Tạp chí Tin học và Điều khiển học, T.25, S.1 (2009), 53–68

TIẾP CẬN ĐẠI SỐ GIA TỬ CHO PHÂN LỚP MỜ

NGUYỄN CÁT HỒ[1], DƯƠNG THĂNG LONG[2], TRẦN THÁI SƠN[1]

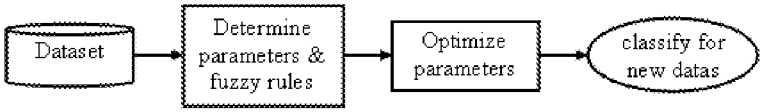
[1] *Viện Công nghệ thông tin, Viện Khoa học và Công nghệ Việt Nam*
[2] *Khoa Công nghệ Tin học, Viện Đại học Mở Hà Nội*

Abstract. In this paper we propose a novel method for fuzzy classification using hedge algebras, which call HA for short. The authors in $[1, 2, 3]$ have pointed out that the results of HA methods depend on the fuzzy measures of some parameters of hedge algebras, so in this paper we use genetic algorithms for optimizing these parameters. We apply the method to iris classification problem, which has been used by many papers and shows that, our results are better in comparing with other methods in $[9, 13-18, 20]$.

Tóm tắt. Bài báo đề xuất một phương pháp mới cho bài toán phân lớp mờ sử dụng đại số gia tử (HA). Các tác giả trong $[1, 2, 3]$ đã chỉ ra rằng kết quả của các phương pháp dựa trên đại số gia tử phụ thuộc đến tham số mờ của các gia tử, ở đây chúng tôi sử dụng giải thuật di truyền để tối ưu các tham số này. Kết quả áp dụng thử nghiệm vào bài toán phân lớp các loài hoa iris cho kết quả rất tốt so với các phương pháp trong $[9, 13-18, 20]$.

1. GIỚI THIỆU

Bài toán phân lớp mờ (fuzzy classification) đã và đang được nhiều tác giả nghiên cứu và ứng dụng khá thành công, các phương pháp được biết đến như hệ mờ $[8, 9, 12-14, 16]$, hoặc hệ mờ $-$nơron $[10, 11, 15]$. Các phương pháp này tiếp cận giải tích đến các tập mờ, sử dụng các phép toán truyền thống trên các tập mờ để lập luận kết quả đầu ra của hệ. Trong đó bao gồm hai giai đoạn chính thiết kế một hệ phân lớp mờ (Hình 1.1). Thứ nhất, xác định một hệ mờ với việc lựa chọn các biến vào, các tham số mờ, phân hoạch các khoảng mờ của biến vào, cách lập luận đầu ra cho hệ mờ và xây dựng các luật mờ; thứ hai, tối ưu các tham số của hệ nhằm tăng hiệu quả phân lớp.



Hình 1.1. Quá trình xây dựng một hệ phân lớp mờ

Xuất phát từ tập dữ liệu mẫu (dataset) $D = \{d_i = (p_{i,1}, ..., p_{i,N}, t_i)\}, i = 1, ..., Q$, trong đó kích thước đầu vào (số thuộc tính $p_{i,1}$,...,$p_{i,N}$) là N, mỗi dữ liệu có nhãn phân lớp t_i tương ứng, kích thước tập mẫu là Q. Các tác giả $[10, 11, 15]$ đã sử dụng phương pháp trích rút luật từ tập dữ liệu mẫu có kết hợp mạng nơron, trong [9] đã áp dụng kỹ thuật tối ưu tham số cho hệ mờ dựa trên giải thuật di truyền.

Bài báo đề xuất một phương pháp mới tiếp cận đại số gia tử (Hedge Algebras – HA). Khai thác từ những đặc trưng về tính có cấu trúc thứ tự của các phần tử là các nhãn ngôn ngữ, sự không phụ thuộc vào biểu diễn giải tích của các tập mờ và các phép toán mờ của đại số gia tử [1–6]. Miền giá trị của mỗi biến thuộc tính đầu vào hình thành một cấu trúc đại số gia tử tương ứng, chúng ta xác các tham số mờ cho mỗi thuộc tính, từ đó xây dựng các hàm định lượng, khoảng mờ, hàm định tính hóa.

Các luật của hệ mờ được biểu diễn dưới dạng Tagaki–Sugeno như sau [17,18]:

$$IF\ X_1\ is\ A_{i,1}\ AND\ X_2\ is\ A_{i,2}\ AND ... AND\ X_N\ is\ A_{i,N}\ THEN\ Y = g_i, \tag{0.1}$$

trong đó, biến vào X_j có miền giá trị là một cấu trúc đại số tương ứng $AX_j, A_{i,j} \in AX_j$ là các hạng từ (phần tử) trong đại số, biến ra $Y = g_i$ tương ứng với nhãn phân lớp, $g_i \in \{c_1, c_2, ..., c_{K_c}\}, K_c$ là số lớp. Sử dụng phép kết gộp N đầu vào cho mỗi luật bằng toán tử nhân, tính độ kích hoạt đầu vào của luật thứ i với một dữ liệu $d = (p_1, ..., p_N)$ như sau

$$\beta_i(p_1, ..., p_N) = \prod j = 1...N sm(p_j, A_{i,j}), i = 1...M, \tag{0.2}$$

trong đó $sm(p_j, A_{i,j})$ hàm đo độ tương tự (tương đương) của thuộc tính đầu vào p_j so với hạng từ $A_{i,j}$ của đại số AX_j, M là số luật. Đầu ra của hệ y^* được xác định bằng kết quả của luật có mức độ đáp ứng đầu vào lớn nhất

$$y^* = g_{i^*},\ i^* = argmax_{i=1,...,M}\beta_i \tag{0.3}$$

Thực tế, xuất phát từ tập dữ liệu mẫu của một bài toán cụ thể chúng ta phải xác định tập các tham số mờ cho đại số gia tử (tham số gia tử) của mỗi thuộc tính vào, từ đó xây dựng tập các luật mờ. Hơn nữa, số luật mờ (M) tìm được càng nhỏ gọn càng tốt nhưng vẫn đảm bảo hiệu quả áp dụng. Các tác giả trong [1–6] đã cho thấy ảnh hưởng của tham số gia tử đến kết quả áp dụng khá lớn, trong bài này chúng tôi sẽ áp dụng giải thuật di truyền để tối ưu các tham số này.

Ngoài ra, các mẫu dữ liệu có thể dạng ngôn ngữ (linguistic) hoặc/và dữ liệu định lượng trên miền thực của bài toán, ký hiệu U⊂R, trong khi mỗi biến ngôn ngữ X có miền ngữ nghĩa định lượng tương ứng trên đoạn $[0, 1]$, ký hiệu U_s. Chúng ta phải chuẩn hóa dữ liệu (chuyển từ miền thực về miền ngữ nghĩa định lượng) bằng ánh xạ tuyến tính $f : U \to U_s$ và chuyển ngược lại bằng ánh xạ ngược của nó $f^{-1} : U_s \to U$.

2. ĐẠI SỐ GIA TỬ VÀ CÁCH TIẾP CẬN ĐỂ PHÂN LỚP MỜ

Như đã trình bày ở trên, ta xem mỗi thuộc tính đầu vào có cấu trúc miền ngữ nghĩa tương ứng là một đại số gia tử. Xét các đại số này là các đại số gia tử tuyến tính đầy đủ [1–6], ký hiệu $AX = (X, G, C, H, \sum, \Phi, \leq)$, trong đó X là tập các hạng từ (terms) của đại số, G tập phần tử sinh, C các giá trị hằng, H là tập các gia tử (hedges),Φ gia tử tới hạn min, $\sum$ gia tử tới hạn max, $\leq$ quan hệ thứ tự ngữ nghĩa giữa các hạng từ. Trước hết, chúng tôi trình bày một số khái niệm, ký hiệu và kiến thức cơ sở của đại số gia tử cũng như cách tiếp cận để ứng dụng cho bài toán phân lớp mờ.

Chúng ta định nghĩa khoảng mờ (fuzziness intervals) của một hạng từ x dựa trên thứ tự nghữ nghĩa của các hạng từ và độ đo tính mờ của chúng.

Định nghĩa 2.1. [5] Khoảng mờ của một hạng từ là một khoảng con của $[0,1]$ có độ dài đúng bằng độ đo tính mờ $fm(x)$, ký hiệu $\Im(x)$, được xác định bằng quy nạp theo độ dài của x như sau:

(i) Với độ dài x bằng 1 ($l(x) = 1$), tức là $x \in \{c^-, c^+\}, \Im_{fm}(c^-)$ và $\Im_{fm}(c^+)$ là các khoảng con và tạo thành một phân hoạch của [0,1], thỏa $\Im_{fm}(c^-) \leq \Im_{fm}(c^+)$, tức là $\forall u \in \Im_{fm}(c^-), \forall v \in \Im_{fm}(c^+) : u \leq v$, điều này phù hợp với thứ tự ngữ nghĩa của c^- và c^+. $|\Im_{fm}(c^-)| = fm(c^-)$ và $|\Im_{fm}(c^+)| = fm(c^+)$, trong đó $|\Im_{fm}(x)|$ ký hiệu độ dài của $\Im_{fm}(x)$.

(ii) Giả sử $\forall x \in X$ độ dài bằng $k(l(x) = k)$ có khoảng mờ là $\Im_{fm}(x)$ và $|\Im_{fm}(x)| = fm(x)$, các khoảng mờ của $y = h_i x, \forall i \in [-p\hat{}q](l(y) = k+1)$ là tập $\{\Im_{fm}(h_i x)\}$ thỏa mãn một phân hoạch của $\Im_{fm}(x), |\Im_{fm}(h_i x)| = fm(h_i x)$ và có thứ tự tuyến tính tương ứng với thứ tự của tập $\{h_{-q}x, h_{-q+1}x, ..., h_p x\}$.

Khi $l(x) = k$, ta ký hiệu $\Im(x)$ thay cho $\Im_{fm}(x), X_k = \{x \in X : l(x) = k\}$ là tập hạng từ độ dài đúng $k, I_k = \{\Im_k(x) : x \in X_k\}$ là tập tất cả các khoảng mờ mức k ($k - khoảng$).

Rõ ràng hai khoảng mờ bằng nhau, ký hiệu $\Im(x) = \Im(y)$, khi chúng được xác định bởi cùng một hạng từ ($x = y$), tức là $\Im_L(x) = \Im_L(y)$ và $\Im_R(x) = \Im_R(y)$, trong đó $\Im_L(x), \Im_R(x)$ ký hiệu điểm trái cùng và phải cùng của khoảng mờ $\Im(x)$. Ngược lại chúng ta gọi hai khoảng mờ khác nhau, $\Im(x) \neq \Im(y)$.

Định đề 2.1. [5] *Cho $AX = (X, G, C, H, \sum, \Phi, \leq)$ là một đại số gia tử tuyến tính tự do, ta có,*

(i) *Nếu $Sign(h_p x) = +1$, thì $\Im(h_{-q}x) \leq \Im(h_{-q+1}x) \leq ... \leq \Im(h_{-1}x) \leq \Im(h_1 x) \leq \Im(h_2 x) \leq ... \leq \Im(h_p x)$, và nếu $Sign(h_p x) = -1$, thì $\Im(h_{-q}x) \geq \Im(h_{-q+1}x) \geq ... \geq \Im(h_{-1}x) \geq \Im(h_1 x) \geq \Im(h_2 x) \geq ... \geq \Im(h_p x)$;*

(ii) *Tập $I_k = \{\Im_k(x) : x \in X_k\}$ là một phân hoạch của khoảng $[0,1]$;*

(iii) *Cho một số $m, \{\Im(y) : y = h_m ... h_1 x, \forall h_m, ..., h_1 \in H\}$ là một phân hoạch của khoảng mờ $\Im(x)$;*

(iv) *Tập $I_k = \{\Im_k(x) : x \in X_k\}$ mịn hơn (finer) tập $I_{k-1} = \{\Im_k(x) : x \in X_{k-1}\}$, tức là mọi khoảng trong I_k đều được chứa trong I_{k-1};*

(v) *Nếu $x < y$ và $l(x) = l(y) = k$ thì $\Im_k(x) \leq \Im_k(y)$ và $\Im_k(x) \neq \Im_k(y)$;*

(vi) *$\forall x, y \in X$ xác định hai khoảng mờ $\Im_k(x)$ và $\Im_l(y)$, chúng hoặc không có quan hệ kế thừa; hoặc có quan hệ kế thừa (ký hiệu $\Im_k(x) \sim \Im_l(y)$) nếu $\exists \Im_v(z) \in I_v, v \leq min(k,l), \Im_L(z) \leq \Im_L(y), \Im_R(z) \geq \Im_R(y)$, và $\Im_L(z) \leq \Im_L(x), \Im_R(z) \geq \Im_R(x)$, hay $\Im_v(z) \supseteq \Im_k(x)$ và $\Im_v(z) \supseteq \Im_l(y)$, tức là x, y được sinh ra từ $z, x = h_{i_n} ... h_{i_1} z, y = k_{j_m} ... k_{j_1} z, \forall h_i, k_j \in H$.*

Các tính chất (i–v) đã được chứng minh trong [5], (vi) dễ dàng được suy ra từ tính phân hoạch của các khoảng mờ.

Chúng ta thấy dựa trên cấu trúc thứ tự của X, phần tử x nằm ở giữa hai tập $\{h_{-i}x : -q \leq i \leq -1\}$ và $\{h_j x : 1 \leq j \leq p\}$, hơn nữa ta có

$$\sum_{i \in [-q,-1]} |\Im(h_i x)| = fm(x) \times \sum_{i \in [-q,-1]} \mu(h_i) = \alpha.fm(x) = \alpha |\Im(x)|. \tag{0.4}$$

Điều này gợi ý chúng ta chọn điểm cuối chung của hai khoảng mờ $\Im(h_{-1}x)$ và $\Im(h_{+1}x)$ là giá trị định lượng ngữ nghĩa $\upsilon(x)$ (xem [3]) của hạng từ x.

Ánh xạ định lượng ngữ nghĩa υ phụ thuộc vào các tham số độ đo tính mờ $fm(c^-), fm(c^+)$, $\mu(h)$. Các tham số này được thiết lập mang tính chủ quan, phụ thuộc ngữ cảnh sử dụng và có thể thay đổi tùy ý. Chúng ta xem giá trị $\upsilon(x) \in [0,1]$, khác với một giá trị số thực trên đoạn [0,1] trong tự nhiên, như một biểu diễn định lượng ngữ nghĩa của x. Như vậy, ngữ nghĩa của đẳng thức $\upsilon(x) = a$ khác với ngữ nghĩa của $b = a$, trong đó a, b là các giá trị thực thuần túy.

Hơn nữa, trong các ứng dụng mà đặc biệt là bài toán phân lớp, xuất phát từ dữ liệu định lượng, ta xây dựng hệ các luật mờ để phân lớp, vì vậy việc định tính hóa các dữ liệu thực này rất cần thiết. Dựa trên phân hoạch các khoảng mờ mức k, I_k, mỗi khoảng mờ $\Im(x)$ mang ý nghĩa định lượng cho hạng từ x, ta sẽ ánh xạ mỗi giá trị trong khoảng mờ này đến hạng từ x tương ứng.

Định nghĩa 2.2. Cho $AX = (X, G, C, H, \sum, \Phi, \leq)$ là một đại số gia tử tuyến tính đầy đủ, các tham số mờ được thiết lập, và một số $k > 0$. Ánh xạ định tính ngữ nghĩa mức k của một giá trị ngữ nghĩa định lượng v, ký hiệu $\pi : [0,1] \to X_k$, được định nghĩa dựa trên phân hoạch các khoảng mờ I_k như sau: $\forall v \in [0,1], \pi(v) = x \in X_k$, sao cho $v \in \Im_k(x)$.

Bây giờ, ta đi xây dựng thủ tục tính phân hoạch các khoảng mờ mức $k+1$ cho một khoảng mờ ở mức k, tính phân hoạch tất cả các khoảng mờ mức k cho một đại số gia tử AX tuyến tính đầy đủ với tham số bất kỳ.

Dựa trên tính thứ tự của các khoảng mờ trong phân hoạch, (xem Định đề 2.1. (i)), ký hiệu $\Im_L(x)$ và $\Im_R(x)$ là điểm trái và phải của khoảng mờ $\Im(x)$, tính khoảng mờ mức $k+1$ thông qua thủ tục sau.

Thủ tục 2.1. Tính phân hoạch các khoảng mờ mức $k+1$ ($\Im_{k+1}(hx) \forall h \in H$) của khoảng mờ mức $k(\Im_k(x))$.

Input: $\Im_k(x), \mu(h) \forall h \in H = \{h_{-q}, h_{-q+1}, ..., h_{-1}, h_1, h_2, ..., h_p\}$.

Output: $\Im_{k+1}(hx) \forall h \in H$.

Actions:

Step 1) Nếu $Sign(h_p x) = 1$ (tức là $\Im(h_{-q}x) \leq ... \leq \Im(h_{-1}x) \leq \Im(h_1 x) \leq ... \leq \Im(h_p x)$), đặt $j = -q, t = 1$. Ngược lại ($Sign(h_p x) = -1$), đặt $j = p, t = -1$.

Step 2) Tính khoảng mờ xuất phát $\Im(h_j x) = (\Im_L(x), \Im_L(x) + \mu(h_j).|\Im(x)|]$, nếu $\Im(x)$ là khoảng mờ đóng trái thì đặt $\Im(h_j x)$ cũng đóng trái.

Step 3) Đặt $i = j$, thay đổi j theo $t, j = j + t$, nếu $j = 0$ thì $j = j + t$ (không xét h_0).

Step 4) Tính khoảng mờ tiếp theo $\Im(h_j x) = (\Im_R(h_i x), \Im_R(h_i x) + \mu(h_j).|\Im(x)|]$.

Step 5) Nếu $j = -q$ hoặc $j = p$ thì dừng, ngược lại quay lên Step 3) tính tiếp.

Thủ tục này được dùng để tính toán phân hoạch tất cả các khoảng mờ I_k với k cho trước của một đại số gia tử AX như sau.

Thủ tục 2.2. Tính phân hoạch tất cả các khoảng mờ cho đến mức k^*, $I_j = \{\Im_j(x) : x \in X_j\}$ với $j = 1, 2, ..., k^*$ của AX.

Input: Các tham số mờ $fm(c^-), fm(c^+), \mu(h), \forall h \in H$. số k^* nguyên dương ($k^* > 0$),

Output: Tất cả các khoảng mờ từ mức 1 đến mức $k^*, I_j = \{\Im_j(x) : \forall x \in X_j\}(j = 1...k^*)$,

Actions:

Step 1) Đặt $X_1 = \{c^-, c^+\}$, tính $I_1 = \{\Im(c^-), \Im(c^+)\}$, với $\Im(c^-) = [0, fm(c^-)], \Im(c^+) = (fm(c^-), 1]$,

Step 2) Lặp với $j = 2$ đến k^*, tính

2.i) $X_j = \{hx : \forall h \in H, \forall x \in X_{j-1}\}$,

2.ii) $I_j = \bigcup_{x \in X_{j-1}} \{\Im_j(hx) : \forall h \in H\}$, trong đó tính $\{\Im_j(hx) : \forall h \in H\}$ dựa trên $\Im_{j-1}(x) \in I_{j-1}$ (sử dụng Thủ tục 2.1).

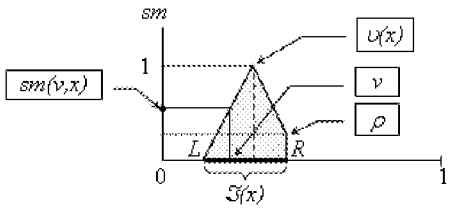
Như vậy, sau khi xác định tham số mờ của HA và một số k, dùng thủ tục trên tính một phân hoạch các khoảng mờ mức k, $\Im_k(x)$, ta sử dụng phân hoạch này làm cơ sở xác định độ đo tương đương giữa một giá trị định lượng với một nhãn ngôn ngữ như sau.

Định nghĩa 2.3. Cho phân hoạch các khoảng mờ mức k, $\Im_k(x)$, của một HA, độ đo tương tự của một giá trị định lượng $v \in [0, 1]$ và một hạng từ $x \in X_k$ trong HA, ký hiệu $sm(v, x)$, xác định dựa trên khoảng mờ mức k của x theo dạng (Hình 2.1) sau

i) $sm(v, x) = (v - \Im_L(x))/(\upsilon(x) - \Im_L(x))$, nếu $\Im_L(x) < v \leq \upsilon(x)$,

ii) $sm(v, x) = (\Im_R(x) + v.(\rho - 1) - \rho.\upsilon(x))/(\Im_R(x) - \upsilon(x))$, nếu $\upsilon(x) < v \leq \Im_R(x)$,

iii) $sm(v, x) = 0$, ngược lại, tức là $v \notin \Im(x)$.



Hình 2.1. Dạng tam giác đo độ tương tự giữa v và x

trong đó, ρ là một tham số đo mức độ tương tự giữa v và x khi $v = \Im_R(x)$, ở đây ta chọn $\rho = 0,5$ với $v = \Im_L(x)$ thì $sm(v, x) = 0$, vì khoảng mờ $\Im(x)$ ở dạng đóng phải và mở trái. Hàm đo độ tương tự $sm(v, x)$ ở đây được dùng để xác định độ kích hoạt đầu vào của một dữ liệu mẫu đối với vế trái của luật trong (1.2), như vậy độ kích hoạt $\beta(p_1, ..., p_N)$ càng lớn nếu các $sm(p_j, A_j), j = 1, ..., N$ càng lớn, và ngược lại. Khi $v = \upsilon(x)$ thì $\beta = 1$, và v càng xa giá trị định lượng của x thì β càng giảm dần.

Thủ tục 2.3. Tính khoảng mờ mức cực đại bao hàm hai khoảng mờ cho trước.

Input: Hai khoảng mờ bất kỳ $\Im_k(x), \Im_l(y)$, và tập tất cả các phân hoạch $I_j(j = 1, 2, ...k^*)$ của AX.

Output: $\Im_v(z)$ chứa hai khoảng mờ $\Im_k(x)$ và $\Im_l(y)$ đã cho khi chúng có quan hệ "kế thừa", rõ ràng $v \leq min(k, l)$.

Actions:

Step 1) Đặt $v = min(k, l)$,

Step 2) Nếu $\exists \Im_v(z) \in I_v, \Im_v(z) \supseteq \Im_k(x)$ và $\Im_v(z) \supseteq \Im_l(y)$ thì kết quả là $\Im_v(z)$, với z là hạng từ tương ứng khoảng mờ $\Im_v$,

Step 3) Ngược lại giảm v đi một ($v = v - 1$) và nếu $v = 0$ thì không có kết quả, tức là hai khoảng mờ $\Im_k(x)$ và $\Im_l(y)$ không có quan hệ kế thừa, nếu $v > 0$ lặp lại Step 2).

Định nghĩa 2.4. Cho một đại số gia tử tuyến tính AX và các tham số gia tử, với bất kỳ hai hạng từ $x, y \in X$ xác định hai khoảng mờ tương tứng $\Im_k(x)$ và $\Im_l(y)$. Ta định nghĩa mức độ gần nhau, ký hiệu $\lambda(.,.)$, của hai khoảng mờ như sau:

(i) Nếu hai khoảng mờ có quan hệ kế thừa $\Im_k(x) \sim \Im_l(y)$ thì $\lambda(\Im_k(x), \Im_l(y)) = v/max(k, l)$, với v là mức khoảng mờ tính trong Thủ tục 2.3.

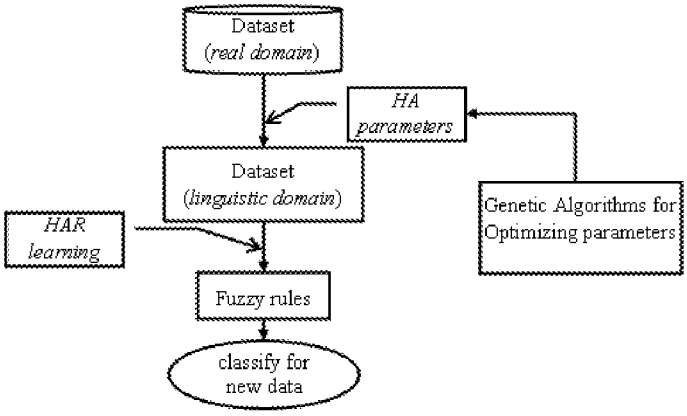
(ii) Ngược lại, tức là hai khoảng mờ không có quan hệ kế thừa, thì $\lambda = 0$.

Định nghĩa này cho phép xác định hai hạng từ x, y có thể kết nhập thành một hạng từ z đại diện mà không làm mất nhiều ý nghĩa của x, y bằng cách dựa vào λ. Nếu $\lambda = 0$ thì không thể kết nhập vì chúng xác định hai khoảng mờ không quan hệ kế thừa, ngược lại chúng ta chọn z là hạng từ có khoảng mờ tương ứng ở mức thấp hơn. Giả sử $k > l$ và $\Im(x) \supseteq \Im(y)$, khi đó $z = y$, khoảng mờ $\Im_k(x)$ được mở rộng và biểu diễn bởi $\Im_l(y)$.

Rõ ràng phép kết nhập trên sẽ không làm mất thông tin, vì ở đây chúng ta mở rộng khoảng mờ. Tuy nhiên, λ càng bé thì phép kết nhập càng làm tăng tính mờ của một khái niệm mờ, và dẫn đến nó phủ các khái niệm khác trong cùng một phân hoạch. Khi $\lambda = 1$, có nghĩa là $x = y$, dẫn đến $\Im_v(z) = \Im_k(x) = \Im_l(y)$.

3. THIẾT KẾ MỘT MÔ HÌNH PHÂN LỚP MỜ THEO *HA*

Như đã trình bày ở trên, mô hình của chúng tôi được chia làm hai giai đoạn (Hình 3.1). Trước hết chúng ta thiết lập bộ tham số gia tử, áp dụng vào giai đoạn thứ nhất để phân hoạch các khoảng mờ trên miền của các thuộc tính vào, chuyển các dữ liệu thực về dạng ngôn ngữ (linguistic domain). Giai đoạn thứ hai áp dụng một phương pháp học để trích rút các luật mờ (dưới dạng ngôn ngữ) từ tập dữ liệu mẫu (dataset) – HAR learning. Tuy nhiên, hiện nay chưa có phương pháp để xác định bộ tham số gia tử thích hợp, chủ yếu bằng thực nghiệm. Trong bài báo này, tác giả đã áp dụng giải thuật di truyền (GA) để tối ưu bộ tham số này nhằm đạt kết quả phân lớp tốt hơn với kích thước tập luật nhỏ hơn.



Hình 3.1. Sơ đồ quá trình xây dựng hệ phân lớp mờ theo HA

Xét tập dữ liệu mẫu của bài toán phân lớp có N thuộc tính vào, đầu ra là chỉ số lớp tương ứng (hoặc nhãn của lớp) có dạng $D = \{(p_i, t_i) : i = 1, ..., Q\}, p_i = (p_{i,1}, p_{i,2}, ..., p_{i,N}) \in R^N, t_i \in \{1, 2, ..., K_c\}$ là nhãn chỉ số lớp, K_c là số lớp của bài toán. Sử dụng ánh xạ tuyến tính để chuyển tập dữ liệu mẫu về miền định lượng $U_{s_j} = [0, 1]$, với mỗi biến vào X_j xác định miền tham chiếu $U_j = [a_j, b_j]$, ta có:

$$f_{u_j}^{-1} : U_{s_j} \to U_j, \forall y \in U_{s_j}, f_{u_j}^{-1}(y) = a_j + y.(b_j - a_j), \tag{0.5}$$

và ánh xạ ngược

$$f_{u_j} : U_j \to U_{s_j}, \forall x \in U_j, f_{u_j}(x) = (x - a_j)/(b_j - a_j). \tag{0.6}$$

Như vậy, tập dữ liệu mẫu được chuyển về dạng định lượng, ký hiệu là DS, mỗi miền định lượng của biến vào được phân hoạch thành các khoảng mờ mức k cho trước, dựa trên ánh xạ định tính trong HA (Định nghĩa 2.3), mỗi dữ liệu trong miền định lượng thuộc một khoảng mờ sẽ được xác định một hạng từ tương ứng, khi đó tập dữ liệu mẫu được chuyển về dạng ngôn ngữ (DL). Giai đoạn học trích rút các luật mờ sẽ thực hiện trên DL.

Thuật toán học trích rút các luật mờ từ DL theo cách xây dựng một tập các vế trái có thể (candidates of left) từ không gian vào $p_i = (p_{i,1}, p_{i,2}, ..., p_{i,N})$, với mỗi vế trái tìm được ta tính tổng mức độ kích hoạt đầu vào của nó theo từng lớp trên các mẫu dữ liệu và sinh luật mới cho vế trái này với đầu ra tương ứng là lớp có tổng mức đáp ứng đầu vào lớn nhất. Thuật toán được trình bày như sau.

Thuật toán 3.1: HAR learning algorithms,

Input:

+ Dataset $D = (x_p, t_p)$ với K_c cụm, $p = 1, ..., Q$,

+ Các tham số mờ gia tử $fm(c^-), fm(h) \forall h \in H$,

+ Các mức khoảng mờ cho các thuộc tính $k_j^*, j = 1...N$.

Output: tập các luật mờ $R_1, ..., R_M$.

Methods:

Step 1) Khởi tạo tập luật $R = \varnothing$, tập vế trái $L = \varnothing$.

Step 2) Tính phân hoạch các khoảng mờ $I_{k_j^*}$ cho các thuộc tính theo tham số mờ gia tử và mức $k_j^*, j = 1, 2, ..., N$.

Step 3) Với mỗi mẫu dữ liệu $(x_p, t_p) \in D$:

i) Với mỗi giá trị của thuộc tính đầu vào $x_{p,j} \in x_p$, tìm $A_j = \pi_{k^*}(x_{p,j})$ theo Định nghĩa 2.2,

ii) Tạo một tuyển vế trái $left = (A_1, A_2, ..., A_N)$,

iii) Nếu left chưa có trong danh sách vế trái, L, thì thêm left: $L = L \cup left$.

Step 4) Tính tổng mức độ kích hoạt đầu vào của các dữ liệu mẫu đối với tuyển các vế trái theo từng lớp, chia cho số mẫu dữ liệu trong mỗi lớp tương ứng, ký hiệu $C_i(t)$,

$$\forall t \in \{1, ..., K_c\}, \ \forall left_i \in L : C_i(t) = \sum_{p=1,...,Q} \{ \prod_{j=1...N} sm(x_{p,j}, A_{i,j}) : t_k = t\}/count(t). \tag{0.7}$$

trong đó hàm $sm(.,.)$ tính theo Định nghĩa 2.3, $count(t)$ là đếm số mẫu dữ liệu có đầu ra mong muốn là t.

Step 5) Với mỗi vế trái $left_i \in L$:

i) Tìm chỉ số lớp có mức đáp ứng lớn nhất, $c = argmax_{t=1...K_c}\{C_i(t)\}$.

ii) Tạo luật mới có vế trái là $left_i$ và vế phải tương ứng là c, $r_i = left_i \to c$.

iii) Thêm luật này vào tập luật R nếu chưa tồn tại, $R = R \cup r_i$.

Step 6) Loại bỏ các luật dư thừa trong tập R thu được ở Step 5, sử dụng Thủ tục 3.1 sau, dựa vào độ đo khả năng kết nhập (integratable) giữa hai luật ($itgr$), thay thế hai luật có $itgr$ cao nhất bằng một luật mới sinh ra từ hai luật đó bằng cách tăng tính mờ của các thuộc tính vào trong luật. Quá trình sẽ dừng khi không tìm được hai luật có $itgr$ lớn nhất và lớn hơn hơn ngưỡng σ cho trước.

Hàm $itgr$ được xây dựng trên cơ sở các thuộc tính vào của luật. Xét hai luật có vế phải giống nhau, vế trái của mỗi luật xác định một tập các khoảng mờ, ký hiệu $\{\Im_{k_1}(x_1), ..., \Im_{k_N}(x_N)\}$ và $\{\Im_{l_1}(y_1), ..., \Im_{l_N}(y_N)\}$, khi đó dựa trên đánh giá mức độ phủ nhau $\lambda_j = \lambda(\Im_{k_j}(x_j), \Im_{l_j}(y_j))$, $j = 1...N$, (Định nghĩa 2.5), ta tính hàm $itgr$ như sau

$$itgr(r_k, r_l) = \begin{cases} 0 & right(r_k) \neq right(r_l), \\ min_{j=1...N}\lambda_j & else, \end{cases} \tag{0.8}$$

trong đó $right(r)$ là vế phải của luật r.

Vídụ 3.1. Cho hai luật $r_i = (VASmall, VLarge, class_1), r_j = (MASmall, Large, class_1)$, ta có $\lambda_1 = 2/3, \lambda_2 = 1/2, itgr(r_i, r_j) = 1/2$, khi đó thay bằng một luật mới

$$r' = (ASmall, Large, class_1).$$

Trường hợp hai luật $r_i = (VALarge, LVLarge, class_1), r_j = (MASmall, LVLarge, class_1)$ có $\lambda_1 = 0, \lambda_2 = 3/3, itgr(r_i, r_j) = 0$, không thể kết nhập hai luật này thành luật mới.

b) *Tối ưu tham số của phương pháp bằng TGA*

Trong [1 – 6] các tác giả chỉ ra mức độ ảnh hưởng của bộ tham số gia tử đến kết quả ứng dụng của một số phương pháp tiếp cận theo đại số gia tử, phương pháp này không nằm ngoài tính chất đó, kết quả trích rút các luật mờ với mục tiêu số luật càng nhỏ càng tốt nhưng đảm bảo sai số phân lớp cho phép. Một cách khá mềm dẻo và linh hoạt là sử dụng giải thuật di truyền (GA) [7, 9, 13, 18] tối ưu các tham số, đặc biệt trong [7] đã chỉ ra cách kết hợp tham số nhiệt (phỏng theo nhiệt độ tôi luyện thép) nhằm tăng tốc độ hội tụ của giải thuật, gọi là TGA.

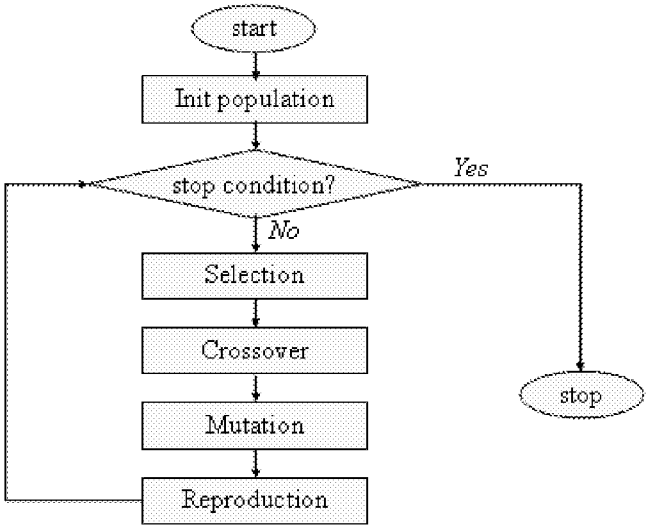
Ở đây, ta chọn số gia tử là 2, $H = \{L, V\}$, trong đại số của mọi thuộc tính. Vậy chúng ta có bộ tham số gia tử của các biến vào $PAR_{ha} = (fm_j(c^-), \mu_j(L)), j = 1, ..., N$, các mức khoảng mờ $PAR_k = (k_1, ..., k_N)$. Để ý rằng $fm_j(c^+) = 1 - fm_j(c^-)$ và $\mu_j(V) = 1 - \mu_j(L)$. Để áp dụng giải thuật GA [7] thực hiện tối ưu các tham số, trước hết, mỗi cá thể trong GA được mã hóa dưới dạng số thực trong đoạn [0,1] biểu diễn các tham số của mô hình, gồm:

• Một nhiễm sắc thể có N gen, $(fm_1(c^-), fm_2(c^-), ..., fm_N(c^-))$, biểu diễn tham số độ đo tính mờ của các phần tử sinh âm trong các đại số gia tử.

• Một nhiễm sắc thể có N gen, $(\mu_1(L), \mu_2(L), ..., \mu_N(L))$, biểu diễn các tham số độ đo tính mờ của gia tử L trong đại số của các biến vào tương ứng $AX_j, j = 1, 2, ..., N$.

• Một nhiễm sắc thể có N gen, $(k1, k_2, ..., k_N)$ biểu diễn mức khoảng mờ cần phân hoạch trên miền các biến vào $X_j, j = 1, 2, ..., N$.

Giải thuật GA xuất phát từ một tập các cá thể ban đầu chọn ngẫu nhiên, sử dụng các phép chọn lọc (selection) để chọn ra những cá thể bố mẹ cho quá trình lai ghép, phép lai ghép (crossover) thực hiện trên các cá thể bố mẹ để sinh ra các cá thể con, phép đột biến (mutation) các cá thể con nhằm tránh rơi vào tối ưu địa phương, cuối cùng phép tái tạo (reproduction) chọn ra những cá thể tốt cho vào thế hệ tiếp theo. Quá trình này được tiếp tục với hy vọng thế hệ sau gồm các cá thể tốt hơn thế hệ trước (Hình 3.2).



Hình 3.2. Sơ đồ hoạt động của giải thuật di truyền

Trong đó, các phép di truyền cũng như điều kiện dừng của giải thuật dựa trên cơ sở hàm đánh giá độ phù hợp của từng cá thể, $FITNESS(.)$. Với mục tiêu tối ưu các tham số nhằm tăng độ chính xác trong phâp lớp và giảm kích thước tập luật sinh ra, hàm $FITNESS(.)$ được thiết kế dựa trên những kết quả này. Thực vậy, với mỗi bộ tham số ứng với một cá thể, sử dụng Thuật toán 3.1 sinh ra tập các luật mờ phân lớp dựa trên tập mẫu dataset. Hàm $FITNESS(.)$ xây dựng trên cơ sở gồm 3 yếu tố sau:

i) Đánh giá của mỗi luật mờ, $Fr(.)$, theo trọng tâm của các mẫu dữ liệu được phân lớp bởi luật đó. Được tính theo độ thuộc của trọng tâm các mẫu dữ liệu vào luật đó, công thức (2),

$$F_r(r) = \prod_{j=1...N} sm(c_j, A_j), \tag{0.9}$$

trong đó, c_j là tâm của các mẫu dữ liệu được phân lớp bởi luật r, A_j là hạng từ HA của biến vào X_j trong phần điều kiện của luật r. Với hệ có M luật, khi đó đánh giá độ phù hợp của tập luật là

$$F_r = \frac{\sum_{i=1...M} F_r(r_i)}{M}. \tag{0.10}$$

ii) Tổng sai số của tập các luật đối với tập dữ liệu mẫu kiểm tra, $E_s(.)$,

$$E_s = Numberoferrors(misclassifieds). \tag{0.11}$$

iii) Kích thước của tập luật, $F_m(.)$,

$$F_m(M) = M. \tag{0.12}$$

Vậy hàm đánh giá của cá thể sẽ là

$$FITNESS(individual) = \frac{w_0}{w_1.E_s + w_2.F_m} + (1 - w_0)F_r, \tag{0.13}$$

trong đó, w_1 là trọng số đánh giá sai số phân lớp của tập luật trên cả dữ liệu huấn luyện và dữ liệu kiểm tra, w_2 là trọng số cho kích thước tập luật, $w_1 + w_2 = 1, w_0$ là trọng số đánh giá của cả hai yếu tố trên so với yếu tố F_r.

Rõ ràng, theo (0.9), ta chọn bộ trọng số w_0, w_1, w_2 thích hợp để kết quả phân lớp càng chính xác và số luật sinh ra ngày càng bé (sẽ được thử nghiệm ở phần sau).

Các phép di truyền của GA đã được các tác giả trong [7] phân tích và đánh giá, họ đã sử dụng tham số mô phỏng nhiệt độ tôi luyện − T, tác động vào các phép trên nhằm cải thiện chất lượng của GA truyền thống, mở rộng thành giải thuật TGA. Trong bài báo này, tác giả đã sử dụng các phép di truyền như sau.

• Phép chọn lọc sử dụng sơ đồ chọn lọc xếp hạng không tuyến tính theo hàm số mũ: các cá thể được sắp xếp theo thứ tự giảm của hàm $FITNESS(.)$, cá thể thứ i (xếp hạng i) sẽ được chọn vào quần thể bố mẹ (parent) theo xác xuất

$$p_i = \frac{(1 - a).a^{-i}}{a^{-N} - 1}, \tag{0.14}$$

$$a = \frac{1 + \gamma(T_k)}{N}, \tag{0.15}$$

$$\gamma(T) = 1 + (\gamma_{max} - 1).\frac{ln(T_0) - ln(T_k)}{ln(T_0) - ln(T_{end})}, \tag{0.16}$$

trong đó N là số cá thể trong quần thể, $T_k = T_0.\alpha^k$ là nhiệt độ tôi luyện hiện tại (tức thế hệ thứ $k = 1, ..., G$) giảm từ nhiệt độ ban đầu T_0 đến $T_{end} = T_0.\alpha^G, 0 < \alpha < 1$ (thường chọn $\alpha = 0,7$), G là số thế hệ cần tiến hóa, hàm $\gamma(T)$ sẽ tăng tuyến tính theo số thế hệ đã tiến hóa từ 1 đến γ_{max}, thường chọn $\gamma_{max} = 10$.

• Phép lai ghép được chọn ngẫu nhiên theo phân bố đều một trong ba phép lai ghép một điểm cắt, lai ghép tuyến tính và lai ghép tuyến tính mở rộng, cụ thể như sau:

*) Lai ghép một điểm cắt: Chọn một vị trí ngẫu nhiên i phân bố đều theo chiều dài chuỗi gen của hai cá thể bố mẹ X, Y, hai cá thể con U, V sinh ra bằng cách lấy phần đầu của cá thể X ghép với phần sau Y và ngược lại, $U = Left(X, i)+Right(Y, i), V = Left(Y, i)+Right(X, i)$.

*) Lai ghép tuyến tính: Chọn a ngẫu nhiên phân bố đều trong khoảng (0,1), từ hai cá thể bố mẹ X, Y sinh hai cá thể con U, V như sau: $U = a.X + (1-a).Y, V = (1-a).X + a.Y$.

*) Lai ghép tuyến tính mở rộng: Chọn một điểm cắt i chia chuỗi gen hai cá thể bố mẹ X, Y thành 4 phần $X_L = Left(X,i), X_R = Right(X,i), Y_L = Left(Y,i), Y_R = Right(Y,i)$, khi đó thực hiện phép lại ghép tuyến tính trên hai nửa của bố mẹ như sau

$$U_L = a.X_L + (1-a).Y_L,\ U_R = a.X_R + (1-a).Y_R, V_L = (1-a).X_L + a.Y_L, V_R = (1-a).X_R + a.Y_R,$$

với a chọn ngẫu nhiên phân phối đều trong đoạn $[\frac{-f2}{f1+f2}, 1 + \frac{f1}{f1+f2}], f1, f2$ là độ phù hợp (FITNESS) của hai cá thể bố mẹ.

• Phép đột biến được tính như sau. Giả sử một gen có giá trị x_i nằm trong khoảng giới hạn giá trị là $[L_i, U_i]$ được chọn để đột biến, giá trị gen sau đột biến là

$$x_i' = \begin{cases} x_i + z.(x_i - L_i) & u < 0,5, \\ x_i + z.(U_i - x_i) & else, \end{cases} \tag{0.17}$$

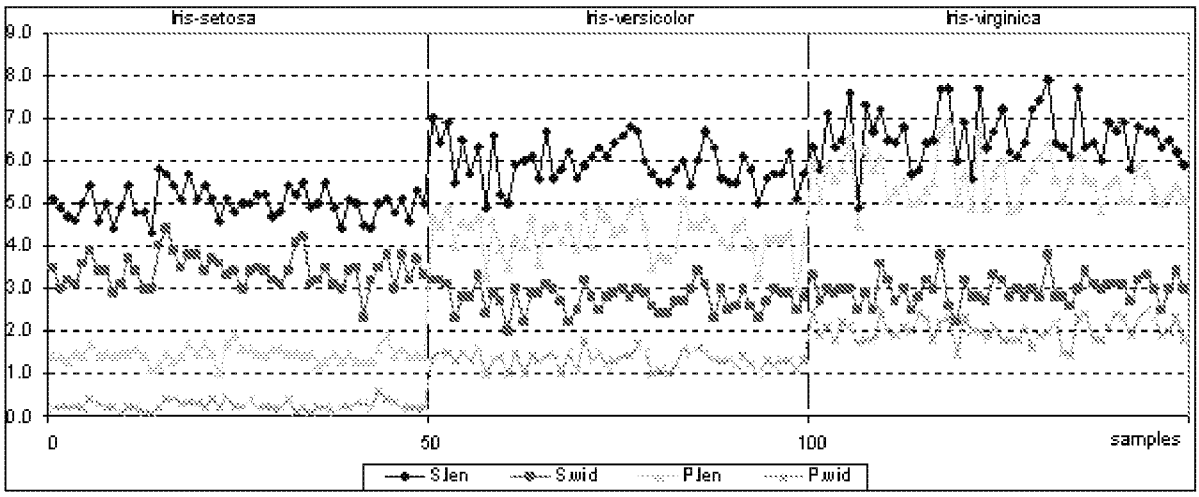
trong đó, u chọn ngẫu nhiên trong đoạn $[0,1]$, $z = sign(u-0.5).T_k.(1 + \frac{1}{T_k})^{|2u-1|} - 1$, T_k là nhiệt độ tôi luyện tại thế hệ thứ k.

• Phép tái tạo là cách thay thế bố mẹ bằng cá thể con, mỗi cá thể con cạnh tranh với cá thể tốt nhất trong hai cá thể bố mẹ. Gọi g_{p_1}, g_{p_2}, g_c tương ứng là giá trị hàm mục tiêu của hai cá thể bố mẹ và cá thể con, $g^* = min\{g_{p_1}, g_{p_2}\}$, khi đó cá thể con được chấp nhận với xác suất $p = min\{1, e^{-\frac{g_c - g^*}{T_k}}\}$. Trong trường hợp cá thể con không được chấp nhận, cá thể bố mẹ tương ứng với g^* được chấp nhận.

Phần tiếp theo sẽ là ứng dụng phương pháp này vào bài toán phân lớp các loài hoa (IRIS dataset).

4. ÁP DỤNG THỬ NGHIỆM VÀO BÀI TOÁN PHÂN LỚP CÁC LOÀI HOA *IRIS*

Bài toán phân lớp IRIS được các tác giả sử dụng rất phổ biến [8−18] và đã được công bố tại [19], tập dữ liệu mẫu có 4 thuộc tính vào (sepal−length, sepal−width, petal−length, petal−width), kích thước 150 mẫu chia đều cho 3 lớp (setosa, versicolor, virginica), mỗi lớp 50 mẫu dữ liệu. Phân bố dữ liệu của các thuộc tính vào trên tập dữ liệu mẫu theo lớp thể hiện trong Hình 4.1 và Bảng 4.1, trực quan ta thấy lớp setosa được phân biệt rõ bởi hai thuộc tính petal−length và petal−width phân bố trong khoảng $[1,\ 1,9]$ và $[0,1,\ 0,6]$ tương ứng. Thuộc tính petal−length phân lớp versicolor (trong khoảng $[3,\ 5,1]$), virginica (trong khoảng $[4,5,\ 6,9]$) rõ hơn so với petal−width, tương ứng trong khoảng $[1,\ 1,8]$ và $[1,4,\ 2,5]$. Hai thuộc tính sepal−length và sepal−width phân bố rộng trong miền $[4,3,\ 7,9]$ và $[2,\ 4,4]$, chúng phân bố chồng chéo lên nhau giữa các lớp dẫn đến vai tròtác động phân lớp của hai thuộc tính này không lớn.

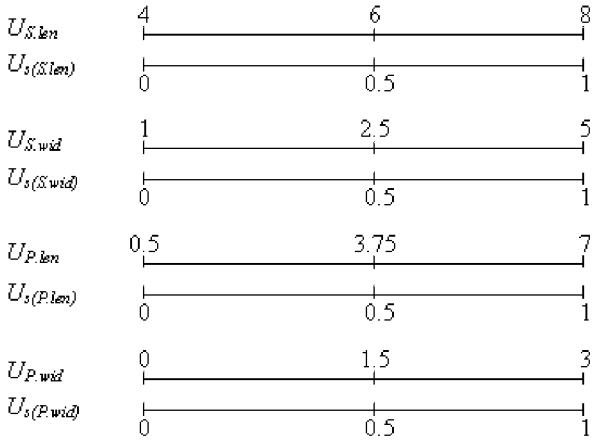


Hình 4.1. Đồ thị phân bố dữ liệu của 4 thuộc tính theo lớp trên tập dữ liệu mẫu

Bảng 4.1. Khoảng phân bố dữ liệu của các thuộc tính theo các lớp

Class name	Sepal−length	Sepal−width	Petal−length	Petal−width
Iris−setosa	[4,3, 5,8]	[2,3, 4,4]	[1, 1,9]	[0,1, 0,6]
Iris−versicolor	[4,9, 7]	[2, 3,4]	[3, 5,1]	[1, 1,8]
Iris−virginica	[4,9, 7,9]	[2,2, 3,8]	[4,5, 6,9]	[1,4, 2,5]

Ta sử dụng các ánh xạ tuyến tính $f_{u_j} : U_j \to [0,\ 1],\ (j = 1...4)$, tương ứng với các thuộc tính vào sepal−length, sepal−width, petal−length, petal−width, để chuyển dữ liệu từ miền thực về miền ngữ nghĩa định lượng (Hình 4.2).



Hình 4.2. Ánh xạ tuyến tính chuyển dữ liệu thực về miền ngữ nghĩa định lượng

Theo phương pháp tiếp cận của đại số gia tử (HAR), ta xây dựng các cấu trúc đại số gia tử tương ứng cho các biến vào, AX_j $(j = 1...4)$, với bộ các tham số gia tử là $\{fm_j(c^-),\ \mu_j(P),\ \mu_j(L),\ \mu_j(M),\ \mu_j(V)\}$, trong đó $H^- = \{L\}$, $H^+ = \{V\}$. Các biến vào X_1, X_2, X_3, X_4 tương

ứng là thuộc tính sepal−length, sepal−width, petal−length, petal−width. Như vậy, chúng ta có vectơ gồm 24 tham số của mỗi cá thể trong GA, chia làm 6 nhiễm sắc thể gồm

*) 01 nhiễm sắc thể tham số mờ của c^-, $chromosome_0 = (fm_1(c^-), fm_2(c^-), fm_3(c^-), fm_4(c^-))$, tương ứng của 4 thuộc tính $X_j, j = 1...4$.

*) 04 nhiễm sắc thể tham số mờ các gia tử P, L, M, V tương ứng là $chromosome_1 = \{\mu_1(P),...,\mu_4(P)\}$, $chromosome_2 = \{\mu_1(L), ..., \mu_4(L)\}$, $chromosome_3 = \{\mu_1(M), ..., \mu_4(M)\}$, $chromosome_4 = \{\mu_1(V), ..., \mu_4(V)\}$.

*) 01 nhiễm sắc thể tham số mức khoảng mờ của mỗi thuộc tính, $chromosome_5 = (k_1, k_2, k_3, k_4)$.

Ở đây, ta chọn 160 cá thể ngẫu nhiên trong quần thể xuất phát, số thế hệ tiến hóa là 20. Hàm $FITNESS(.)$ tính theo công thức (9) với bộ trọng số (w_0, w_1, w_2) được chọn khác nhau. Thứ nhất, tất cả các mẫu dữ liệu được sử dụng để huấn luyện (training) và cũng dùng để kiểm tra (testing), chạy thử nghiệm lần thứ nhất (E1) với bộ trọng số $FITNESS$ $w_{E1} = (w_0 = 0,6,\ w_1 = 0,3,\ w_2 = 0,7)$, lần thứ hai (E2) với bộ trọng số $w_{E2} = (0,9,\ 0,5,\ 0,5)$. kết quả thu được các tham số gia tử trong Bảng 4.1, số luật sinh ra trong trường hợp E1 là 5 với số lỗi phân lớp là 0/150 mẫu dữ liệu, trường hợp E2 có số luật là 4, số lỗi phân lớp 3/150 (Bảng 4.4). So với [9], số bước tiến hóa của phương pháp này giảm 80% (20/100), số luật trong E1 giảm 25% (6/8) và cả hai đều có kết quả phân lớp chính xác 100%. Trong khi, số luật ở E2 giảm đến 50% (4/8) nhưng kết quả phân lớp chính xác đạt 98%, chỉ giảm 2% so với [9].

Bảng 4.2. Kết quả các tham số trong hai trường hợp E1, E2

Thuộc tính	$fm(c^-)$	$\mu(P)$	$\mu(L)$	$\mu(M)$	$\mu(V)$	k
Trường hợp thử nghiệm E1						
S.len	0.010206	0.075309	0.168202	0.547296	0.209191	1
S.wid	0.069836	0.129965	0.515546	0.127315	0.227172	2
P.len	0.687520	0.022336	0.447423	0.134606	0.395633	2
P.wid	0.265453	0.115678	0.111313	0.009905	0.763102	3
Trường hợp thử nghiệm E2						
S.len	0.063385	0.005110	0.194986	0.135107	0.664796	1
S.wid	0.009477	0.320817	0.137353	0.145616	0.396213	1
P.len	0.946530	0.334347	0.271895	0.006914	0.386842	3
P.wid	0.007080	0.182840	0.195990	0.012080	0.609088	3

Bảng 4.3. Bảng FAM kết quả các luật mờ trong hai trường hợp E1, E2

Rules	Left of rules				Right of rules
	Sepal length	Sepal width	Petal length	Petal width	
Trường hợp thửnghiệm E1					
1	Large	Large	V Small	V Small	setosa
2	Large	Large	L Small	Large	versicolor
4	Large	L Large	L Large	PV Large	versicolor
3	Large	Large	Large	V Large	virginica
5	Large	L Large	L Small	V Large	virginica

Hai thuộc tính sepal−length, sepal−width có tham số mờ $fm(Small)$ (Bảng 4.2) rất nhỏ, đối chiếu với Bảng 4.3 thì các luật mờ sinh ra chỉ chứa các hạng từ sinh bởi Large và giống nhau ở các luật trên hai thuộc tính này. Giải thích rằng khi hai thuộc tính có các mẫu dữ liệu được phân bố chồng chéo giữa các lớp và rộng khắp trên miền định lượng nên tác động rất ít đến sự phân lớp, chúng ta có thể không cần đến thuộc tính này. Do vậy, chỉ lấy một khoảng mờ khá lớn phủ kín khoảng phân bố của các mẫu dữ liệu, tức là độ đo tính mờ của hạng từ cơ sở tương ứng khá lớn. Ngược lại, khi một thuộc tính (vídụ petal−length) có tham số mờ $fm(c^-)$ thực tế hơn (không quá nhỏvà không quá lớn) thì các luật mờ sinh ra chứa các hạng từ sinh bởi cả c^- và c^+. Cũng theo Bảng 4.3, các luật mờ sinh ra chứa các hạng từ khác nhau ở thuộc tính petal−length và petal−width, tức là hai thuộc tính này quyết định chủ yếu đến sự phân lớp. Điều này rất phù hợp với sự phân bố của tập dữ liệu mẫu như đã phân tích ở phần trên, khẳng định kết quả đúng đắn của phương pháp này.

Tiếp theo, ta sử dụng chiến lược leave−one−out. Mỗi lần thử nghiệm lấy ra một mẫu để kiểm tra mô hình, còn lại 149 mẫu dùng cho huấn luyện trích rút luật. Tiến hành chạy 150 lần thử nghiệm cho lần lượt mỗi mẫu dữ liệu được lấy ra kiểm thử, kết quả lỗi phân lớp nhỏ nhất là 0, lớn nhất 3 lỗi. Số luật sinh ra nhỏ nhất là 3, lớn nhất 8 luật cho mỗi lần thử nghiệm. Tính trung bình cộng kết quả của 150 lần thử nghiệm là $1,14$ lỗi/150 mẫu, đạt độ chính xác $99,24\%$, tăng đáng kể so với [9] là $98,67\%$ (2 lỗi/150 mẫu), hơn nữa số luật trung bình là $5,27$, giảm $2,4\%$ so với [9] là $5,4$ luật.

Đặc biệt, trong trường hợp thử nghiệm này cho kết quả rất tốt tại 40 lần thử nghiệm tương ứng với các mẫu dữ liệu được dùng để kiểm tra, các kết quả này đều sinh ra 5 luật mờ và số lỗi phân lớp 0/150. Bảng 4.3 thể hiện so sánh của phương pháp này với các kết quả trước (ký hiệu "\" là không xác định).

Bảng 4.4. Kết quả trung bình của leave−one−out và so sánh với các phương pháp trước

	Our method	Method [9]	Method [13]	Bayes method	Method [20]
Số luật(TB)	5.27	5.4	4.75	\	\
Kết quả	99.24%	98.67%	98%	97.33%	94.33%

Trường hợp thứ 3, sử dụng chiến lược k−cross−validation, các mẫu dữ liệu trong mỗi lớp được chia ngẫu nhiên theo tỷlệ 1:1, một nửa dành cho huấn luyện, nửa còn lại để kiểm tra. Chạy thử nghiệm 150 lần, tính kết quả trung bình cộng của 150 lần thử nghiệm đạt chính xác $98,02\%$ ($2,97$ lỗi/150 mẫu), tốt hơn so với [9, 16, 17] tương ứng là $95,5\%$, $97,84\%$ và $96,7\%$, số luật sinh ra bénhất là 3 luật và lớn nhất là 5 luật, tính trung bình là $3,83$ luật, giảm $19,03\%$ so với [17] ($3,83/4,73$), giảm $56,72\%$ so với [16] ($3,83/8,85$), tăng $8,62\%$ so với [19] ($3,83/3,5$). Kết quả tốt nhất trong 150 lần thử nghiệm là 3 luật mờ và đạt tỷ lệ chính xác $98,67\%$, tốt hơn so với [18] là 3 luật mờ và tỷ lệ chính xác $96,4\%$. Bảng sau thể hiện so sánh kết quả trung bình của phương pháp này với các phương pháp trước (ký hiệu "\" là không xác định).

Bảng 4.5. Kết quả trung bình của k−cross−validation so sánh với $[9, 14 - 17, 21]$

	Our method	[9]	[17]	[16]	[14]	[15]	[21]
số luật(TB)	3,89	3,5	4,73	8,85	\	\	\
kết quả& 98,03%	95,5%	97,84%	96,7%	96,21%	97,34%	98%	

5. KẾT LUẬN

Bài báo đã đề xuất một phương pháp mới giải quyết bài toán phân lớp mờ dựa trên tiếp cận đại số gia tử. Miền của các thuộc tính vào được cấu hóa dựa theo đại số của các hạng từ, phân hoạch miền này thành các khoảng mờ mức k, chọn hạng từ thích hợp cho mỗi khoảng mờ từ tập dữ liệu mẫu. Sau đó dùng Thuật toán 3.1 để học và trích rút các luật mờ, hơn nữa kỹ thuật rút gọn tập luật mờ cũng dựa trên cấu trúc đại số của miền thuộc tính vào bằng cách giảm tính rõ (hoặc tăng tính mờ) của một giá trị đầu vào trong luật. Cách làm này sẽ phải thỏa hiệp giữa sai số phân lớp với số luật sinh ra.

Kết quả cho thấy cách tiếp cận này chỉ phụ thuộc bộ tham số gia tử và chọn mức khoảng mờ cho mỗi thuộc tính vào, một giải pháp được hầu hết các tác giả sử dụng để tối ưu các tham số cho mô hình là giải thuật di truyền, sử dụng giải thuật di truyền có tích hợp thêm tham số nhiệt, phỏng theo nhiệt độ tôi luyện thép [7], nhằm tăng tốc độ hội tụ của giải thuật.

Khác với các phương pháp tiếp cận hệ mờ theo giải tích, phương pháp đại số gia tử này dựa trên sự phân hoạch các khoảng mờ của miền các thuộc tính, đạt được hai ưu điểm nổi bật sau. Thứ nhất, mỗi khoảng mờ được xác định bởi một hạng từ, khai thác đặc tính kế thừa của các hạng từ (theo Định đề 2.1. (vi)) đi xây dựng giải thuật sinh luật mờ (Thuật toán 3.1) chủ yếu xử lý trên các hạng từ dưới dạng cấu trúc, giảm bớt các quá trình tính toán theo giải tích, dẫn đến tốc độ nhanh hơn. Thứ hai, kết quả phân lớp của phương pháp này luôn đạt độ chính xác cao vì theo tính phân hoạch của khoảng mờ, mỗi dữ liệu trong tập mẫu chỉ thuộc vào một khoảng mờ nhất định và được xử lý phân lớp theo khoảng mờ đó. Khác với cách tiếp cận giải tích, mỗi dữ liệu có thể thuộc vào hai tập mờ vì các tập mờ có giao nhau, dẫn đến dữ liệu này sẽ được xem xét phân lớp theo hai tập mờ.

Kết quả áp dụng vào bài toán phân lớp các loài hoa (iris) cho thấy phương pháp này tốt hơn kết quả của các tác giả [9, 13-18, 20]. Đặc biệt tốc độ hội tụ nhanh hơn, các trường hợp thử nghiệm chỉ cần tiến hóa qua 20 thế hệ, giảm 80% so với [9] (20/100).

Tuy nhiên, có thể thấy rằng quá trình phân hoạch miền các thuộc tính vào thành các khoảng mờ sẽ còn phụ thuộc vào một yếu tố nữa đó là số các gia tử tham gia trong đại số tương ứng. Rõ ràng, nếu càng nhiều gia tử thì mỗi mức khoảng mờ được phân hoạch thành nhiều khoảng mờ hơn, tăng tính mềm dẻo trong quá trình áp dụng. Nhưng khi đó chúng ta phải thỏa hiệp với tốc độ tụ, tăng số lượng gia tử lên cũng có nghĩa số lượng tham số mờ của gia tử cũng tăng lên và không gian tìm kiếm trong giải thuật di tuyền sẽ lớn hơn, dẫn đến tốc độ hội tụ ngày càng giảm.

TÀI LIỆU THAM KHẢO

[1] Ho N.C, A topological completion of refined hedge algebras and a model of fuzziness of linguistic terms and hedges, *Fuzzy Sets and Systems* **158** (2007) 436 – 451.

[2] Ho N.C and Lan V.N, Hedge algebras: An algebraic approach to domains of linguistic varibles and their applicability, *AJSTD* **23** (Issues 1&2) (2006) 1–18.

[3] Ho N.C, Long N.V, Fuzziness measure on complete hedge algebras and quantifying semantics of terms in linear hedge algebras, *Fuzzy Sets and Systems* **158** (2007) 452 – 471.

[4] Ho N.C, Lan V.N, and Viet L.X, Optimal hedge-algebras-based controller: Design and application, *Fuzzy Sets and Systems* **159** (2008) 968– 989.

[5] Ho N. C, CSDL mờ với ngữ nghĩa đại số gia tử "Lectures on the Fuzzy Systems & Applications Autumn School", 9/2008.

[6] Ho N.C, Son T.T, Long D.T, "Một mô hình lập luận mờ dựa trên cấu trúc đại số của các biến ngôn ngừ, mạng nơron và giải thuật di truyền", Semina về đại số gia tử và ứng dụng, Viện Công nghệ thông tin, 5/2009.

[7] Trần Ngọc Hà, "Các hệ thống thông minh lai và ứng dụng trong xử lý dữ liệu, Luận án tiến sỹ", Đại học Bách khoa Hà Nội, 2002.

[8] Johannes A. Roubos, Magne Setnes, Janos Abonyi, Learning fuzzy classification rules from labeled data, *Information Sciences* **150** (2003) 77–93.

[9] Enwang Zhou, Alireza Khotanzad, Fuzzy classifier design using genetic algorithms, *Pattern Recognition section* **40** (12) (December 2007) Elsevier Science (3401–3414).

[10] Diyar Akay, M. Ali Akcayol, Mustafa Kurt, NEFCLASS based extraction of fuzzy rules and classification of risks of low back disorders, *Expert Systems with Applications* **35** (2008) 2107–2112.

[11] R. Kruse, "Design and Implementation of a Neuro-Fuzzy Data Analysis Tool in Java", Institute of Operating Systems and Computer Networks - Technical University Braunschweig, 1999.

[12] H. Ishibuchi and T. Nakashima, Voting in fuzzy rule-based systems for pattern classification problems, *Fuzzy Set Syst.* **103** (2) (1999) 223–238.

[13] X.G. Chang and J.H. Lilly, Evolutionary design of a fuzzy classifier from data, *IEEE Trans. Systems, Man, and Cybernetics*, part B **34** (4) (2004) 1894–1906.

[14] T.P. Wu and S.M. Chen, A new method for constructing membership functions and fuzzy rules form training examples, *IEEE Trans. System, Man, and Cybernetics*, part B **29** (1) (1999) 25–40.

[15] J.S. Wang and G.C.S. Lee, Self–adaptive neuro-fuzzy inference system for classification application, *IEEE Trans. Fuzzy Systems* **10** (6) (2002) 790–802.

[16] Yung-Chou Chena, Li-HuiWangb, and Shyi-Ming Chenc, Generating weighted fuzzy rules from training data for dealing with the iris data classification problem, *International Journal of Applied Science and Engineering* **22** (1) (2006) 175–188.

[17] Chia-Chong Chen, Design of PSO-based Fuzzy Classification Systems, *Tamkang Journal of Science and Engineering* **9** (1) (2006) 63–70.

[18] Hisao Ishibuchi* and Takashi Yamamoto, "Fuzzy rule selection by multi-objective genetic local search algorithms and rule evaluation measures in data mining", Department of Industrial Engineering, Osaka Prefecture University, 1-1 Gakuen-cho, Sakai, Osaka, Japan (599–8531).

[19] UCI machine learning repository via an anonymous ftp server at the address, ftp://ftp.ics.uci.edu/pub/machine-learning-databases/iris.

[20] M. Grabisch and F. Dispot, A comparison of some methods of fuzzy classification on real data, *Proc. of IIZUKA 92*, Iizuka, Japan, Jul., 1992 (659–662).

[21] Cheng-Jian Lin, Chi-Yung Lee, and Shang-Jin Hong, An efficient fuzzy classifier based on hierarchical huzzy entropy, *International Journal of Information Technology* **12** (6) (2006).

Nhận bài ngày 9 - 3 - 2009
Nhận lại sau sửa ngày 26 - 8 -2009