Tạp chí Tin học và Điều khiển học, T.25, S.1 (2009), 55–71

MỘT PHƯƠNG PHÁP XÂY DỰNG HỆ LUẬT MỜ CÓ TRỌNG SỐ ĐỂ PHÂN LỚP DỰA TRÊN ĐẠI SỐ GIA TỬ

DƯƠNG THĂNG LONG

Khoa Công nghệ Tin học, Viện Đại học mở Hà nội

Abstract. In this paper, we propose a method for designing a classification system, which uses weighted fuzzy rules base on hedge algerbas (HA). The antecedents of fuzzy rules are generated from datasets base on the partitioning of fuzziness intervals of terms. Rule weights and reasoning methods are defined base on these intervals. Next, we apply genetic algorithms (GA) for selecting rule sets with high performance and small average length of the rules, then we apply GA to optimize fuzzy parameters of hedges. We experiment our method to a popular classification problem (wine, [24]), and getting good results in comparing with [13,14].

Tóm tắt. Bài báo này đề xuất một phương pháp xây dựng hệ luật mờ có trọng số cho bài toán phân lớp $[12-14]$ dựa trên đại số gia tử (HA) $[1-2, 10, 15-17]$. Phần điều kiện của các luật mờ được sinh từ dữ liệu mẫu dựa trên phân hoạch các khoảng tính mờ theo HA của các thuộc tính, các tính toán trọng số luật và lập luận phân lớp cũng dựa trên khoảng tính mờ của các hạng từ tương ứng [2]. Các tham số mờ gia tử áp dụng cho mô hình được tối ưu theo giải thuật di truyền, hơn nữa tập luật sinh ra từ dữ liệu mẫu của bài toán khá lớn, chúng ta chọn tập tối ưu các luật cũng bằng giải thuật di truyền $[2, 3, 13, 25]$. Mô hình sẽ được thử nghiệm cho bài toán phân lớp các loại rượu (wine, [24]) phổ biến rộng rãi và nhiều tác giả sử dụng $[8, 9, 11-14]$, cho kết quả tốt so với $[13, 14]$.

1. GIỚI THIỆU

Hệ mờ được nghiên cứu phát triển và ứng dụng rộng rãi cho bài toán phân lớp $[6-9, 11-14, 18-23, 25-27]$, nhiều tác giả sử dụng hệ các luật mờ dạng IF antecedents THEN consequent $[2, 7, 9, 11-14, 19, 20, 23]$. Đặc biệt, các tác giả trong $[8, 12, 14, 27]$ đã cho thấy mỗi luật dựa trên tập dữ liệu mẫu được đánh giá mức độ tác động đến hiệu quả phân lớp là khác nhau. Như vậy, mỗi luật sẽ mang theo nó một trọng số (weight) đánh giá độ quan trọng trong lập luận. Luật mờ có trọng số thường được biểu diễn như sau:

$$R_q : IF\ X_1\ is\ A_{q,1}\ AND\ X_2\ is\ A_{q,2}\ AND ... AND\ X_N\ is\ A_{q,N}\ THEN\ C_q with CF_q, \quad (0.1)$$

trong đó, $X = (X_1, ..., X_N)$ là véctơ dữ liệu mẫu kích thước N (N thuộc tính), $A_{q,j}$ là một hạng từ (linguistic term) của phần điều kiện luật tương ứng với thuộc tính j, C_q là nhãn phân lớp đầu ra của luật, và CF_q là trọng số của luật (weight).

*Nghiên cứu được hoàn thành với sự hỗ trợ từ quỹ NAFOSTED

Bài toán phân lớp thường được cho bởi một tập dữ liệu mẫu, ký hiệu $P = \{p_i = (d_{i1}, ..., d_{iN}; C_i) \mid p_i \in D, i = 1, ..., M\}$, D là tập dữ liệu, $C_i \in C_1, ..., C_m$ là nhãn phân lớp cho dữ liệu p_i, M là số mẫu, m là số lớp. Dựa trên tập mẫu này, chúng ta xây dựng một hệ luật mờ có trọng số theo dạng (1) để phân lớp. Với mỗi mẫu dữ liệu $p = (d_1, ..., d_n)$ được phân lớp vào một $C' \in C_1, ..., C_m$ theo một phương pháp lập luận phân lớp được chọn trên hệ luật. Yêu cầu đặt ra đối với hệ luật này là hiệu quả phân lớp càng cao càng tốt và hệ luật phải nhỏ gọn, dễ hiểu đối với người dùng.

Với những nghiên cứu về hệ phân lớp dựa trên các luật mờ không có trọng số (tức là $CF_q = 1, \forall q$), các tác giả thường điều chỉnh tham số mờ bằng các kỹ thuật học máy như giải thuật di truyền [2, 9, 19, 26] hoặc kết hợp mạng nơron [18, 22]. Tuy nhiên, điều này có thể làm mất tính trực quan và dễ hiểu của các tập mờ và các luật mờ [14]. Một số tác giả khắc phục bằng cách đưa vào các ràng buộc về việc điều chỉnh các tham số mờ [14, 9]. Trong [14] sử dụng trọng số luật thay vì điều chỉnh tham số mờ để làm tăng độ chính xác của hệ, đảm bảo tính trực quan của luật. Rõ ràng, chúng ta phải thỏa hiệp giữa hai yếu tố hiệu quả phân lớp và tính trực quan của các luật mờ cũng như kích thước tập luật. Bài báo này đề xuất phương pháp xây dựng hệ luật có trọng số dựa trên đại số gia tử nhằm tăng hiệu quả phân lớp nhưng vẫn đảm bảo tính trực quan của các luật bằng những ràng buộc về tham số mờ gia tử.

Bài báo được trình bày gồm 5 mục. Trong Mục 2 giới thiệu luật mờ dựa trên đại số gia tử, các đánh giá trọng số của luật [11 − 14] và phương pháp lập luận cho bài toán phân lớp [14]. Mục 3 đề xuất mô hình phân lớp mờ, trong đó phân tích ảnh hưởng của tham số θ trong việc kết nhập các luật mờ [3] và phương pháp sàng (lựa chọn) các luật theo các tiêu chuẩn khác nhau [11 − 14]. Mục 4 áp dụng thử nghiệm cho mô hình đến bài toán phân lớp các loại rượu (wine), được công bố trong [24] và nhiều tác giả sử dụng. Kết quả của mô hình rất tốt so với [2, 9, 14]. Cuối cùng là phần kết luận.

2. LUẬT MỜ CÓ TRỌNG SỐ DỰA TRÊN ĐẠI SỐ GIA TỬ VÀ PHƯƠNG PHÁP LẬP LUẬN

Theo tiếp cận ĐSGT, hệ luật mờ dạng (0.1) có phần điều kiện ở vế trái là các hạng từ trong ĐSGT. Trong đó, miền của mỗi thuộc tính sẽ được phân hoạch mờ bởi một tập các khái niệm mờ tương ứng là các khoảng tính mờ tại mức k cho trước. Phương pháp phân hoạch này đã được nhiều tác giả sử dụng cho quá trình sinh luật [4 − 8, 11 − 14], trong đó các khái niệm mờ được đặt cố định cả về số lượng lẫn hàm thuộc tương ứng. Chúng tôi sử dụng phân hoạch dựa trên phân hoạch hệ các khoảng tính mờ của ĐSGT sẽ khắc phục được các nhược điểm này (Hình 3.1).

Một số vấn đề liên quan đến ĐSGT

Xét đại số gia tử $AX = (X, G, C, H, \Sigma, \Phi, \leq)$ [2]. Trong đó X là tập các hạng từ (terms) của đại số, $G = c^-, c^+$ là tập phần tử sinh, $C = 0, 1, W$ các giá trị hằng, H là tập các gia tử

(hedges),Φ gia tử tới hạn min, Σ gia tử tới hạn max, ≤ quan hệ thứ tự ngữ nghĩa giữa các hạng từ. Ta ký hiệu $X_k \in X$ là tập các hạng từ có độ dài đúng k, $I_k = \{\Im_k(x) \mid x \in X_k\}$ là tập các khoảng tính mờ mức k (k-intervals) xác định bởi các hạng từ trong X_k [1, 2]. Dựa trên tính dàn của các hạng từ trong đại số và tính phân hoạch của các khoảng tính mờ, định nghĩa về tính kế thừa của các khoảng tính mờ như sau:

Định nghĩa 2.1. [2] Với $\forall x, y \in X$ xác định hai khoảng tính mờ $\Im_p(x) \in I_p$ và $\Im_q(y) \in I_q$, chúng có quan hệ kế thừa (ký hiệu $\Im_p(x)\Im_q(y)$) nếu $\exists \Im_v(z) \in I_v, v \leq min(p, q), \Im_v(z) \supset \Im_p(x)$ và $\Im_v(z) \supset \Im_q(y)$, tức x, y được sinh ra từ $z, x = h_{kn}...h_{k1}z, y = h_{kn}...h_{k1}z, \forall h_{ki}, h'_{ki} \in H$.

Khi $\Im(x)$ và $\Im(y)$ có quan hệ kế thừa, ta nói rằng $\Im(z)$ bao hàm hai khoảng tính mờ trên. Để ý rằng, 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 $h_jx : 1 \leq j \leq p$, hơn nữa ta có.

$$\Sigma_{i\in[-q,-1]}|\Im(h_ix)| = fm(x) \times \Sigma_{i\in[-q,-1]}\mu(h_i) = \alpha.fm(x) = \alpha|\Im(x)|. \tag{0.2}$$

Điều này gợi ý chọn điểm cuối chung của hai khoảng tính mờ $\Im(h_{-1}x)$ và $\Im(h_{+1}x)$ là giá trị định lượng ngữ nghĩa u(x) (xem [17]) của hạng từ x.

Rõ ràng, giá trị định lượng ngữ nghĩa u của một hạng từ cũng như khoảng tính mờ của nó phụ thuộc đầy đủ vào các tham số mờ $fm(c^-), fm(c^+), \mu(h) \forall h \in H$.

Định nghĩa 2.2. Với $\forall x, y \in X$ xác định hai khoảng tính mờ $\Im_p(x) \in I_p$ và $\Im_q(y) \in I_q$, nếu chúng có quan hệ kế thừa thì $\Im_{v^*}(z^*) = argmax_v\{\Im_v(z) \in I_v | \Im_v(z) \supset \Im_p(x)$ và $\Im_v(z) \supset \Im_q(y)\}$, gọi là khoảng tính mờ bao hàm cực tiểu hai khoảng tính mờ trên. Ngược lại, khi chúng không có quan hệ kế thừa thì $v^* = 0$.

Thủ tục 2.1. [2] Tính khoảng tính mờ bao hàm nhỏ nhất (cực tiểu) hai khoảng tính mờ cho trước

Input: Hai khoảng tính mờ bất kỳ $\Im_p(x), \Im_q(y)$, và tập tất cả các phân hoạch $I_k(k = 1, 2, ..., k_{max})$.

Output: $\Im_v(z)$ bao hàm hai khoảng tính mờ $\Im_p(x), \Im_q(y)$ khi chúng có quan hệ kế thừa.

Actions:

Step 1) Đặt $v = min(p, q)$,

Step 2) Nếu $\exists \Im_v(z) \in I_v, \Im_v(z) \supset \Im_p(x)$ và $\Im_v(z) \supset \Im_q(y)$ thì kết quả là $\Im_v(z)$, với z là hạng từ tương ứng khoảng tính 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 tính mờ $\Im_p(x)$ và $\Im_q(y)$ không có quan hệ kế thừa, nếu $v > 0$ lặp lại Step2).

Thủ tục này được dùng để tính toán hạng từ kết nhập khi hợp các luật mờ với nhau (trình bày trong Thuật toán 3.1). Định nghĩa sau đánh giá mức độ gần nhau của hai hạng từ, và từ đó ta định nghĩa mức độ gần nhau giữa hai luật mờ trong Định nghĩa 2.4.

Định nghĩa 2.3. Cho $AX, \forall x, y \in X$ xác định hai khoảng tính mờ $\Im_k(x), \Im_l(y)$ (trong đó k là độ dài của x, l là độ dài của y), độ kết nhập của x và y dựa trên khoảng tính mờ bao hàm cực tiểu $\Im_v(z)$ và giá trị định lượng ngữ nghĩa của chúng là ánh xạ $sm : X \times X \to [0, 1]$ được xác định như sau:

$$sm(x, y) = \frac{v}{max(k, l)}(1 - |v(x) - v(y)|) \tag{0.3}$$

Định đề 2.1. Độ kết nhập sm được định nghĩa như trên (Định nghĩa 2.3), ta có:

i) Hàm sm là đối xứng, $sm(x, y) = sm(y, x)$.

ii) Nếu x, y không có quan hệ kế thừa thì $sm(x, y) = 0$.

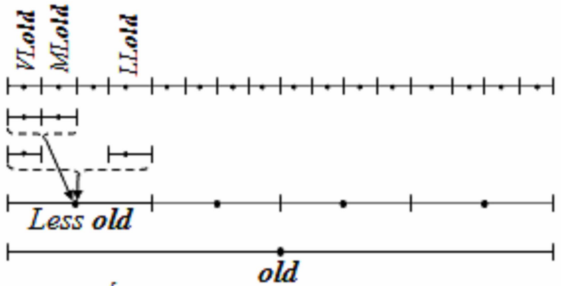
iii) $sm(x, y) = 1$ khi và chỉ khi $x = y$.

iv) $c, c' \in G, c \neq c', x \in H(c), y \in H(c') \Leftrightarrow sm(x, y) = 0$, và $c \in G, x, y \in H(c) \Leftrightarrow sm(x, y) > 0$.

v) $\forall x, y, z \in X_k, x \leq y \leq z \Rightarrow sm(x, z) \leq sm(x, y)$ và $sm(x, z) \leq sm(y, z)$.

Khi x, y cùng kế thừa ngữ nghĩa từ z (tức z bao hàm x, y) thì có thể sử dụng z đại diện cho cả x và y, tuy nhiên giá trị $max(k, l) - v$ càng lớn càng làm suy giảm ngữ nghĩa của x, y. Ví dụ, $x = very\ less\ old$ và $y = more\ less\ old$ đều kế thừa ngữ nghĩa của $z = less\ old$. Đây được gọi là sự mở rộng ngữ nghĩa, hay là sự tăng tính mờ của các hạng từ.

Sử dụng độ đo kết nhập trong Định nghĩa 2.3 để kết nhập hai hạng từ thành một hạng từ mới trong các ứng dụng. Rõ ràng, giá trị sm càng lớn thì việc kết nhập càng không làm mất ngữ nghĩa của các hạng từ. Ví dụ, $x = very\ less\ old, y = more\ less\ old, z = less\ less\ old$ đều kế thừa ngữ nghĩa của $w = less\ old$, tuy nhiên $sm(x, y) > sm(x, z)$ (Định đề 2.1. (v)), việc kết nhập x, y thành w tốt hơn x, z thành w (Hình 2.1).



Hình 2.1. Minh họa kết nhập hai hạng từ thành một hạng từ đại diện

Theo tiếp cận ĐSGT, hệ các luật mờ có trọng số theo dạng (1) có phần điều kiện (antecedents) ở vế trái của mỗi luật là các hạng từ trong ĐSGT và chúng xác định một tập các khoảng tính mờ $\{\Im(x_{q,1}), ..., \Im(x_{q,n})\}$. Ta định nghĩa độ kết nhập của hai luật làm cơ sở để kết nhập chúng thành một luật mới như sau:

Định nghĩa 2.4. Cho tập luật R dạng (1), độ kết nhập của hai luật $r_p, r_q \in R$ là hàm $itg : R \times R \to [0,1]$, được xác định dựa trên độ tương hợp của các thành phần như sau:

$$itg(r_p, r_q) = T_{ex}(sm(A_{p,1}, A_{q,1}), ..., sm(A_{p,n}, A_{q,n})), \tag{0.4}$$

trong đó, T_{ex} là một t-norm mở rộng, trong bài này T_{ex} được chọn là toán tử min. Hàm sm tính theo Định nghĩa 2.3.

Định nghĩa 2.5. Cho $AX, \forall x \in X$, định nghĩa độ tương hợp của giá trị định lượng v đối với hạng từ x là một ánh xạ $\lambda : [0,1] \times X \to [0,1]$, được xác định dựa trên khoảng cách từ v đến $\upsilon(x)$ và khoảng tính mờ $\Im(x)$ như sau:

i) Với $v \in \Im(x), \lambda(v,x) = (1 + \frac{(1-\rho).(v-\upsilon(x))^2}{(\rho.\sigma(v)^2)})^{-1}$,

ii) Ngược lại (tức là $v \notin \Im(x)$), $\lambda(v,x) = 0$,

trong đó, $\sigma(v)$ là khoảng cách từ tâm (giá trị định lượng ngữ nghĩa của x) đến một trong hai đầu mút của khoảng tính mờ $\Im(x)$ tính theo v, ρ là độ tương hợp tại hai điểm đầu mút (thường chọn 0,3), với $Left(\Im)$ và $Right(\Im)$ là điểm mút trái và điểm mút phải của khoảng tính mờ $\Im(x)$.

$$\sigma(\nu) = \begin{cases} \upsilon(x) - Left(\Im), & \nu < \upsilon(x) \\ \upsilon(x) - Right(\Im), & \nu \geq \upsilon(x) \end{cases} \tag{0.5}$$

Hàm λ đóng vai trò như một hàm định lượng ngữ nghĩa (hàm thuộc) cho giá trị ngôn ngữ x, do đó chúng tôi muốn rằng độ thuộc tại hai đầu mút của khoảng tính mờ là khác 0 và đặt bằng ρ.

Tiếp theo chúng ta trình bày các phương pháp đánh giá trọng số của luật mờ theo dạng (0.1). Áp dụng độ hỗ trợ và độ tin cậy (support, confidence) trong khai phá luật kết hợp. Luật mờ dạng (0.1) cũng được xem như một luật kết hợp $A_q \Rightarrow C_q$, trong đó $A_q = (A_{q,1}, ..., A_{q,N})$ là phần điều kiện luật, mỗi điều kiện $A_{q,j}$ trong vế trái luật xác định một khoảng tính mờ tương ứng $\Im(A_{q,1}), ..., \Im(A_{q,N})$. Hai độ đo này được tính như sau:

$$s(A_q \Rightarrow C_q) = \frac{\sum_{p_i \in classC_q} \mu_{A_q}(p_i)}{M}, \tag{0.6}$$

$$c(A_q \Rightarrow C_q) = \frac{\sum_{p_i \in classC_q} \mu_{A_q}(p_i)}{\sum_{i=1}^{M} \mu_{A_q}(p_i)}, \tag{0.7}$$

trong đó M là số mẫu, C_i là nhãn phân lớp cho mẫu pi, m là số lớp. Mức độ đáp ứng đầu vào của mẫu p_i đối với luật $A_q \Rightarrow C_q$ được tính như sau:

$$\mu_{A_q}(p_i) = \lambda(d_{i1}, A_{q1}) \times ... \times \lambda(d_{iN}, A_{qN}), \tag{0.8}$$

với hàm $\lambda(.,.)$ được xác định theo Định nghĩa 2.5.

Ta sẽ áp dụng một trong các phương pháp đánh giá trọng số luật được đề xuất trong [14]:

$$CF_q^I = c(A_q \Rightarrow C_q), \tag{0.9}$$

$$CF_q^{II} = c(A_q \Rightarrow C_q) - c_{Ave}, \tag{0.10}$$

$$CF_q^{III} = c(A_q \Rightarrow C_q) - c_{2nd}, \tag{0.11}$$

trong đó chỉ số I, II, III thể hiện từng định nghĩa của trọng số luật, c_{Ave} là trọng số trung bình trên các luật có cùng điều kiện A_q nhưng phần kết luận là các lớp còn lại khác C_q:

$$c_{Ave} = \frac{1}{m-1} \sum_{h=1, class_h \neq C_q}^{m} c(A_q \Rightarrow C_q), \tag{0.12}$$

c_{2nd} là trọng số cao nhất trong số các luật có điều kiện A_q và kết luận là các lớp còn lại:

$$c_{2nd} = max\{c(A_q \Rightarrow C_q) | h = 1, 2, ..., m; class_h \neq C_q\}, \tag{0.13}$$

$$CF_q^{IV} = 1, \forall q, \tag{0.14}$$

tức là luật không có trọng số.

Phương pháp lập luận đầu ra để phân lớp đối với một dữ liệu đầu vào p dựa trên hệ luật mờ S dạng (0.1), chúng ta có hai phương pháp đó là single winner rule ($Classify^I$) và weighted vote ($Classify^{II}$). Công thức xác định hai phương pháp này như sau [14]:

$$Classify^I(p) = argmax_{C_q}\{\mu_{A_q}(p).CF_q | A_q \Rightarrow C_q \in S\}, \tag{0.15}$$

$$Classify^{II}(p) = argmax_{class_h}\{V_{class_h}(p) | h = 1, 2, ..., m\}, \tag{0.16}$$

trong đó $V_{class_h}(p)$ là giá trị *vote* của mỗi lớp h đối với mẫu p, được tính là:

$$V_{class_h}(x_p) = \sum_{R_q \in S, C_q = h} \mu_{A_q}(x_p).CF_q. \tag{0.17}$$

3. XÂY DỰNG HỆ LUẬT MỜ CHO BÀI TOÁN PHÂN LỚP

3.1. Thuật toán sinh luật từ tập dữ liệu mẫu

Cho tập dữ liệu mẫu $P = p_i = (d_{i1}, ..., d_{iN}; C_i) | i = 1, ..., M$. Theo tiếp cận ĐSGT, miền của mỗi thuộc tính X_j sẽ được phân hoạch bởi hệ các khoảng tính mờ

$$I_{kj} = \{\Im_{k_j}(x_{k_j,1}), \Im_{k_j}(x_{k_j,2}), ..., \Im_{k_j}(x_{k_j,|X_{k_j}|})\}$$

tương ứng với tập hạng từ $X_{kj} = \{x_{k_j,1}, x_{k_j,2}, ..., x_{k_j,|X_{kj}|}\}$ của một ĐSGT AX_j, k_j là mức phân hoạch khoảng tính mờ của thuộc tính thứ $j, j = 1, 2, ..., N$.

Như vậy, các hệ I_{k_j} của N thuộc tính tạo nên một không gian các siêu hộp $\{B = (\Im_{k1}(x_{k1,i1}), \Im_{k2}(x_{k2,i2}), ..., \Im_{kN}(x_{kN,iN}))|\Im_{kj}(x_{kj,ij}) \in I_{kj}\}$ và $1 \leq i_j \leq |X_{kj}|$. Quá trình sinh luật sẽ xuất phát từ các siêu hộp có chứa mẫu dữ liệu, mỗi siêu hộp như vậy xem xét sinh một luật mờ gồm điều kiện vế trái là các hạng từ tương ứng với khoảng tính mờ trong siêu hộp, tức là $A_q = (x_{k1,i1}, x_{k2,i2}, ..., x_{kN,iN})$, vế phải được xác định là nhãn phân lớp C_q sao cho luật đó đạt độ tin cao nhất và được xác định như sau :

$$C_q = argmax_{C_h}\{c(A_q \Rightarrow C_h)|C_h \in C_1, ..., C_m, C_h \neq C_q\} \tag{0.18}$$

Với mục tiêu là sự đơn giản và dễ hiểu đối với các luật mờ sinh ra, chúng ta áp dụng phương pháp rút gọn vế trái luật bằng cách loại bỏ các thuộc tính với hy vọng sẽ ít tác động đến việc phân lớp. Mỗi luật sinh ra dạng $A_q \Rightarrow C_q$ sẽ được sinh tiếp các luật con $A_{q_i} \Rightarrow C_{q_i}$ như sau:

$$A_{q_i} \subseteq A_q, |A_{q_i}| \leq \alpha_{max} \tag{0.19}$$

trong đó α_{max} là độ dài (số thuộc tính trong điều kiện vế trái) tối đa của luật và được cho trước (thường bằng 3). Nhãn phân lớp ở vế phải được xác định theo (3.1).

Phương pháp lấy tổ hợp theo số thuộc tính trong vế trái luật như trên sẽ sinh một hệ luật với số lượng rất lớn, do sự bùng nổ tổ hợp của tập thuộc tính. Chẳng hạn với N thuộc tính và độ dài tối đa α_{max} thì mỗi luật $A_q \Rightarrow C_q$ sinh ra $C_N^1 + C_N^2 + ... + C_N^{\alpha_{max}}$ luật con. Tiếp theo chúng ta áp dụng phương pháp rút gọn hệ luật bằng hai phương pháp: thứ nhất, dựa trên mức độ gần nhau giữa các luật (Định nghĩa 2.4) để hợp chúng lại với nhau theo ngưỡng kết nhập θ_{itg}. Thứ hai, áp dụng phương pháp sàng luật dựa trên giá trị của một trong các tiêu chuẩn sàng được đề xuất bởi [14] gồm c, s và c.s, tức là chúng ta sẽ lấy các luật có giá trị tiêu chuẩn từ cao xuống thấp.

Ví dụ minh họa phương pháp sinh luật dựa trên phân hoạch mờ trên miền của các thuộc tính đối với tập dữ liệu mẫu có hai thuộc tính như sau (hình vẽ 3.1). Mỗi siêu hộp chứa mẫu dữ liệu sẽ sinh một luật tương ứng, chẳng hạn luật

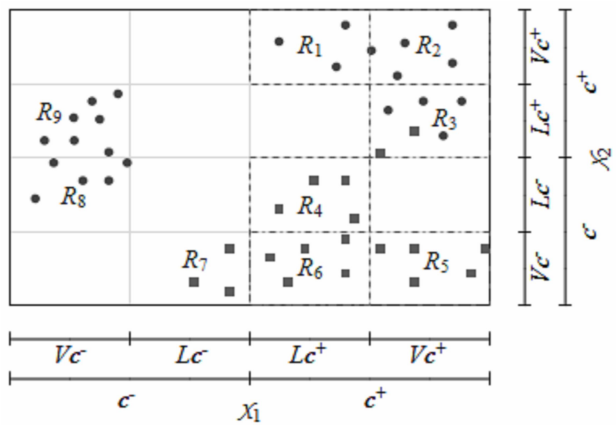
$$R_9 = (IF\ X_1\ is\ L.c^-\ and\ X_2\ is\ L.c^+\ THEN\ circle),$$

ta có các luật con sinh ra sẽ là

$$R_{9,1} = (IF\ X_1\ is\ L.c^-\ THEN\ circle), R_{9,2} = (IF\ X_2\ is\ L.c^+\ THEN\ circle).$$

Rõ ràng luật $R_{9,1}$ bao trùm lên siêu hộp tương ứng luật R_8, khi đó các luật sinh ra từ hai siêu hộp này có thể được loại bỏ.

Ngoài ra, dựa trên quan hệ kế thừa ngữ nghĩa, hai luật $R_1 = (Lc^+, Vc^+) \Rightarrow' cirlce'$ và $R_2 = (Vc^+, Vc^+) \Rightarrow' cirlce'$ có thể được kết nhập thành một luật mới (bằng cách hợp hai hạng từ Lc^+ và Vc^+ của thuộc tính X_1 thành c^+) có dạng $R_{12} = (c^+, Vc^+) \Rightarrow' cirlce'$. Quá trình hợp này sẽ được tiếp tục theo một ngưỡng qitg cho trước với hàm đánh giá mức độ gần nhau giữa các luật trong Định nghĩa 2.4, thậm chí việc kết nhập này thực hiện ngay cả khi các luật được sinh ra.



Hình 3.1. Phân hoạch mờ trên miền của hai thuộc tính và sinh luật

Như vậy mỗi thuộc tính, trước hết phải tính toán hệ phân hoạch các khoảng tính mờ cùng với các hạng từ dựa trên các tham số mờ gia tử theo các mức từ $1, 2, ..., k_j$ (k_j cho trước), ký hiệu $X_{(k_j)} = X_1 \cup ... \cup X_{k_j}$ là tập các hạng từ độ dài không quá k_j và $I_{(k_j)} = I_1 \cup ... \cup I_{k_j}$ là tập các khoảng tính mờ tương ứng. Quá trình sinh hệ luật mờ theo phương pháp trên được trình bày dưới dạng một thuật toán như sau:

Thuật toán 3.1. Sinh luật từ tập mẫu

Vào:

+ Tập mẫu $P = p_i = (d_{i1}, ..., d_{iN}; C_i)|i = 1, ..., M$ có N thuộc tính, M là số mẫu huấn luyện,

+ Các tham số mờ gia tử và mức khoảng tính mờ k_j của mỗi thuộc tính thứ j (có thể áp dụng k_j giống nhau cho mọi thuộc tính),

+ Tiêu chuẩn sàng luật PS (sử dụng 1 trong 3 tiêu chuẩn c, s hoặc $c.s$),

+ Phương pháp đánh giá trọng số luật CF (sử dụng 1 trong 4 đánh giá bởi các công thức (2.5) đến (2.13)).

+ Ngưỡng kết nhập các luật $\theta_{itg}(0 < \theta_{itg} < 1)$,

+ Độ dài luật tối đa, α_{max}, tức số điều kiện trong luật (thường đặt là 3),

+ Số luật cần sinh, K.

Ra:

+ Tập các luật mờ S_r.

Các bước:

Step1) Tính các khoảng tính mờ tương ứng với các hạng từ cho tất cả các thuộc tính từ độ dài 1 đến k_j.

Step2) Dùng thuật toán $HAFRE$ trong [2] sinh tập luật thô với độ dài N tại mức khoảng tính mờ cao nhất, k_j, gọi tập này là S_0.

Step3) Với mỗi độ dài luật $a(a = 1, 2, ..., \alpha_{max})$ thực hiện.

3.a) Sinh luật độ dài α từ tập S_0 bằng cách xóa bỏ $(N - a)$ điều kiện trong mỗi luật, số luật sinh ra trong bước này là $|S0|$, ký hiệu bởi tập S_a.

3.b) Kết nhập mỗi hai luật trong tập S_a có độ kết nhập (Định nghĩa 2.4) $itg > \theta_{itg}$ thành một luật mới, tập luật mới S^{α^*} có kích thước nhỏ hơn S_α (rõ ràng θ_{itg} càng bé thì tập S_{α^*} càng nhỏ). Bước này phải sử dụng thủ tục 2.1 để tính toán khoảng tính mờ bao hàm và Định nghĩa 2.3 để xác định độ kết nhập của các hạng từ.

3.c) Với mỗi tập S_{α^*}, loại bỏ các luật cùng vế trái nhưng khác vế phải, chỉ giữ lại một trong số chúng để đảm bảo tập luật không bị nhập nhằng trong lập luận (để đơn giản chúng tôi chọn một luật ngẫu nhiên trong số này).

Step4) Tính hợp các tập luật: $S^* = \bigcup_{\alpha=1,2,...,\alpha_{max}} S^{\alpha^*}$.

Step5) Tính độ tin cậy, độ hỗ trợ và trọng số CF của các luật trong tập S^* (sử dụng các công thức (2.5)-(2.13)).

Step6) Sàng các luật trong tập S^* theo tiêu chuẩn PS đã cho. Sắp các luật theo nhóm (có cùng lớp đầu ra) thứ tự giảm của tiêu chuẩn sàng, sau đó chọn $[K/M]$ luật đầu trong mỗi nhóm. Các luật còn lại $K\%M$ (số dư của K chia M) được chọn theo giá trị tiêu chuẩn cao nhất trên tất cả các nhóm. $[K/M]$ là số nguyên, và $K\%M$ là số dư phép chia K/M.

3.2. Thiết kế giải thuật di truyền để chọn các luật - GA1

Việc áp dụng các tiêu chuẩn sàng để chọn ra một số ít các luật có thể đạt kết quả phân lớp không cao, cơ may bỏ qua các luật tốt khá nhiều [14]. Trước hết, áp dụng Thuật toán 3.1 để sinh một tập luật thô, ký hiệu S_r. Chọn ra K^* luật trong tập này sao cho độ chính xác phân lớp càng cao càng tốt và độ dài luật trung bình nhỏ. Thông thường ta đặt K^* nhỏ hơn $|S_r|$ rất nhiều, vậy sẽ có $S^L_{|S_r|}$ tập con luật có thể chọn và số này là rất lớn.

Các tác giả trong [13] đã cho thấy việc áp dụng giải thuật di truyền (GA) chọn tập luật tối ưu từ một tập cho trước Sr đạt kết quả rất tốt. Tuy nhiên, việc áp dụng GA theo sơ đồ mã hóa nhị phân gặp những hạn chế nhất định, đặc biệt thời gian tối ưu rất lớn. Ở đây, sử dụng sơ đồ mã hóa số thực có tích hợp tham số nhiệt mô phỏng quá trình tôi luyện thép, gọi là TGA [3, 4]. Mỗi cá thể độ dài K^* biểu diễn tập chỉ số các luật được chọn trong S_r, giá trị các gen gi giới hạn trong khoảng đơn vị [0,1], ánh xạ mỗi gen này thành một số nguyên từ 1 đến K^* bằng hàm sau:

$$\forall g_i \in [0,1], i = 1,2,...,K^*, index = 1 + g_i.(|S_r| - 1). \tag{0.20}$$

Để đánh giá độ phù hợp của cá thể trong GA, sử dụng hàm gồm hai tham số: thứ nhất, độ dài luật trung bình avg; thứ hai, độ chính xác phân lớp trong tập mẫu học perf. Hàm có dạng sau:

$$fitness(individual) = w_1.(\alpha_{max} - avg)/(\alpha_{max} - 1) + w_2.perf, \tag{0.21}$$

trong đó (w_1, w_2) là trọng số của độ dài trung bình và độ chính xác phân lớp, $w_1 + w_2 - 1$. α_{max} là độ dài (số điều kiện) tối đa của luật, biểu thức $(\alpha_{max} - avg)/(\alpha_{max} - 1)$ để chuẩn hóa giá trị avg về khoảng đơn vị $[1, 0]$.

Các phép toán di truyền được áp dụng theo [2].

4. THỬ NGHIỆM

Bài toán phân lớp các loại rượu (wine) được các tác giả sử dụng rất phổ biến để thử nghiệm mô hình phân lớp [9, 13 – 14, 20], tập dữ liệu này được công bố tại [24], bao gồm 178 mẫu chia thành 3 lớp theo tỷ lệ 59:71:48. Bài báo thử nghiệm tập mẫu này vì nó chứa 13 thuộc tính, là một trong những khó khăn cho các mô hình phân lớp khi có nhiều thuộc tính tham gia.

Trước hết chuẩn hóa tập mẫu này từ miền tham chiếu về miền định lượng ngữ nghĩa $[0, 1]$ bằng các ánh xạ tuyến tính dạng:

$$f_s : \ [a,b] \to [0,1], \quad x \to f_s(x) = (x - a)/(b - a), \tag{0.22}$$

trong đó $[a, b]$ là miền tham chiếu của thuộc tính. Mười ba thuộc tính của tập mẫu wine như sau:

Thuộc tính	min	max	a	b
0	11,03	14,83	10,03	15,83
1	0,74	5,8	0,64	5,9
2	1,36	3,23	1,26	3,33
3	10,6	30	9,6	31
4	70	162	69	163
5	0,98	3,88	0,88	3,98
6	0,34	5,08	0,24	5,18
7	0,13	0,66	0,03	0,76
8	0,41	3,58	0,31	3,68
9	1,28	13	1,18	13,1
10	0,48	1,71	0,38	1,81
11	1,27	4	1,17	4,1
12	278	1680	277	1681

Mô hình được thử nghiệm theo 2 chiến lược: thứ nhất, tất cả các tập mẫu được dùng để sinh luật và đánh giá sai số phân lớp; thứ hai (leave-one-out), mỗi lần thử nghiệm lấy ra một mẫu để đánh giá sai số phân lớp, số mẫu còn lại dùng để sinh luật. Chiến lược thứ hai được lặp lại với mỗi lần lặp lấy ra một mẫu kiểm tra theo thứ tự trong tập dữ liệu ban đầu. Kết quả thể hiện ở các lần thử nghiệm, và so sánh với [13, 14] cho thấy hiệu quả của phương pháp được đề xuất.

Áp dụng Giải thuật 3.1 để sinh luật

Để áp dụng Giải thuật 3.1, chúng ta đặt các tham số mờ gia tử giống nhau cho mọi thuộc tính như sau: $\mu_j(c^-) = 0,32, \mu_j(L) = 0,68, \theta_{itg} = 0,37, k_j = k_{max} = 3, \forall j$. Độ dài luật tối đa là $\alpha_{max} = 3$. Thử nghiệm trên 3 tiêu chuẩn sàng luật và 4 định nghĩa khác nhau về trọng số luật với tập các luật lần lượt là 3, 6, 9, 30, 60, 90, 300, 600, 900 luật.

Trường hợp 1. Tất cả các mẫu dữ liệu dùng để sinh luật, kết quả thể hiện trong bảng sau với phương pháp lập luận single winner rule. So sánh cho thấy kết quả của mô hình (chữ đậm) hầu hết tốt hơn so với [13] (chữ nghiêng), đặc biệt trong hai tiêu chuẩn sàng s và $c.s$, ô đánh dấu '*' là kết quả tốt nhất của mỗi tập luật.

Bảng 4.1. Kết quả trường hợp 1 với lập luận single winner rule và so sánh với [13]

Sàng	Trọng số luật	Số luật								
		3	6	9	30	60	90	300	600	900
Độ dài trung bình tập luật		3,00	3,00	3,00	3,00	3,00	3,00	3,00	3,00	3,00
c	CF^I(confidence)	2,25	3,93	15,73	39,89	66,29	77,53	94,38	96,63*	100*
		8,50	15,80	22,00	49,90	68,10	77,20	93,60	97,8*	99,0*
c	CF^{II}(confidence-average)	2,25	3,93	15,73	39,89	66,29	77,53	94,38	96,63*	100*
		8,50	15,80	22,00	49,90	68,10	77,20	93,60	97,8*	99,0*
c	CF^{III}(confidence-2nd max)	2,25	3,93	15,73	39,89	66,29	77,53	94,38	96,63*	100*
		8,50	15,80	22,00	49,90	68,10	77,20	93,60	97,8*	99,0*
c	CF^{IV} (no weight)	2,25	3,93	15,73	39,89	66,29	77,53	94,38	96,63	100*
		8,50	15,80	22,00	49,90	68,10	77,20	93,60	97,8*	99,0*
Độ dài trung bình tập luật		1,00	1,17	1,22	1,57	1,77	1,97	2,42	2,65	2,72
s	CF^I(confidence)	89,33	91,57	91,01	91,01	92,13	92,13	92,13	93,26	93,82
		49,40	52,20	78,10	84,30	89,30	89,90	91,60	92,70	92,10
s	CF^{II}(confidence-average)	87,64	91,01	91,57	94,94	96,07*	95,51	94,38	95,51	94,38
		60,70	57,30	88,80	89,90	92,70	92,70	93,30	92,70	93,30
s	CF^{III}(confidence-2nd max)	85,39	87,08	87,64	94,38	94,94	94,38	95,51	96,07	96,07
		54,50	48,90	88,80	91,00	94,40	94,9*	96,1*	94,40	96,10
s	CF^{IV} (no weight)	84,27	79,21	79,21	75,84	74,16	74,16	73,60	72,47	71,91
		39,90	39,90	39,90	39,90	39,90	39,90	39,90	39,30	39,30
Độ dài trung bình tập luật		1,33	1,50	1,56	1,90	2,00	2,13	2,48	2,67	2,71
c,s	CF^I(confidence)	92,13	92,70	91,57	93,82	92,70	92,70	92,70	93,26	93,82
		87,60	82,00	91,00	93,80	91,00	91,60	92,10	92,70	92,10
c,s	CF^{II}(confidence-average)	93,26*	92,70	92,13	95,51*	96,07*	96,07*	94,94	95,51	94,38
		89,30	88,80	93,80	94,90	92,70	93,80	93,30	93,30	93,80
c,s	CF^{III}(confidence-2nd max)	92,13	94,38*	93,82*	94,94	94,94	95,51	96,07*	96,07	96,07
		91,0*	91,0*	94,9*	96,1*	95,5*	94,9*	95,50	96,10	96,10
c,s	CF^{IV} (no weight)	91,01	87,08	86,52	87,08	81,46	80,90	74,72	73,03	72,47
		81,50	39,90	39,90	39,90	39,90	39,90	39,90	39,90	39,90

Trường hợp 2, Thử nghiệm theo chiến lược leave-one-out, kết quả trung bình sau 178 lần lặp đánh giá trên tập kiểm tra (testing set) thể hiện trong bảng sau với phương pháp lập luận weighted vote. So sánh cho thấy kết quả của mô hình (chữ đậm) hầu hết tốt hơn so với [13] (chữ nghiêng), đặc biệt trong hai tiêu chuẩn sàng s và $c.s$. Đối với tiêu chuẩn sàng c thì kết quả [13] tốt hơn, tuy nhiên tiêu chuẩn này cho kết quả phân lớp thấp hơn so với hai tiêu chuẩn còn lại (ô đánh dấu '*' là kết quả tốt nhất của mỗi tập luật).

Bảng 4.3. Kết quả trung bình trường hợp 2 và so sánh với [13]

Sàng	Trọng số luật	Số luật								
		3	6	9	30	60	90	300	600	900
Độ dài trung bình tập luật		3,00	3,00	3,00	3,00	3,00	3,00	3,00	3,00	3,00
c	CF^{I}(confidence)	2,25	5,06	15,17	35,96	55,62	62,36	84,27	89,33	93,82
		7,10	14,50	19,90	47,10	65,10	73,80	89,50	92,8*	93,8*
c	CF^{II}(confidence-average)	2,25	5,06	15,17	35,96	55,62	62,36	84,27	89,33	93,82
		7,10	14,50	19,90	47,10	65,10	73,80	89,50	92,8*	93,8*
c	CF^{III}(confidence-2nd max)	2,25	5,06	15,17	35,96	55,62	62,36	84,27	89,33	93,82
		7,10	14,50	19,90	47,10	65,10	73,80	89,50	92,8*	93,8*
c	CF^{IV} (no weight)	2,25	5,06	15,17	35,96	55,62	62,36	84,27	89,33	93,82
		7,10	14,50	19,90	47,10	65,10	73,80	89,50	92,8*	93,8*
Độ dài trung bình tập luật		1,00	1,17	1,22	1,57	1,77	1,97	2,42	2,65	2,71
c	CF^{I}(confidence)	88,20	92,13	94,94	95,51*	95,51	94,38	97,75*	94,94	95,51*
		36,00	45,50	71,30	82,00	88,20	88,20	89,30	90,40	90,40
c	CF^{II}(confidence-average)	87,08	94,38	95,51*	95,51*	95,51	95,51	97,19	94,94	95,51*
		47,20	57,30	76,40	89,30	92,10	92,10	92,10	92,10	91,60
c	CF^{III}(confidence-2nd max)	85,39	84,83	91,57	92,70	96,07	97,19	96,07	94,38	94,94
		21,30	36,50	77,00	89,30	92,10	93,3*	93,3*	92,70	92,10
c	CF^{IV} (no weight)	84,27	86,52	92,13	94,38	93,82	93,26	96,07	95,51*	95,51*
		39,90	39,90	39,90	39,90	39,90	39,90	39,90	39,30	39,30
Độ dài trung bình tập luật		1,33	1,50	1,56	1,90	2,02	2,13	2,48	2,67	2,71
c	CF^{I}(confidence)	91,57	92,13	92,13	95,51*	96,63*	97,75*	95,51	94,38	95,51*
		87,10	79,80	86,50	89,90	89,30	88,80	89,90	90,40	90,40
c	CF^{II}(confidence-average)	92,13*	92,70	92,70	95,51*	96,63*	97,75*	94,94	94,38	95,51*
		88,80	89,30	93,3*	94,90	92,10	91,60	92,10	92,70	91,60
c	CF^{III}(confidence-2nd max)	91,57	94,38*	93,82	95,51*	96,63*	97,19	94,94	94,38	94,94
		90,4*	90,4*	93,3*	95,5*	93,8*	92,70	93,3*	92,70	92,10
c	CF^{IV} (no weight)	90,45	91,57	92,13	95,51*	96,63*	98,31	96,07	94,94	95,51*
		81,50	28,70	39,90	39,90	39,90	39,90	39,90	39,90	39,30

Trường hợp 3, Chiến lược thử nghiệm giống trường hợp 2 (leave-one-out) nhưng tập luật ít hơn, cả hai phương pháp lập luận được dùng với tiêu chuẩn sàng là *c.s*. So sánh trong bảng sau cho thấy kết quả của mô hình (chữ đậm) tốt hơn đáng kể so với [14] (chữ nghiêng) trong cả hai phương pháp lập luận, ô đánh dấu '*' là kết quả tốt nhất của mỗi tập luật và tương ứng với mỗi phương pháp lập luận.

Bảng 4.4. Kết quả trung bình trường hợp 3 và so sánh với [14]

Sàng	Trọng số luật	Số luật				
		3	6	9	12	15
Phương pháp lập luận *single winner rule*						
Đọ dài trung bình tập luật		1,33	1,50	1,56	1,50	1,60
c.s	CF^{I}(confidence)	91,57	89,33	89,33	88,76	91,01
		89,89*	83,15	91,57	93,26	91,57
c.s	CF^{II}(confidence-average)	92,13*	90,45	89,89	90,45	94,38*
		89,89*	85,96*	92,13	92,7	91,57
c.s	CF^{III}(confidence-2nd max)	91,57	91,01*	92,70*	92,13*	91,01
		89,33	84,83	93,26*	93,26*	94,38*
c.s	CF^{IV}(no weight)	90,45	82,58	81,46	83,15	85,96
		89,89*	80,34	88,76	93,26*	88,76
Phương pháp lập luận *weighted vote*						
Đọ dài trung bình tập luật		1,33	1,50	1,56	1,50	1,60
c.s	CF^{I}(confidence)	91,57	92,13	92,13	93,82	96,07
		89,89*	87,64	93,26	94,94*	95,51*
c.s	CF^{II}(confidence-average)	92,13*	92,70	92,70	94,38*	96,07
		89,89*	88,76	93,26	94,38	94,38
c.s	CF^{III}(confidence-2nd max)	91,57	94,38*	93,82*	94,38*	96,63*
		89,33	89,33*	94,38*	94,38	94,38
c.s	CF^{IV}(no weight)	90,45	91,57	92,13	93,26	96,07
		89,89*	87,08	93,82	94,38	95,51*

Áp dụng giải thuật di truyền GA1 để chọn luật

Áp dụng tiêu chuẩn sàng *c.s*, trọng số luật là CF^{III} và cách lập luận weighted vote, sử dụng Thuật toán 3.1 để sinh tập 900 luật. Dùng GA1 để chọn một tập luật tối ưu (hiệu quả phân lớp cao) trong số 900 luật này. Khởi tạo quần thể xuất phát 500 cá thể ngẫu nhiên,

tiến hóa 150 thế hệ với trọng số hàm $fitness w_1 = 0.05, w_2 = 0.95$. Thử nghiệm GA1 này 5 lần với số luật chọn tương ứng là 3,4,5,6 và 7. Kết quả trong bảng 4.5, so sánh với [13] cho thấy hiệu quả mô hình này cao hơn trong trường hợp 5,6 và 7 luật được chọn (Avg-len là độ dài trung bình tập luật, Perf là độ chính xác phân lớp).

Bảng 4.5. Kết quả áp dụng GA1 và so sánh với [13]

Số luật	Mô hình của chúng tôi theo GA1		Mô hình [13]	
	Avg-len	Perf	Avg-len	Perf
3	1,33	92,70	1,33	93,3
4	1,75	96,07	1,5	97,2
5	1,4	98,88	1,6	98,3
6	2,33	100	2	99,4
7	1,86	100	2	100

4.c. Áp dụng giải thuật di truyền GA2 để tối ưu tham số mờ gia tử

Có hai cách áp dụng GA2 để tối ưu tham số mờ: thứ nhất, tối ưu một bộ tham số mờ gia tử và áp dụng cho tất cả các thuộc tính; thứ hai, tối ưu tham số mờ cho từng thuộc tính là khác nhau. Với cách thứ hai, các thuộc tính khác nhau sẽ có các tham số mờ khác nhau, điều này đúng với thực tế định lượng ngữ nghĩa của các hạng từ là khác nhau ở các thuộc tính. Tuy vậy, cách này tạo ra không gian tìm kiếm của GA khá lớn, thay vì tìm một bộ gồm 2 tham số $\mu(c^-), \mu(L)$ trong cách 1, ở cách 2 phải tìm 13 bộ gồm 13 bộ (26 tham số). Để đảm bảo tính trực quan của các luật cũng như tránh trường hợp tối ưu đưa các tham số mờ vào giá trị cực đoan, giới hạn giá trị các gen mã hóa tham số mờ trong đoạn $[0,2,\ 0,8]$. Trong cả hai cách đều chọn ra 3 luật để đánh giá theo tiêu chuẩn sàng $c.s$, trọng số luật là CF^{III}, phương pháp lập luận weighted vote và ngưỡng để kết nhập các luật $q = 0,1$. Sau đó, sử dụng bộ tham số tìm được của GA2 để sinh tập 900 luật và áp dụng GA1 chọn ra tập con các luật để phân lớp.

• **Cách 1 ($GA2^*$):** áp dụng GA2 tối ưu một bộ tham số mờ gồm $\mu(c^-), \mu(L)$. Khởi tạo quần thể xuất phát gồm 10 cá thể, lặp 20 thế hệ với trọng số hàm fitness $w_1 = 0.005, w_2 = 0.995$. Thu được kết quả $\mu(c-) = 0.306, \mu(L) = 0.787$. Áp dụng tiếp GA1 chọn ra tập ít luật ta được kết quả trong bảng sau.

Bảng 4.6. Kết quả của $GA2^* + GA1$ và so sánh với GA1

Số luật	Mô hình của chúng tôi				Mô hình [13]	
	Áp dụng $GA2^* + GA1$		Áp dụng $GA1$			
	Avg-len	Perf	Avg-len	Perf	Avg-len	Perf
3	1	96,07	1,33	92,70	1,33	93,3
4	1,5	98,31	1,75	96,07	1,5	97,2
5	1,6	100	1,4	98,88	1,6	98,3
6	1,67	100	2,33	100	2	99,4
7	1,57	100	1,86	100	2	100

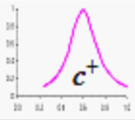
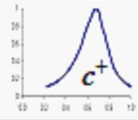
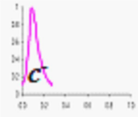
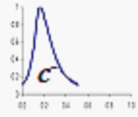
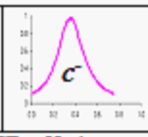
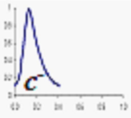
Bảng 4.7. Kết quả các tham số mờ của từng thuộc tính trong GA2

Thuộc tính	$\mu(c^-)$	$\mu(L)$
0	0,749965	0,534655
1	0,366686	0,733218
2	0,414166	0,669746
3	0,677943	0,725562
4	0,310051	0,285119
5	0,232626	0,469307
6	0,272307	0,705672
7	0,335617	0,330898
8	0,292057	0,542233
9	0,406566	0,721982
10	0,510205	0,695906
11	0,714500	0,376053
12	0,218494	0,595137

Bảng 4.8. Kết quả của $GA2^{**} + GA1$, so sánh với các kết quả trước

Số luật	Mô hình của chúng tôi						Mô hình [13]	
	$GA2^{**} + GA1$		$GA^{*} + GA1$		$GA1$			
	Avg len	Perf	Avg len	Perf	Avg len	Perf	Avg len	Perf
3	2	100	1	96,07	1,33	92,70	2,33	100
4	1,75	100	1,5	98,31	1,75	96,07	1,25	98,9
5	1,8	100	1,6	100	1,4	98,88	-	-
6	1,83	100	1,67	100	2,33	100	-	-
7	1,71	100	1,57	100	1,86	100	-	-

Bảng 4,9. Hệ 3 luật được chọn với độ chính xác 100% trong Bảng 4.8

Rules	Antecedents						Class
	X_0	X_5	X_6	X_9	X_{10}	X_{12}	
1	DC		DC	DC	DC		0
	R_1 : IF X_5 is c^+ AND X_{12} is c^+ THEN $class_0$						
2		DC	DC		DC	DC	1
	R_2 : IF X_0 is c^- AND X_9 is c^- THEN $class_1$						
3	DC	DC		DC		DC	2
	R_3 : IF X_6 is c^- AND X_{10} is c^- THEN $class_2$						

• **Cách 2 ($GA2^{**}$):** áp dụng GA2 tối ưu các tham số mờ $\mu_j(c^-), \mu_j(L)$ cho các thuộc tính khác nhau. Trong cách này đòi hỏi không gian tìm kiếm lớn, khởi tạo quần thể xuất phát gồm 50 cá thể, lặp 100 thế hệ với trọng số hàm $fitness w_1 = 0.005, w_2 = 0.995$. Thu được kết quả tham số mờ trong bảng 4.7 (thuộc tính gạch chân không được dùng trong việc sinh luật) với kết quả phân lớp 100% trong 3 luật. Áp dụng tiếp GA1 chọn ra tập ít luật ta được kết quả trong bảng 4.8 (dấu '-' ký hiệu không có kết quả thử nghiệm, DC ký hiệu thuộc tính không tham gia vào phần điều kiện luật (Don't Care)).

5. KẾT LUẬN

Trong bài báo này, chúng tôi đã đề xuất một mô hình phân lớp dựa trên đại số gia tử, trong đó các luật mờ có trọng số với hai phương pháp lập luận phân lớp *single winner rule* và *weighted vote* [13, 14]. Rõ ràng, một hệ luật mờ mà mỗi luật với trọng số của mình đóng vai trò mức độ tác động đến kết quả lập luận của hệ, điều này rất phù hợp với thực tế các ứng dụng.

Khác với mô hình trong [3] sinh luật từ dữ liệu khi các thuộc tính phải xác định trước. Ở đây, mục tiêu chọn một số ít các thuộc tính có ảnh hưởng lớn đến phân lớp cho mỗi luật, bài báo đề xuất mô hình sinh luật từ dữ liệu nhưng có chọn lọc thuộc tính dựa trên tiêu chuẩn sàng các luật [13, 14]. Theo tiếp cận đại số gia tử, miền của mỗi thuộc tính được phân hoạch các khoảng tính mờ mức k_j và xác định các hạng từ tương ứng. Các mẫu dữ liệu sẽ xác định các siêu khối (hypercube) trong không gian các khoảng tính mờ, các luật mờ sinh từ các siêu khối này và sau đó được kết nhập tạo ra các luật mới. Áp dụng các tiêu chuẩn sàng, đánh giá hiệu quả phân lớp trên tập luật này đối với dữ liệu mẫu.

Kết quả thử nghiệm cho thấy mô hình này, sau khi đã tối ưu tham số mờ gia tử, tốt hơn nhiều so với [13, 14]. Đối với trường hợp chỉ áp dụng tiêu chuẩn sàng luật, độ chính xác phân lớp tốt nhất của các tiêu chuẩn và trọng số luật tăng 6,06% so với [14] trên tất cả các tập luật đánh giá trong Bảng 4.1, tăng 25,99% trong Bảng 4.3, tăng 8,13% theo phương pháp lập luận weighted vote và 6,55% theo phương pháp lập luận single winner rule trong Bảng 4.4. Khi áp dụng GA để chọn luật (GA1), mô hình này đạt 100% tại 6 luật với độ dài luật trung bình 2,33, trong khi phương pháp [13] đạt 100% tại 7 luật với độ dài trung bình 2 trong Bảng 4.5. Đặc biệt, khi áp dụng GA2 để tối ưu tham số mờ gia tử, mô hình cho kết quả 100% tại 3 luật (Bảng 4.9) với độ dài trung bình 2, trong khi phương pháp tối ưu của [13] với số phân hoạch khoảng tính mờ của mỗi thuộc tính lên đến 14 cũng đạt 100% tại 3 luật nhưng độ dài trung bình 2,33 (Bảng 4.8). Hơn nữa, tại 4 luật thì mô hình này đạt 100% trong khi [13] chỉ đạt 98,9%. Điều này cho thấy mô hình này đảm bảo tính ổn định khi số luật tăng.

TÀI LIỆU THAM KHẢO

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

[2] N.C. Ho, T.T. Son, D.T. Long, Tiếp cận đại số gia tử cho phân lớp mờ, *Tạp chí Tin học và Điều khiển học* **25** (1) (2009).

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

[4] Chen Ji-lin, Hou Yuan-long, Xing Zong-yi, Jia Li-min, and Tong Zhong-zhi, A multi-objective genetic-based methodfor design fuzzy classification systems, *IJCSNS International Journal of Computer Science and Network Security* vol. **6** (8A) (August 2006).

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

[6] Chia-Chong Chen, Design of PSO-based fuzzy classification systems, *Tamkang Journal of Science and Engineering* Vol. **9** (1) (2006) 63–70.

[7] 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.

[8] Eghbal G. Mansoori, Mansoor J. Zolghadri, Seraj D. Katebi, A weighting function for improving fuzzy classification systems performance, *Fuzzy Sets and Systems* Vol. **158** (2007) 583–591.

[9] Enwang Zhou, Alireza Khotanzad, "Fuzzy classifier design using genetic algorithms", Southern Methodist University, 2007.

[10] N.C. Ho, 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.

[11] H. Ishibuchi, T. Nakashima, T Murata, Three-objective genetics-based machine learning for linguistic rule extraction, *Information Scienc* vol. **136** (2001) 109–133.

[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] 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, Japan, 2004.

[14] Hisao Ishibuchi, Takashi Yamamoto, Rule weight specification in fuzzy rule-based classification systems, *IEEE Trans. On Fuzzy Systems* **13** (4) (August 2005) 428–435.

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

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

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

[18] 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.

[19] Jiri Kubalik, Leon Rothkrantz, Jiri Lazansky, Genetic Programming Fuzzy Rule Extractor Using Class Preserving Representation, *Proceedings of the 13th Belgium-Netherlands Conference on Artificial Intelligence*, 2001 (167–174).

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

[21] 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).

[22] R. Kruse, and U. Nauck, "Design and implementation of a neuro-fuzzy data analysis tool in java", Technische Universitt Braunschweig, Braunschweig, 1999.

[23] 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.

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

[25] Ulrich Bodenhofer, Genetic Algorithms- Theory and Applications, Fuzzy Logic Laboratorium Linz-Hagenberg, (2003).

[26] 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.

[27] 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* (2006) 41–52.

Nhận bài ngày 16 - 11 - 2009

Nhận lại sau sửa ngày 24 - 3 -2010