Kỷ yếu Hội nghị Quốc gia lần thứ X về Nghiên cứu cơ bản và ứng dụng Công nghệ thông tin (FAIR), Đà Nẵng, ngày 17-18/08/2017
DOI: 10.15625/vap.2017.000105

THIẾT KẾ HỆ MỜ TỐI ƯU BẰNG CÁCH KẾT HỢP PHƯƠNG PHÁP TẠO LUẬT CỦA WANG-MENDEL VỚI PSO

Phan Anh Phong[1], Võ Duy Thanh[2]

[1] Trường Đại học Vinh, [2] Trường Đại học Đồng Tháp

phongpa@gmail.com, vdthanhdhdt@gmail.com

TÓM TẮT: *Bài báo đề xuất một phương pháp xây dựng hệ mờ tối ưu bằng cách kết hợp phương pháp one-pass của Wang-Mendel và thuật toán PSO. Phương pháp one-pass được sử dụng để tạo cơ sở luật, còn thuật toán PSO được dùng để tối ưu các tham số cho các tập mờ. Thử nghiệm phương pháp đề nghị cho bài toán dự báo thời gian sống của bệnh nhân viêm tủy cho kết quả tin cậy và ổn định.*

Từ khóa: *Phương pháp thiết kế hệ mờ, Thuật toán tối ưu bầy đàn, Hệ logic mờ, Ứng dụng hệ mờ trong y tế.*

I. GIỚI THIỆU

Hệ mờ là một trong những công cụ phù hợp để mô hình hóa các bài toán mà dữ liệu thu được là mờ, không chắc chắn [6, 8, 12]. Hiệu năng của hệ mờ thường phụ thuộc vào các thành phần cơ bản trong hệ thống, trong đó, hàm thuộc các tập mờ và cơ sở luật đóng vai trò cốt yếu. Để cực tiểu hóa sự phụ thuộc này, các nghiên cứu thường áp dụng các phương pháp học máy nhằm tạo ra một mô hình "ăn khớp" với ứng dụng. Đã có nhiều tiếp cận được đề xuất, một số thực hiện tối ưu cơ sở luật, một số khác tối ưu hàm thuộc và một số thì kết hợp cả hai, tức là vừa tối ưu cả hàm thuộc lẫn cơ sở luật [3, 6, 7, 9, 11, 12].

Để xây dựng và tối ưu hàm thuộc của tập mờ các thuật toán phân cụm và những cải tiến của chúng được sử dụng nhiều [3, 6, 8, 12]. Cơ sở luật của hệ mờ thường được xây dựng hoặc dựa vào kinh nghiệm chuyên gia hoặc/và dựa vào dữ liệu vào – ra [1, 6, 7, 8, 12]. Phương pháp tạo luật của Wang – Mendel (WM) được đề xuất lần đầu trong [1] là sự kết hợp cả hai kỹ thuật trên. Phương pháp này được rất nhiều nghiên cứu đề cập vì sự đơn giản của thuật toán và tránh được sự xung đột của các luật [6]. Năm 2003, Wang tiếp tục hoàn thiện phương pháp nhằm nâng cao khả năng ứng dụng vào thực tiễn [8], tuy nhiên, các tập mờ được xây dựng là ngẫu nhiên hoặc dựa vào chuyên gia nên đã ảnh hưởng nhiều đến sự ổn định và hiệu năng của hệ mờ. Về bản chất, phương pháp WM thực hiện chia lưới dữ liệu để tạo ra các tập mờ, sau đó sinh luật mờ mà không tối ưu hàm thuộc tập mờ.

Gần đây có nhiều mở rộng phương pháp này bằng cách sử dụng các thuật toán phân cụm dữ liệu như FCM, C-mean để tạo hàm thuộc tối ưu cho tập mờ, theo đó làm tăng hiệu năng cho hệ mờ [12]. Một số khác kết hợp phương pháp WM với các thuật toán tối ưu tuyến tính và tối ưu phi tuyến [6, 10, 13].

Sự tiến hóa sinh học hay các phương pháp tìm kiếm ngẫu nhiên của loài vật được các nhà khoa học mô phỏng thành các thuật toán tối ưu, như thuật toán di truyền (GA), thuật toán tối ưu đàn kiến (ACO), thuật toán tối ưu bầy đàn (PSO)… Một ưu điểm chung của các thuật toán dạng này là lời giải có chất lượng tốt. Trong số những kỹ thuật tối ưu dựa vào tìm kiếm trên quần thể, thuật toán PSO và GA được xem là các thuật toán tốt nhất [11]. Tuy nhiên, GA thường có tính toán phức tạp vì quá trình lai ghép, đột biến, trong khi đó PSO lại thích hợp với việc giải và tối ưu các bài toán phi tuyến trong không gian tìm kiếm rời rạc và có ít mối liên hệ với nhau [2, 4]. Hơn nữa, PSO có thời gian hội tụ sớm hơn so với các thuật toán tiến hóa, quần thể khác [12].

Bài báo này đề xuất một phương pháp xây dựng hệ mờ tối ưu bằng cách kết hợp phương pháp WM với thuật toán PSO. Phương pháp WM được sử dụng để tạo cơ sở luật còn PSO được dùng để tối ưu các tham số cho các tập mờ. Sau đó thử nghiệm phương pháp đề xuất cho bài toán dự báo thời gian sống của bệnh nhân viêm tủy.

Trong nghiên cứu này chỉ xét hệ mờ Mamdani với các luật mờ có dạng: IF x_1 is A_{i1} and … and x_n is A_{in} THEN y is B, trong đó, x_1,…, x_n là các biến vào, y là biến ra, A_{i1},..., A_{in}, B là các nhãn ngôn ngữ của các biến vào, biến ra của luật thứ *i.*

Bài báo được tổ chức như sau, tiếp theo phần mở đầu là cơ sở lý thuyết về phương pháp tạo luật mờ của Wang-Mendel và thuật toán PSO. Phương pháp đề xuất thiết kế hệ mờ tối ưu được trình bày trong Phần III. Phần IV là một thử nghiệm của phương pháp đề nghị cho bài toán dự báo thời gian sống của bệnh nhân viên tủy cùng với các kết quả đánh giá mô hình và cuối cùng là kết luận bài báo.

II. KIẾN THỨC CHUẨN BỊ

A. Xây dựng cơ sở luật bằng phương pháp Wang – Mendel

Phương pháp tạo luật mờ của Wang – Mendel được sử dụng nhiều bởi tính đơn giản của cách tiếp cận, cơ sở luật cho hệ mờ sẽ được tạo ra chỉ sau hai lần duyệt tập dữ liệu [1, 8]. Bài báo này chỉ sử dụng cách xây dựng cơ sở luật từ dữ liệu theo của phương pháp WM

Giả sử bộ dữ liệu có *n* các cặp dữ liệu vào – ra như ở công thức 1

$$(x_1^{(i)}, x_2^{(i)}; y^{(i)}) \text{ với } i = 1, 2, ..., n \quad (1)$$

Trong đó x_1, x_2 là các thuộc tính đầu vào, y là thuộc tính đầu ra. Khi đó, quá trình tạo cơ sở luật của WM được thực hiện theo 4 bước sau:

Bước 1 - Phân chia không gian tham chiếu của các biến vào-ra:

- Xác định không gian tham chiếu của các biến vào - ra, giả sử: x1: [x1- , x1+]; x2 : [x2- , x2+] và y: [y- , y+]
- Xác định các nhãn ngôn ngữ có ngữ nghĩa từ nhỏ đến lớn cho mỗi biến, số lượng các nhãn ngôn ngữ là một số lẻ, thường là 3, 5 hoặc 7. Chẳng hạn x_1, x_2 và y có tương ứng là *l*, *u* và *v* giá trị ngôn ngữ: x_1: $\{A_1, A_2, \ldots, A_l\}$; x_2 : $\{B_1, B_2, \ldots, B_u\}$; y : $\{C_1, C_2, \ldots, C_v\}$
- Phân chia không gian tham chiếu của mỗi biến theo số nhãn ngôn ngữ tương ứng, tức là theo *l*, *u* và *v*. Việc phân chia này hoặc là sử dụng ý kiến chuyên gia hoặc là chia ngẫu nhiên (có thể chia đều hoặc không đều). Theo đó, mỗi nhãn ngôn ngữ sẽ được gán một hàm thuộc tương ứng, tùy vào hình dạng hàm thuộc mà xác định bộ tham số đi kèm. Trong thử nghiệm ở mục V sử dụng tập mờ hình thang biểu diễn mỗi nhãn ngôn ngữ.

Như vậy, bản chất Bước 1 của phương pháp WM là gán hàm thuộc cho các nhãn ngôn ngữ của các biến vào-ra.

Bước 2 - Tạo các luật mờ tiềm năng:

Căn cứ vào hàm thuộc của từng biến ngôn ngữ ta tính độ thuộc cho các giá trị $x_1^{(i)}$, $x_2^{(i)}$, $y^{(i)}$.

$x_1^{(i)}$: $\{m_{A1}(x_1^{(i)}), m_{A2}(x_1^{(i)}),\ldots, m_{Al}(x_1^{(i)})\}$ với m_{Al} là độ thuộc của giá trị $x_1^{(i)}$ vào giá trị ngôn ngữ thứ A_l cho thuộc tính x_1

$x_2^{(i)}$: $\{m_{B1}(x_2^{(i)}), m_{B2}(x_2^{(i)}),\ldots, m_{Bu}(x_2^{(i)})\}$ với m_{Bu} là độ thuộc của giá trị $x_2^{(i)}$ vào giá trị ngôn ngữ thứ B_u cho thuộc tính x_2

$y^{(i)}$: $\{m_{C1}(y^{(i)}), m_{C2}(y^{(i)}),\ldots, m_{Cv}(y^{(i)})\}$ với m_{Cv} là độ thuộc của giá trị $y^{(i)}$ vào giá trị ngôn ngữ thứ C_v cho thuộc tính y

Chọn ra giá trị ngôn ngữ có độ thuộc lớn nhất tương ứng với từng giá trị $x_1^{(i)}$, $x_2^{(i)}$, $y^{(i)}$

$$\begin{aligned} &x_1^{(i)} : \max\{m_{A1}(x_1^{(i)}), m_{A2}(x_1^{(i)}),\ldots, m_{Al}(x_1^{(i)})\} \\ &x_2^{(i)} : \max\{m_{B1}(x_2^{(i)}), m_{B2}(x_2^{(i)}),\ldots, m_{Bu}(x_2^{(i)})\} \\ &y^{(i)} : \max\{m_{C1}(y^{(i)}), m_{C2}(y^{(i)}),\ldots, m_{Cv}(y^{(i)})\} \end{aligned} \quad (2)$$

Từ công thức (2) ta suy ra được luật thứ *i* ứng với cặp giá trị vào – ra $(x_1^{(i)}, x_2^{(i)}, y^{(i)})$ có dạng như sau:

$$\text{IF } x_1 \text{ is } A_l \text{ and } x_2 \text{ is } B_u \text{ THEN y is } C_v \quad (3)$$

Với A_l, B_u, C_v là các giá trị ngôn ngữ mà x_1, x_2 và y có độ thuộc lớn nhất, tương ứng.

Kết thúc Bước 2 một tập luật được tạo ra, số luật đúng bằng số bản ghi trong tập dữ liệu huấn luyện. Các luật này cần được tinh chỉnh để tạo ra cơ sở luật. Bước 3 thực hiện tính trọng số mỗi luật để làm căn cứ chọn các luật thích hợp.

Bước 3 - Xác định trọng số của mỗi luật:

Tính trọng số của từng luật trong bộ dữ liệu có n luật ứng với n cặp giá trị $(x_1^{(i)}, x_2^{(i)}, y^{(i)})$

Cho luật thứ i có dạng *IF x_1 is A and x_2 is B THEN y is C* thì công thức tính trọng số được xác định như sau:

$$D(\text{luật } i) = m_A(x_1^{(i)}) \times m_B(x_2^{(i)}) \times m_C(y^{(i)}) \quad (4)$$

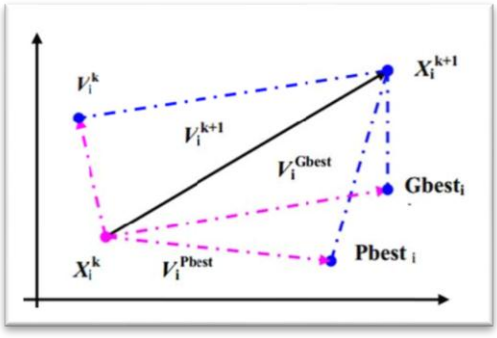
Bước 4 - Loại bỏ luật không phù hợp:

Trong tập luật vừa tìm được ở Bước 2 với các luật giống nhau phần IF nhưng khác nhau phần THEN chỉ giữ lại luật có trọng số D lớn nhất tính theo công thức (4). Như vậy, kết thúc Bước 4 ta thu được một cơ sở luật cho hệ mờ.

B. Thuật toán tối ưu bầy đàn

Thuật toán PSO là một phương pháp tối ưu toàn cục heuristic, được giới thiệu vào năm 1995 bởi J. Kennedy và R. Eberhart nhằm mô phỏng quá trình tìm kiếm thức ăn của bầy chim [2]. Lúc đầu cả đàn bay theo một hướng ngẫu nhiên nào đó. Sau một khoảng thời gian tìm kiếm, một số cá thể sẽ tìm ra những nơi có nguồn thức ăn. Tín hiệu về lượng thức ăn tìm được và vị trí hiện tại sẽ được mỗi cá thể thông báo đến toàn bộ quần thể. Dựa vào thông tin nhận được mỗi cá thể sẽ điều chỉnh hướng bay và vận tốc về vị trí có nhiều thức ăn nhất. Quá trình trên cứ tiếp tục và đến một lúc nào đó cả đàn chim sẽ tìm được nơi có nhiều thức ăn nhất trong không gian tìm kiếm.

PSO được khởi tạo ngẫu nhiên một nhóm cá thể và sau đó tìm nghiệm tối ưu bằng cách cập nhật các thế hệ. Trong mỗi thế hệ, mỗi cá thể được cập nhật theo 2 vị trí tốt nhất. Một là vị trí tốt nhất mà nó từng đạt được đến thời điểm hiện tại, gọi là Pbest. Một nghiệm tối ưu khác mà cá thể này bám theo là nghiệm tối ưu toàn cục Gbest, đó là vị trí tốt nhất trong tất cả quá trình tìm kiếm cả quần thể từ trước thời điểm hiện tại. Nói một cách khác, mỗi cá thể trong quần thể cập nhật vị trí của nó theo vị trí tốt nhất của nó và của cả quần thể tính đến thời điểm hiện tại (xem Hình 1).



Hình 1. Sơ đồ minh họa một điểm tìm kiếm bằng PSO

Các bước thực hiện thuật toán PSO:

Bước 1: Khởi tạo:

Xác định hàm mục tiêu. Gọi maxite là số lần cập nhật vị trí của bầy đàn. Khởi tạo kích thước quần thể với N là số cá thể của quần thể. Khởi tạo các cá thể với vị trí ngẫu nhiên và nằm trong 1 miền giá trị xác định. Mỗi cá thể là một bộ có m biến. Khởi tạo vận tốc ngẫu nhiên cho từng cá thể.

Bước 2: Tính kết quả của hàm mục tiêu đối với từng cá thể.

Bước 3: Dựa trên giá trị hàm mục tiêu của từng cá thể:

Xác định kết quả của Pbest. Pbest là vị trí của cá thể cho kết quả hàm mục tiêu tốt nhất trong k lần cập nhật.

Xác định kết quả của Gbest. Gbest là vị trí của cá thể cho kết quả hàm mục tiêu tốt nhất của quần thể trong lần cập nhật vị trí thứ k.

Bước 4: Nếu kết quả hàm mục tiêu thỏa yêu cầu hoặc số lần cập nhật vị trí = maxite thì dừng.

Bước 5: Cập nhật lại vị trí và vận tốc của các cá thể dựa vào công thức (5) và công thức (6). Quay lại Bước 2.

Công thức tính vị trí và vận tốc của từng cá thể:

$$X_i^{k+1} = X_i^k + V_i^{k+1} \quad (5)$$

$$V_i^{k+1} = w*V_i^k + c1*rand1()*(pbesti - X_i^k) + c2*rand2()*(gbest - X_i^k) \quad (6)$$

$$w = 0.9 - \frac{0.5*\text{lần cập nhật vị trí thứ k}}{\text{số lần cập nhật vị trí}} \quad (7)$$

Trong đó:

X_i^k : Vị trí cá thể thứ i tại thế hệ thứ k; V_i^k : Vận tốc cá thể i tại thế hệ thứ k

X_i^{k+1} : Vị trí cá thể thứ i tại thế hệ k +1; V_i^{k+1} : Vận tốc cá thể i tại thế hệ thứ k + 1

Pbest i : Vị trí tốt nhất của cá thể thứ I; Gbest i : Vị trí tốt nhất của cá thể trong quần thể

w là trọng số quán tính. w thường được tính theo công thức (7).

c1, c2 : các hệ số gia tốc. Thường được sử dụng với giá trị là 2.

rand1, rand2 là các tham số được lấy ngẫu nhiên trong đoạn [0, 1].

III. PHƯƠNG PHÁP ĐỀ NGHỊ

Phương pháp tạo cơ sở luật của WM có thời gian thực hiện nhanh, giải quyết được xung đột của các luật, tuy nhiên hàm thuộc tập mờ được xây dựng tùy ý nên hiệu năng của hệ mờ không cao. Động cơ của thuật toán đề xuất là làm sao xác định các tập mờ phù hợp với đặc trưng của dữ liệu, theo đó chọn ra hệ mờ tốt nhất có thể. Phương pháp này sử dụng PSO để tạo ra hàm thuộc tập mờ, sau đó sử dụng cách tạo luật của WM, tiếp theo tiến hành suy diễn mờ, tính toán theo sai số để tìm hệ mờ tốt nhất.

Gần đây, một số nghiên cứu trong [12, 13] có kết hợp PSO với phương pháp tạo luật WM, tuy nhiên, tài liệu [12] dùng kỹ thuật phân cụm để tạo hàm thuộc cho tập mờ, còn tài liệu [13] sử dụng PSO để tối ưu trọng tâm tập mờ đầu ra của cơ sở luật. Như vậy, có thể thấy rằng, thuật toán đề xuất trong bài báo này là khác biệt với các tiếp cận đó.

Thuật toán đề xuất:

Bước 1. Khởi tạo các tham số

1.1. Khởi tạo các tham số cho hệ mờ:

- Xác định hình dạng và số lượng hàm thuộc cho các biến vào, ra;
- Xác định không gian tham chiếu của các biến (sử dụng bộ dữ liệu)

1.2. Khởi tạo các tham số cho PSO: Kích thước quần thể: *N*; Số lần cập nhật vị trí của bầy đàn: *maxite*; Sai số chấp nhận: *eps*

Bước 2. Xây dựng N hệ mờ đầy đủ

2.1. Gán hàm thuộc cho các tập mờ

2.2. Tạo các luật mờ tiềm năng (áp dụng Bước 2 của phương pháp WM)

2.3. Xác định trọng số từng luật theo công thức (4)

2.4. Loại các luật không phù hợp (áp dụng bước 4 của phương pháp WM)

Bước 3. Lựa chọn hệ mờ tốt nhất

3.1. Tính sai số của N hệ mờ của quần thể ở lần cập nhật thứ *k* (tính giá trị hàm mục tiêu của N hệ mờ theo công thức tính sai số, chẳng hạn dùng LSE)

3.2. Xác định *Pbest* qua các lần cập nhật - Tìm hệ mờ có sai số tốt nhất của mỗi cá thể qua k lần cập nhật

3.3. Xác định *Gbest* – Tìm hệ mờ có sai số tốt nhất của quần thể được tạo từ các hệ mờ có sai số tốt nhất của mỗi cá thể qua *k* lần cập nhật

3.4. Nếu có hệ mờ thỏa điều kiện về sai số hoặc *k* > maxite thì chuyển xuống Bước 5.

Bước 4. Cập nhật giá trị các cá thể

4.1. Cập nhật ví trị và vận tốc các cá thể theo công thức (5), (6)

4.2. k = k +1 và Quay lên Bước 2

Bước 5. Kết thúc

Hệ mờ tốt nhất tìm được với tập tham số ở Gbest.

Thuật toán đề xuất có độ phức tạp tính toán là O(N*$\prod_1^n M_i$)*O(k*m), trong đó, N - số bản ghi dữ liệu huấn luyện; n - số biến vào; M - số tập mờ của mỗi biến; k - số lần cập nhật vị trí bầy đàn; m - số cá thể của quần thể. Đây là một nhược điểm của thuật toán so với phương pháp WM, tuy nhiên đây chỉ là độ phức tạp khi xây dựng mô hình, còn quá trình tính toán được thực hiện như các hệ mờ thông thường.

Để thực hiện thuật toán đề xuất cần làm rõ bốn khía cạnh sau: cách biểu diễn hàm thuộc cho mỗi tập mờ; cách biểu diễn cá thể; cách lựa chọn cá thể tốt nhất và làm sao để cập nhật giá trị các cá thể.

Biểu diễn hàm thuộc

Mỗi nhãn ngôn ngữ được biểu diễn bằng một số mờ hoặc là tam giác, hình thang hay dạng Gauss, khi đó sẽ có *np* tham số đi kèm với mỗi tập mờ. Bộ tham số này có ý nghĩa rất quan trọng đến hiệu năng của hệ mờ. Trong bài báo này sử dụng tập mờ hình thang, do đó, mỗi tập mờ được biểu diễn bằng bộ bốn (a_i, a_{i+1}, a_{i+2}, a_{i+3}). Ví dụ, tập mờ tương ứng với nhãn *Medium* có thể được biểu diễn như công thức (8).

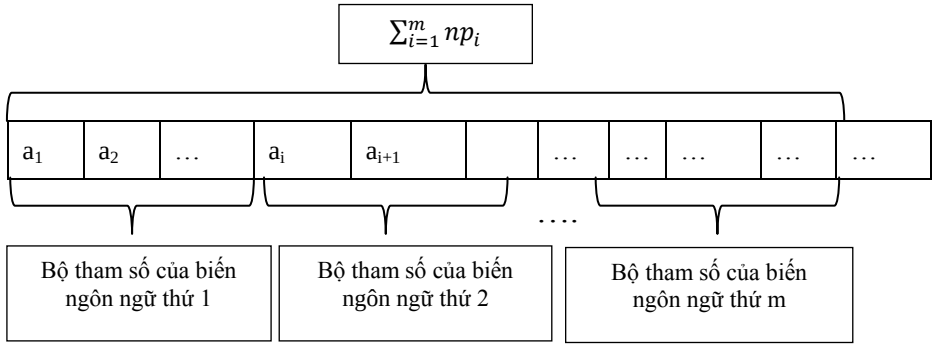
$$\mu_{\text{Medium}}(\text{x}) = \begin{cases} 0\,,\text{x} \le \text{a1} \\ \frac{\text{x}-\text{a1}}{\text{a2}-\text{a1}}\,,\text{a1} < x < a2 \\ 1\,,\text{a2} \le \text{x} \le \text{a3} \\ \frac{\text{a4}-\text{x}}{\text{a4}-\text{a3}}\,,\text{a3} < x < a4 \\ 0\,,\text{x} \ge \text{a4} \end{cases} \quad (8)$$

Biểu diễn cá thể

Mỗi cá thể được biểu diễn bởi $\sum_{i=1}^{m} np_i$ thành phần, trong đó *np* là số các tham số của mỗi tập mờ còn *m* là số biến vào, biến ra của hệ mờ. Hình 2 minh họa một cá thể trong quần thể.

Trong Hình 2, bộ tham số a_i, a_{i+1}… là các nghiệm của một cá thể. Mỗi nghiệm của cá thể được xác định thông qua 2 giá trị:

- X là vị trí và là giá trị cần tìm của nghiệm.
- V là vận tốc. V dùng để cập nhật vị trí mới cho nghiệm của cá thể.



Hình 2. Biểu diễn một cá thể trong quần thể

Lựa chọn cá thể tốt nhất

Duyệt qua tập giá trị của hàm mục tiêu f^k. Ở đây f^k là tập giá trị của các hàm mục tiêu ở lần cập nhật vị trí thứ k đã tìm được ở Bước 3 của thuật toán, gồm có N giá trị tương ứng với N cá thể.

- Xác định kết quả của Pbest: Nếu $f_i^k < f\_Pbest_i$ thì $f\_Pbest_i = f_i^k$ và $\text{Pbest}_i = X_i^k$
  - k là lần cập nhật vị trí thứ k, k ∈ [1, maxite]
  - i là cá thể thứ i, i ∈ [1, N]
  - f_Pbest_i là hàm mục tiêu tốt nhất của cá thể thứ i
  - Pbest_i là vị trí tốt nhất của cá thể thứ i
  - X_i^k là vị trí của cá thể thứ i tại lần cập nhật thứ k
- Xác định kết quả của Gbest: [f_Gbest, index] = min(f_Pbest) và $\text{Gbest} = \text{Pbest}_{\text{index}}$
  - Với *index* là thứ tự của cá thể có hàm mục tiêu nhỏ nhất. Giá trị *index* được xác định qua việc tìm giá trị nhỏ nhất của tập f_Pbest.
  - f_Gbest là hàm mục tiêu tốt nhất
  - Gbest là vị trí của cá thể có hàm mục tiêu tốt nhất trong quần thể.

Cập nhật vị trí và vận tốc các cá thể

Trường hợp kết quả hiện tại xấu thì các cá thể trong quần thể sẽ tiến hành cập nhật vị trí và vận tốc để mong muốn tìm được kết quả tốt hơn. Việc cập nhật của mỗi cá thể sẽ thực hiện bằng cách cập nhật lại vị trí và vận tốc cho từng nghiệm của từng cá thể theo công thức (5) và (6).

IV. THỬ NGHIỆM VÀ BÀN LUẬN

A. Bộ dữ liệu

Bộ dữ liệu về bệnh nhân viêm tủy được tìm thấy trong tài liệu SAS/STAT 9.2 User's Guide The PHREG Procedure (2008) trang 3272 [14]. Trong số 65 bệnh nhân này, có 48 người đã chết và 17 người còn sống sau quá trình nghiên cứu. Mặc dù trong bài toán bệnh nhân viêm tủy có nhiều thuộc tính liên quan đến việc dự đoán thời gian sống của bệnh nhân như: LOGBUN, HGB, PLATELET, AGE, LOGWBC, FRAC, LOGPBM, PROTEIN, SCALC. Tuy nhiên, theo các chuyên gia thuộc tính LOGBUN và HGB mang tính chất quyết định đến kết quả dự báo. Hơn nữa để thuận lợi cho việc thử nghiệm và so sánh với các nghiên cứu trước đây, trong bài báo này chỉ sử dụng các thuộc tính LogBun và HGB làm đầu vào cho hệ mờ và đầu ra là LogTime (giá trị logarithm thời gian sống của bệnh nhân).

B. Thử nghiệm và bàn luận

1. Thiết lập tham số hệ mờ

Thuộc tính đầu vào LogBun chọn 3 giá trị ngôn ngữ: *Low, Medium, High*. Thuộc tính đầu vào HGB chọn 3 giá trị: *Low, Medium, High* và thuộc tính đầu ra LogTime chọn 3 giá trị: *Short, Medium, Long*. Sử dụng tập mờ hình thang để biểu diễn mỗi giá trị ngôn ngữ. Từ bộ dữ liệu ta có thể giới hạn được miền giá trị của mỗi thuộc tính, để tránh rơi vào các trường hợp biên ta cộng (+), trừ (-) một giá trị α_i nào đó vào cận dưới *min* và cận trên *max* của các giá trị dữ liệu, chẳng hạn $\alpha_i = (max - min)/100$. Trong thử nghiệm này α_i được chọn là 0.0005 cho cả 3 thuộc tính, khi đó miền trị của LogBun, HGB và LogTime tương ứng là [0.7776, 2.2361], [4.8994, 14.6006], và [0.0963, 1.9644].

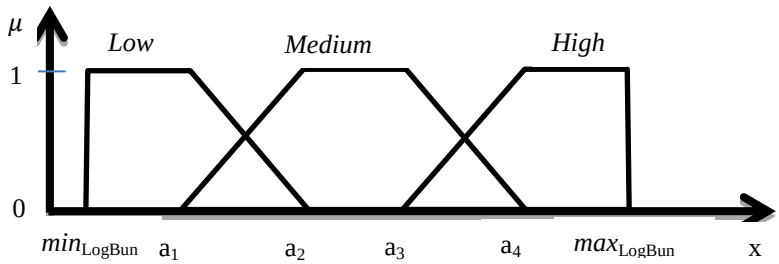
Chia miền trị mỗi biến thành 5 đoạn bằng nhau, theo đó xác định được hàm thuộc của các giá trị ngôn ngữ. Ví dụ hàm thuộc tập mờ của biến *LogBun* được tính theo các công thức (9 - 11):

$$\mu_{\text{Low}}(\text{x}) = \begin{cases} 1, \text{x} \le \text{a1} \\ \dfrac{\text{a2} - \text{x}}{\text{a2} - \text{a1}}, \text{a1} < x < a2 \\ 0, \text{x} \ge \text{a2} \end{cases} \tag{9}$$

$$\mu_{\text{Medium}}(\text{x}) = \begin{cases} 0\ , \text{x} \le \text{a1} \\ \dfrac{\text{x} - \text{a1}}{\text{a2} - \text{a1}}, \text{a1} < x < a2 \\ 1\ , \text{a2} \le \text{x} \le \text{a3} \\ \dfrac{\text{a4} - \text{x}}{\text{a4} - \text{a3}}\ , \text{a3} < x < a4 \\ 0\ , \text{x} \ge \text{a4} \end{cases} \tag{10}$$

$$\mu_{High}(x) = \begin{cases} 0, x \le a3 \\ \dfrac{x - a3}{a4 - a3}, a3 < x < a4 \\ 1, x \ge a4 \end{cases} \tag{11}$$

Với a_1, a_2, a_3, a_4 là các tham số tương ứng với vị trí của các đỉnh như Hình 3.



Hình 3. Tập tham số của các hàm thuộc cho biến ngôn ngữ LogBun

Hàm thuộc các tập mờ của biến HGB được biểu diễn tương tự hàm thuộc của LogBun, lần lượt thay a_1, a_2, a_3 và a_4 bởi a_5, a_6, a_7 và a_8 còn hàm thuộc các tập mờ của LogTime được biểu diễn bởi a_9, a_{10}, a_{11} và a_{12}.

2. Hệ mờ cho bài toán dự báo sử dụng phương pháp WM

Miền trị mỗi biến được chia thành 5 đoạn bằng nhau, khi đó các tham số của tập mờ được thiết kế theo phương pháp WM được thể hiện trong Bảng 1.

Bảng 1. Tham số các tập mờ hình thang của các biến vào - ra

a_1	a_2	a_3	a_4	a_5	a_6	a_7	a_8	a_9	a_{10}	a_{11}	a_{12}
1.0693	1.3610	1.6527	1.9444	6.8396	8.7799	10.7201	12.6604	0.4699	0.8435	1.2172	1.5908

Dựa trên 45 bản ghi huấn luyện, thực hiện các tính toán ở Bước 2 và Bước 3 của phương pháp WM ta tạo được 45 luật và trọng số của chúng. Trong mỗi nhóm luật có cùng phần tiền đề ta chọn luật có trọng số lớn nhất, nếu trọng số bằng nhau thì căn cứ vào tần suất xuất hiện để chọn luật, từ đó ta có cơ sở luật cho hệ mờ.

Thực hiện suy diễn và khử mờ thể với t-norm *min*, t-conorm *max* và phép giải mờ trọng tâm với bộ dữ liệu kiểm tra (20 bản ghi) ta thu được sai số LSE kiểm tra là 2.59.

3. Tối ưu các tham số của hệ mờ bằng PSO

Việc ứng dụng thuật toán tối ưu bầy đàn để tối ưu các tham số cho hàm thuộc của hệ mờ đòi hỏi phải xác định được hàm mục tiêu cần tính cho bài toán. Tiếp theo là việc xác định miền giá trị của từng biến cho mỗi cá thể trong quần thể. Và cuối cùng là việc lựa chọn số cá thể và số lần cập nhật vị trí trong quần thể, việc lựa chọn này tùy thuộc vào kinh nghiệm chủ quan của người thực hiện nhưng sẽ ảnh hưởng đến mức độ chính xác của bài toán và thời gian thực hiện.

Hàm mục tiêu:

Mục tiêu mong muốn là giảm sai số của mô hình đến mức thấp nhất có thể, do đó sử dụng công thức (12) tính LSE để làm hàm mục tiêu:

$$\mathrm{LSE} = \sqrt{\sum_{p=1}^{N}\left(Y'_p - Y_t\right)^2} \tag{12}$$

Trong công thức (12), Y'_p là đầu ra rõ của hệ mờ ứng với bản ghi thứ *p* và Y_t là đầu ra mong muốn.

Biểu biến cá thể:

Mỗi cá thể X trong quần thể là một tập hợp bao gồm *m* biến. Mỗi biến tương ứng với một bộ tham số của hàm thuộc trong hệ mờ. Mỗi tham số nằm trong một miền giá trị nhất định được xác định thông qua bộ dữ liệu mẫu. Giá trị *m* được xác định bằng tổng số lượng các tham số của tất cả các hàm thuộc trong hệ mờ. Ví dụ:

$$\mu_{Low}(x) = \begin{cases} 1, x \le a1 \\ \frac{a2-x}{a2-a1}, a1 < x < a2 \\ 0, x \ge a2 \end{cases} \tag{13}$$

Trong hàm thuộc μ_{Low} có 2 tham số a_1, a_2 tương ứng với 2 biến trong tập các biến của cá thể X.

Tập giá trị tham số của các hàm thuộc tối ưu

Triển khai thuật toán đề xuất trên Matlab 2013a thu được bộ giá trị cho cá thể tốt nhất gồm 12 tham số Bảng 2.

Bảng 2. Tham số tối ưu các tập mờ hình thang

a_1	a_2	a_3	a_4	a_5	a_6	a_7	a_8	a_9	a_{10}	a_{11}	a_{12}
0.7781	2.2356	2.2357	2.2358	10.5108	10.6148	10.7899	12.4336	1.1051	1.1106	1.3382	1.5148

Tập luật trong trường hợp tối ưu.

Từ các tham số ở Bảng 2 tiến hành tinh chỉnh tập luật thu được cơ sở luật tối ưu (Bảng 3)

Bảng 3. Cơ sở luật tối ưu

TT	LogBun	HGB	LogTime
1	*Low*	*Low*	*Medium*
2	*Low*	*Medium*	*Long*
3	*Medium*	*Low*	*Short*
4	*Medium*	*Medium*	*Long*
5	*Medium*	*High*	*Long*
6	*Low*	*High*	*Medium*

Để thuận lợi trong so sánh, bài báo sử dụng dữ liệu tương tự bài báo [3, 7], 45 bản ghi dùng để huấn luyện và 20 bản ghi dùng để dự báo, sai số kiểm tra LSE của phương pháp đề xuất là 1.35 (dòng cuối Bảng 4)

Bảng 4. Sai số kiểm tra LSE các hệ mờ

Phương pháp	Cách sinh luật	Đầu ra hệ mờ	Kỹ thuật tối ưu	LSE-Testing
ANFIS1	Grid partition	Constant	Hybrid	2.08
ANFIS2	Grid partition	Constant	Backpropagation	1.99
ANFIS3	Sub clustering	Linear	Hybrid	2.22
ANFIS4	Sub clustering	Linear	Backpropagation	2.43
Y.Qui, Y. Zhang [3]	Chuyên gia	Type 2 – Fuzzy sets	GA	1.78
Phong-Khang-Đông [7]	FCM	HA – T2FS	GA	1.75
Phương pháp WM [1]	WM	Fuzzy sets	Không sử dụng	2.59
Phương pháp đề nghị	**WM**	**Fuzzy sets**	**PSO**	**1.35**

Thuật toán đề xuất cho sai số kiểm tra LSE bằng 1.35, đây là một kết quả dự báo tốt hơn so với các nghiên cứu gần đây, nhóm Y. Qui, Y. Zhang có LSE là 1.78 [3]; nhóm Phong – Khang - Đông có LSE là 1.75 [7]. Ngoài ra, chúng tôi dùng bộ dữ liệu trên xây dựng hệ mờ ANFIS bằng Fuzzy ToolBox của Matlab 2013a với các phương pháp sinh luật, kỹ thuật tối ưu và đầu ra khác nhau, sau đó đánh giá sai số, một lần nữa phương pháp đề xuất tỏ ra có ưu thế hơn nhiều so với ANFIS, cụ thể là ANFIS2 có sai số LSE nhỏ nhất trong các hệ mờ ANFIS nhưng cũng chỉ đạt LSE là 1.99 (dòng 2, Bảng 4).

Để kiểm tra sự ổn định của phương pháp, trong nghiên cứu này đã thực hiện kỹ thuật kiểm tra chéo kiểu 5-fold, sai số LSE_testing lớn nhất là 1.39 và sai số LSE_testing nhỏ nhất là 0.86. Ngoài ra, bài báo còn xáo trộn ngẫu nhiên tập dữ liệu thành 100 bộ dữ liệu khác nhau theo tỷ lệ 45:20. Sử dụng 45 bản ghi huấn luyện và 20 bản ghi còn lại đánh giá mô hình. Tiến hành xây dựng 100 hệ mờ tốt nhất theo phương pháp đề xuất tương ứng với 100 bộ dữ liệu này với 3 nhãn ngôn ngữ. Qua thực nghiệm cho thấy sai số kiểm tra LSE của phương pháp đề xuất ổn định, với trung bình LSE_testing của 100 hệ mờ là 1.262 và độ lệch chuẩn là 0.124.

Tương tự như trên, bài báo tiếp tục thử nghiệm phương pháp đề nghị với 130 bộ dữ liệu được chia ngẫu nhiên theo tỷ lệ 45:20, sử dụng 5 và 7 nhãn ngôn ngữ cho mỗi biến vào, biến ra của hệ thống. Kết quả tương ứng LSE_testing trung bình của 130 hệ mờ tốt nhất là 1.335 và 1.337, còn độ lệch chuẩn δ = 0.178 và 0.187. Hệ mờ cho kết quả LSE_testing lớn nhất với 5 nhãn ngôn ngữ là 1.851 và với 7 nhãn ngôn ngữ là 1.769 và quả LSE_testing nhỏ nhất với 5 nhãn ngôn ngữ là 0.923 và với 7 nhãn ngôn ngữ là 0.882.

V. KẾT LUẬN

Bài báo đã đề xuất một phương pháp xây dựng hệ mờ từ tập dữ liệu vào-ra. Phương pháp vừa tận dụng được tính đơn giản trong cách tạo luật của Wang-Mendel đồng thời khai thác được khả năng tìm nghiệm tối ưu của thuật toán PSO, do vậy, hệ mờ thu được khá phù hợp với ứng dụng. Kết quả thử nghiệm trên bộ dữ liệu y tế, bước đầu cho thấy sự tin cậy và ổn định của phương pháp đề nghị. Bộ tham số và cơ sở luật tìm được trong thử nghiệm được kiểm tra lại trên công cụ suy diễn mờ FIS, có sẵn trong Fuzzy ToolBox của Matlab 2013a cho kết quả dự báo trùng khớp với hệ mờ đã xây dựng. Ngoài ra, kết quả dự báo trong thử nghiệm còn nổi trội hơn so với hệ mờ nơ-ron ANFIS cũng như một số nghiên cứu gần đây.

Năm 2016, J. Gou và cộng sự đã cho thấy phương pháp WM phụ thuộc nhiều vào đặc trưng của tập dữ liệu, nếu tập dữ liệu huấn luyện không trải đều thì cơ sở luật tạo sẽ không đầy đủ [10]. Trong tương lai, chúng tôi tập trung nghiên cứu sâu theo hướng này, sau đó thử nghiệm phương pháp trên những ứng dụng tiêu biểu.

TÀI LIỆU THAM KHẢO

[1] L. X. Wang and J. M. Mendel, “Generating fuzzy rules by learning from examples”. IEEE Trans. On Systems, Man Cybernet, 22: 1414 – 1427, 1992.

[2] J. Kennedy, R. Eberhart, “Particle swarm optimization”, In proc. of IEEE International Conference on Neural Networks, Vol. 4, pp. 1942–1948, 1995.

[3] Y. Qiu and Y. Qing Zhang, “Statistical Genetic Interval-Value Fuzzy Systems with Prediction in Clinical Trials”, IEEE International Conference on Granular Computing, Pg: 129-132, 2007.

[4] N. Siddique and H. Adeli, Computational Intelligence, John Wiley & Sons, Ltd., Publication, UK, 2013.

[5] K. E. Parsopoulos, M. N. Vrahatis, Particle Swarm Optimization and Intelligence: Advances and Applications, 2010.

[6] J. M. Mendel, Uncertain Rule-Based Fuzzy Logic Systems: Introduction and New Directions. Prentice-Hall, Upper-Saddle River, NJ, 2001.

[7] Phan Anh Phong, Dinh Khac Dong and Tran Dinh Khang, “Predicting Survival Time of Myeloma Patients with Hedge Algebra based Type-2 Fuzzy Logic System”, Proceedings of KSE 2009.

[8] L. X. Wang, “The WM method completed: a flexible fuzzy system approach to data mining”, IEEE Transactions on Fuzzy Systems, 2003, 11(6): 768-782.

[9] Phan Anh Phong, Tran Dinh Khang and Dinh Khac Dong, “A fuzzy rule-based classification system using Hedge Algebraic Type-2 Fuzzy Sets”. In proc. the 35th NAFIPS, 265-270, USA, 2016.

[10] J. Gou et al, “An improved Wang-Mendel method based on the FSFDP clustering algorithm and sample correlation”, Journal of Intelligent and Fuzzy Systems 31(6):2839-2850, 2016.

[11] Ganesan et al, “Optimization of machining parameters in turning process using genetic algorithm and particle swarm optimization with experimental verfification”, IJEST, 3: 1091-1102, 2011.

[12] N. Nedjah, S. O. Costa, L. M. Mourelle, L. S. Coelho, V. C. Mariani, “PSO in Building Fuzzy Systems”, Vol. 357 of the series Studies in Computational Intelligence pp 37-52, 2011.

[13] X. M. Yang et al, “An improved WM method based on PSO for electric load forecasting”, Expert Systems and Applications 37 (12), 8036-8041, 2010.

[14] SAS/STAT 9.2 User’s Guide The PHREG Procedure, 2008.

A COMBINATION OF WANG-MENDEL'S METHOD AND PSO FOR DESIGNING OPTIMAL FUZZY SYSTEMS

Phan Anh Phong, Vo Duy Thanh

ABSTRACT: *In this paper, we proposed a new method for designing fuzzy systems by combining the one - pass Wang – Mendel's method with PSO. The one-pass method is used to create the rule base and PSO is used to optimize the parameters for the membership functions. The proposed method was used for predicting survival time of myeloma patients. The experiment demonstrates that the proposed method performs well for the dataset.*