	17.830	61.810	1.474	0.343	+	1.647	0.343	+
	(2+2)M-PAES(C)	**13.120**	83.550	1.441	0.276	+	1.556	0.243	+
	HA-(2+2)M-PAES	24.167	**57.833**	**0.873**	**0.102**	*	**1.034**	**0.161**	*
MPG6	(2+2)M-PAES(I)	**40.360**	130.280	2.565	0.341	+	4.185	1.352	=
	(2+2)M-PAES(C)	48.030	121.660	2.820	0.428	+	4.304	1.365	=
	HA-(2+2)M-PAES	47.700	**112.033**	**2.153**	**0.192**	*	**4.036**	**1.117**	*
STP	(2+2)M-PAES(I)	**48.530**	184.000	0.748	**0.098**	+	0.934	**0.175**	=
	(2+2)M-PAES(C)	49.420	181.730	0.795	0.225	+	1.046	0.309	+
	HA-(2+2)M-PAES	49.100	**146.700**	**0.567**	0.109	*	**0.720**	0.192	*
TR	(2+2)M-PAES(I)	25.100	103.920	0.056	0.020	=	0.100	0.097	=
	(2+2)M-PAES(C)	**19.100**	147.000	0.066	0.025	=	0.132	0.132	=
	HA-(2+2)M-PAES	29.267	**62.267**	**0.038**	**0.014**	*	**0.068**	**0.094**	*

Bảng 3: So sánh kết quả thử nghiệm thuật toán HA-(2+2)M-PAES với các thuật toán trong [8] tại điểm FIRST

Bảng 3 so sánh kết quả giữa lời giải tốt nhất của thuật toán HA-(2+2)M-PAES và hai thuật toán (2+2)M-PAES(I) và (2+2)M-PAES(C) trong [8]. Từ bảng này cho thấy giá trị MSE của thuật toán HA-(2+2)M-PAES tốt hơn trên 5 bài toán trên cả tập huấn luyện và tập kiểm tra, ngoại trừ bài toán ELE thấp hơn trên tập huấn luyện. Tuy nhiên kết quả phân tích thống kê cho thấy thuật toán HA-(2+2)M-PAES và thuật toán tốt nhất không cho thấy sự khác biệt. Ở đây có sự khác biệt lớn về độ chính xác giữa thuật toán của chúng tôi với các thuật toán được so sánh. Như bài toán WA giá trị MSE của thuật toán của chúng tôi là 1.265 so với (1.911 và 1.694) trên tập dữ liệu huấn luyện và 1.383 so với (1.997 và 2.094) trên tập kiểm tra. Trên các bài toán WA, WI, STP và TR thuật toán của chúng tôi tạo ra các hệ luật có độ phức tạp thấp nhưng độ chính xác cao hơn nhiều như bài toán STR giá trị MSE là 0.720 so với (0.934, 1.046) trên tập kiểm tra hoặc bài toán TR giá trị MSE là 0.068 so với (0.100, 0.132). Các hệ luật được tạo ra có chiều dài luật ngắn hơn so với các thuật toán trong [8], với bài toán WA phương pháp của chúng tôi trung bình chiều dài luật là 2.4 so với (3.72 và 6.46) hoặc bài toán WI chiều dài trung bình luật là 2.393 so với (3.467 và 6.368),.... Từ đây chúng tôi thấy rằng thuật toán của chúng tôi sinh ra các luật có tính khái quát cao hơn, do đó dễ hiểu hơn với người dùng.

Từ Bảng 3 cho thấy kết quả phân tích thống kê t-test với độ tin cậy 95% thì thuật toán chúng tôi đề xuất khác biệt so với các thuật toán trong [8] trên 3 bài toán (WA, WI, STR) trên tập dữ liệu kiểm tra, các bài toán còn lại không có sự khác biệt nào. Phân tích các kết quả trong Bảng 4 cũng cho kết quả tương tự như trong Bảng 3.

Bài toán	Thuật toán	$\overline{\#R}$	$\overline{Comp}$	$\overline{MSE}_{Tr}$	σ_{tr}	t_{tr}	$\overline{MSE}_{Ts}$	σ_{ts}	t_{ts}
ELE	(2+2)M-PAES(I)	**13.560**	**29.560**	**16358.500**	**2713.600**	*	**18896.000**	**3672.500**	*
	(2+2)M-PAES(C)	20.000	45.100	16595.800	5556.400	=	18977.300	5816.400	=
	HA-(2+2)M-PAES	22.448	41.655	19296.504	5920.099	+	21042.461	9578.943	+
WA	(2+2)M-PAES(I)	**7.380**	**19.110**	2.142	0.449	+	2.244	0.529	+
	(2+2)M-PAES(C)	10.700	32.250	1.877	0.733	=	2.119	0.937	=
	HA-(2+2)M-PAES	15.300	28.167	**1.686**	**0.359**	*	**1.795**	**0.401**	*
WI	(2+2)M-PAES(I)	**6.370**	**12.370**	1.670	0.539	+	1.827	0.566	+
	(2+2)M-PAES(C)	10.380	36.270	1.577	0.377	+	1.678	0.325	+
	HA-(2+2)M-PAES	14.767	28.033	**1.080**	**0.330**	*	**1.176**	**0.342**	*
MPG6	(2+2)M-PAES(I)	32.600	**73.960**	2.829	0.350	+	**4.109**	1.321	=
	(2+2)M-PAES(C)	**31.160**	84.900	2.985	0.457	+	4.327	1.410	=
	HA-(2+2)M-PAES	37.267	79.700	**2.295**	**0.204**	*	4.114	**1.065**	*
STP	(2+2)M-PAES(I)	47.530	163.420	0.849	**0.164**	=	0.958	**0.183**	=
	(2+2)M-PAES(C)	44.460	144.400	0.881	0.225	=	1.102	0.323	+
	HA-(2+2)M-PAES	**37.900**	**103.833**	**0.645**	0.195	*	**0.763**	0.201	*
TR	(2+2)M-PAES(I)	**10.850**	**41.500**	0.070	**0.025**	=	0.123	0.125	=
	(2+2)M-PAES(C)	15.650	73.030	0.076	0.027	=	0.148	0.135	=
	HA-(2+2)M-PAES	23.500	43.600	**0.048**	0.031	*	**0.063**	**0.034**	*

Bảng 4: So sánh kết quả thử nghiệm thuật toán HA-(2+2)M-PAES với thuật toán trong [8] tại điểm **LAST**

6. KẾT LUẬN

Bài báo đề xuất một thuật toán tiến hóa HA-(2+2)M-PAES xây dựng hệ luật mờ Mamdani dựa trên ngữ nghĩa và phương pháp luận của ĐSGT và lược đồ tiến hóa (2+2)M-PAES trong [6] giải bài toán hồi qui. Thuật toán cho phép học đồng thời cơ sở luật, phân hoạch và điều chỉnh các tập mờ với số tham số cố định $2 * (F + 1)$ thay vì biến thiên theo số tập mờ được sử dụng trong phân hoạch như trong [8]. Chúng tôi đề xuất phương pháp cá thể dựa trên tính chất của tập từ ngôn ngữ sinh ra bằng ĐSGT. Phương pháp mã hóa đề xuất làm giảm thời gian tính giá trị hàm mục tiêu và không gian tìm kiếm so với các thuật toán trong [8]. Thêm vào đó thuật toán phát triển phương pháp sinh luật từ mẫu dữ liệu dựa trên những thông tin mới nhất của chính cá thể thay vì sinh luật ngẫu nhiên như trong [8]. Với thuật toán sinh luật như vậy sẽ tạo ra các luật có tiềm năng đốt cháy dữ liệu cao hơn.

Kết quả thử nghiệm thuật toán HA-(2+2)M-PAES trên sáu bài toán hồi qui cho thấy mặt Pareto tạo ra trội hơn so với mặt Pareto tạo ra trong [8] trừ bài toán ELE. Trên các điểm được xem xét của mặt Pareto (FIRST, LAST) độ chính xác của hệ luật được sinh ra từ thuật toán HA-(2+2)M-PAES tốt hơn trên cả tập huấn luyện và tập kiểm tra. Thuật toán tạo ra các luật có tính khái quát cao hơn làm tăng tính dễ hiểu của hệ luật. Từ kết quả thử nghiệm, chúng tôi có thể kết luật rằng ĐSGT giúp cho việc tối ưu tốt hơn nhờ giảm bớt không gian tìm. Hệ luật được tạo ra dễ hiểu hơn với người dùng do sử dụng các từ ngôn ngữ có ngữ nghĩa thay vì các nhãn.

TÀI LIỆU THAM KHẢO

[1] Nguyễn Cát Hồ, Hoàng Văn Thông, Nguyễn Văn Long, "Một phương pháp tiến hóa sinh hệ luật mờ cho bài toán phân lớp với ngữ nghĩa thứ tự ngôn ngữ", *Tạp chí Tin học và*

Điều khiển học, T.28, S.4, tr. 333-345, 2012.

[2] Cat Ho Nguyen, Witold Pedrycz, Thang Long Duong, ThaiSon Tran, "A genetic design of linguistic terms for fuzzy rule based classifiers", *International Journal of Approximate Reasoning*, vol. 54, pp. 1–21, 2012.

[3] Hisao Ishibuchi and Takashi Yamamoto, "Fuzzy rule selection by multi-objective genetic local search algorithms and rule evaluation measures in data mining", *Fuzzy Sets and Systems* vol. 141, pp. 59–88, 2004.

[4] Joshua D. Knowles and David W.corne, "Approximating the Nondominated Front Using the Pareto Archived Evolution Strategy", *Evolutionary Computation* vol. 8, No. 2, pp. 149-172, 2000.

[5] Rafael Alcalá, Jesús Alcalá-Fdeza, Francisco Herrera, and José Otero, "Genetic learning of accurate and Compact fuzzy rule based systems based on the 2-tuples linguistic representation", *International Journal of Approximate Reasoning*, vol. 44, pp. 45–64, 2007.

[6] Marco Cococcioni, Pietro Ducange, Beatrice Lazzerini, and Francesco Marcelloni, "A Pareto-based multi-objective evolutionary approach to the identification of Mamdani fuzzy systems", *Soft Comput.*, vol. 11, pp. 1013–1031, 2007.

[7] Michela Antonelli, Pietro Ducange, Beatrice Lazzerini, and Francesco Marcelloni, "Learning concurrently partition granularities and rule bases of Mamdani fuzzy systems in a multi-objective evolutionary framework", *International Journal of Approximate Reasoning*, vol. 50, pp.1066–1080, 2009.

[8] Michela Antonelli, Pietro Ducange, Beatrice Lazzerini, and Francesco Marcelloni, "Learning concurrently data and rule bases of Mamdani fuzzy rule-based systems by exploiting a novel interpretability index", *Soft Comput.*, vol. 15, pp. 1981–1998, 2011.

[9] E. H. Mamdani and S. Assilian, "An experiment in linguistic synthesis with a fuzzy logic controller", *Int. J. Man-Mach. Stud.* vol. 7, pp. 1–13, 1975.

[10] KEEL-dataset repository *http://sci2s.ugr.es/keel/datasets.php.*

[11] Pietari Pulkkinen and Hannu Koivisto, "A Dynamically constrained multiobjective genetic fuzzy system for regression problems", *IEEE Transactions on Fuzzy Systems*, vol. 8, no. 1, pp. 161-177, 2010.

Ngày nhận bài 26 – 9 – 2013
Nhận lại sau sửa 26 – 7 – 2014