Tạp chí Tin học và Điều khiển học, T.29, S.4 (2013), 325–337

MỘT PHƯƠNG PHÁP THIẾT KẾ HỆ PHÂN LỚP MỜ DỰA TRÊN VIỆC MỞ RỘNG LƯỢNG HÓA ĐẠI SỐ GIA TỬ*

PHẠM ĐÌNH PHONG[1], NGUYỄN CÁT HỒ[2], TRẦN THÁI SƠN[2], NGUYỄN THANH THỦY[3]

[1] *Công ty Prévoir Việt Nam; Email: dinhphong_pham@gmail.com*
[2] *Viện Công nghệ thông tin, Viện Hàn lâm KH và CNVN; ncho@gmai.com; trn_thaison@yahoo.com*
[3] *Trường Đại học Công nghệ, Đại học Quốc gia Hà Nội; nguyenthanhthuy@vnu.edu.vn*

Tóm tắt. Phương pháp lượng hóa đại số gia tử (ĐSGT) theo cách truyền thống [1-2] đã đem lại những thành công bước đầu trong ứng dụng ĐSGT vào bài toán thiết kế các hệ phân lớp mờ (HPLM) với ngữ nghĩa tập mờ tam giác của các từ ngôn ngữ [3, 4, 11]. Ngữ nghĩa số của mỗi từ ngôn ngữ được xác định bởi giá trị định lượng ngữ nghĩa chỉ là một điểm, cho phép xác định đỉnh của tập mờ hình tam giác. Phương pháp lượng hóa ĐSGT mở rộng đã được công bố trong [10], được bổ sung thêm khả năng mô hình hóa ngữ nghĩa lõi của từ, cho phép xây dựng các phân hoạch trên miền các thuộc tính dựa trên chính các khoảng tính mờ mức k và cho phép định lượng ngữ nghĩa lõi của các từ dưới dạng khoảng được sử dụng để xác định đáy nhỏ của các tập mờ hình thang. Bài báo đề xuất một phương pháp thiết kế các từ ngôn ngữ và HPLM dạng luật với ngữ nghĩa tập mờ của các từ mong muốn dạng hình thang dựa trên phương pháp lượng hóa ĐSGT mở rộng này và khảo sát tính hiệu quả của phương pháp lượng hóa mới khi giải quyết bài toán phân lớp. Kết quả thực nghiệm với 10 tập dữ liệu mẫu chuẩn cho thấy phương pháp lượng hóa ĐSGT mới mềm dẻo hơn và cho kết quả tốt hơn.

Từ khóa. Hệ luật mờ phân lớp, phân hoạch mờ, đại số gia tử, ánh xạ định lượng khoảng.

Abstract. The conventional method of quantification of hedge algebras [1-2] has achieved some effective successes in its application to hedge algebras to the fuzzy classifier design problem using fuzzy set based semantics of linguistic terms in the form of triangle fuzzy sets [3, 4, 11]. The numeric semantic of a term defined by its semantically quantifying mapping value is a point which is relevant to define the vertex of the triangular fuzzy set. An extended quantification method of hedge algebras proposed in [10], using partitions of the feature spaces, based on a degree k semantically quantifying mapping intervals, allows to model the core semantics of a terms in the form of an interval, which is upper base of a trapezoidal fuzzy set. This paper proposes a method for designing linguistic terms and fuzzy rule based classifiers with trapezoidal fuzzy set semantics, based on this extended hedge algebras quantification and examines the effectiveness of the new quantification method in solving classification problems. The experimental results over 10 datasets have shown that the proposed method is more flexible and produces better results.

Key words. Fuzzy classification system, hedge algebras, fuzzy partition, interval semantically quantifying mapping.

1. MỞ ĐẦU

Trong những năm gần đây, hệ mờ dựa trên luật đã có những ứng dụng thành công trong nhiều lĩnh vực khác nhau. Hệ phân lớp mờ (HPLM) là trường hợp đơn giản nhất của hệ mờ

*Bài báo được thực hiện với sự hỗ trợ từ quỹ phát triển KHCNVN (Nafosted), mã số 102.01-2011.06.

dựa trên luật. Một hướng nghiên cứu trong lĩnh vực này đang được quan tâm là xây dựng HLPM dựa trên hệ luật mờ dạng if-then [3–9] và đã cho kết quả khá tốt so với các phương pháp khác.

Trong tiếp cận lý thuyết tập mờ [5–9], các tập mờ của các giá trị ngôn ngữ hầu hết là dạng tam giác cố định và các giá trị ngôn ngữ chỉ là nhãn được người thiết kế gán cho dựa trên cảm nhận trực giác, do đó chúng không thể phản ánh ngữ nghĩa của các giá trị ngôn ngữ tương ứng một cách thích đáng. Mặc dù hệ luật chứa các từ ngôn ngữ với ngữ nghĩa biểu thị bằng tập mờ, nhưng bài toán thiết kế các từ ngôn ngữ cùng với ngữ nghĩa của chúng trong phạm vi lý thuyết tập mờ lại chưa được đặt ra một cách rõ ràng.

ĐSGT cung cấp một cơ sở hình thức toán học cho việc mô hình hóa và thiết kế các từ ngôn ngữ cùng với ngữ nghĩa dựa trên tập mờ của chúng và có thể ứng dụng trong quá trình thiết kế tập giá trị ngôn ngữ cùng với ngữ nghĩa dựa trên tập mờ cho việc xây dựng tự động cơ sở luật của hệ phân lớp mờ. Phương pháp thiết HPLM với ngữ nghĩa ĐSGT [11] được chia làm hai giai đoạn. Giai đoạn thứ nhất là giai đoạn thiết kế tự động các từ ngôn ngữ cùng với ngữ nghĩa dựa trên tập mờ của chúng. Giai đoạn hai là giai đoạn xây dựng phương pháp trích rút hệ luật mờ tối ưu từ tập dữ liệu mẫu dựa trên ngữ nghĩa tích hợp với ngôn ngữ thu được ở giai đoạn trên. Giai đoạn thứ nhất đòi hỏi phải giải bài toán xây dựng các khoảng lân cận ngữ nghĩa của tập các từ trong $X_{(k)}$ có độ dài không lớn hơn k sao cho chúng lập thành phân hoạch của [0, 1]. Do cách lượng hóa truyền thống [3, 4] chấp nhận tiên đề $\sum\{fm(hx) : h \in H\} = fm(x)$ dẫn đến các khoảng tính mờ của các hạng từ có độ dài k trong X_k đã đủ lấp đầy miền tham chiếu [0, 1] nên không còn không gian cho việc xây dựng các khoảng lân cận ngữ nghĩa cho các hạng từ có độ dài nhỏ hơn k, dẫn đến việc phải xây dựng hệ khoảng tương tự mức k [3, 4, 11]. Với phương pháp lượng hóa như vậy, các giá trị định lượng ngữ nghĩa có thể được xem như là ngữ nghĩa lõi của các từ và phù hợp với ý nghĩa của các đỉnh của các tập mờ tam giác tương ứng.

Để có thêm không gian cho việc xây dựng các khoảng lân cận ngữ nghĩa cho các hạng từ có độ dài nhỏ hơn k, trong [10] đã chấp nhận giả thiết $\sum\{fm(hx) : h \in H\} < fm(x)$. Khi đó, gia tử h_0 được sử dụng trong việc mô tả ngữ nghĩa của hạng từ phụ thuộc ngữ cảnh, nó đáp ứng được yêu cầu hình thức hóa trong trình bày và mô tả được sự thay đổi ngữ nghĩa theo ngữ cảnh xác định bởi sự hiện diện đồng thời với các hạng từ khác. Cách lượng hóa ĐSGT mở rộng này phép xây dựng các phân hoạch trên miền các thuộc tính dựa trên chính các khoảng tính mờ mức k do có đủ không gian để biểu diễn lân cận ngữ nghĩa của các hạng từ có độ dài nhỏ hơn k. Gia tử h_0 được sử dụng để xây dựng giá trị định lượng khoảng của các từ ngôn ngữ, h_0x xác định ngữ nghĩa lõi của chúng. Đây chính là cơ sở cho phép xây dựng các tập mờ hình thang với đáy nhỏ là giá trị định lượng khoảng của các từ ngôn ngữ tương ứng.

Mục tiêu bài báo là chứng minh khả năng ứng dụng hiệu quả của phương pháp lượng hóa ĐSGT mở rộng vào bài toán thiết kế các từ ngôn ngữ cùng với ngữ nghĩa dựa trên tập mờ hình thang và xây dựng HPLM dạng luật. Việc phát triển phương pháp thiết kế tự động các từ ngôn ngữ sẽ dựa trên việc xây dựng các phân hoạch mờ của miền các thuộc tính, được thiết lập từ các khoảng tính mờ mức k trong ĐSGT mở rộng. Thuật toán xây dựng các luật mờ khởi sinh cũng được cải tiến cho phù hợp với cách lượng hóa mới.

Bài báo được bố cục gồm 5 mục. Sau phần mở đầu, Mục 2 trình bày tóm tắt phương pháp lượng hóa ĐSGT mở rộng được đề xuất trong [10]. Mục 3 là trình bày phương pháp thiết kế ngôn ngữ gắn kết với ngữ nghĩa định lượng hình thang và thiết kế HPLM dạng luật dựa trên phương pháp định lượng ĐSGT mở rộng. Mục 4 trình bày các thử nghiệm và đánh giá. Và cuối cùng là kết luận.

2. MỞ RỘNG LƯỢNG HÓA ĐẠI SỐ GIA TỬ

ĐSGT là mô hình định tính nên để tính toán số được cần phải lượng hóa các đặc trưng của nó trên cơ sở ngữ nghĩa định tính. Có ba đặc trưng lượng hóa ĐSGT là độ đo tính mờ của ĐSGT, các khoảng tính mờ của các hạng từ và ánh xạ định lượng ngữ nghĩa. Trong [10] đề xuất phương pháp mở rộng các khái niệm này nhằm xây dựng một cơ sở hình thức hóa linh hoạt hơn mô tả được ngữ nghĩa phụ thuộc ngữ cảnh, cho phép mở rộng khả năng ứng dụng của ĐSGT. Mục này tóm tắt lại phương pháp lượng hóa ba đặc trưng này, làm cơ sở cho việc xây dựng phương pháp thiết kế các từ ngôn ngữ với ngữ nghĩa định lượng khoảng và tập mờ hình thang.

Cho ĐSGT mở rộng $\mathcal{AX}^* = (X^*, C, G, H^{ex}, \leq)$ của một ĐSGT $\mathcal{AX} = (X, C, G, H, \leq)$, trong đó $H^{ex} = H \cup \{h_0\}, h_0 \notin H$, (xem [10]).

Một hàm $fm : X^* \to [0,1]$ được gọi là độ đo tính mờ của ĐSGT $\mathcal{AX}^*$ nếu nó thỏa các tính chất:

(fm1) $fm(c^-) + fm(W) + fm(c^+) = 1$;

(fm2) $\sum_{h \in H^{ex}} fm(hu) = fm(u),\ \forall u \in H(G)$;

(fm3) $\forall h \in H^{ex},\ \forall x, y \in H(\{c^-, c^+\})$ thỏa $x, y \neq h_0 z$ với một z bất kì, $\dfrac{fm(hx)}{fm(x)} = \dfrac{fm(hy)}{fm(y)}$.

Giả sử $\mathcal{AX}^*$ là một ĐSGT tuyến tính mở rộng và một độ đo tính mờ $fm : X^* \to [0,1]$ thỏa các tính chất trên. Khi đó, với mỗi $k > 0$, mỗi hạng từ x của $X^*_{(k)}$ được liên kết với một khoảng trong $PI([0,1])$, tập tất cả các khoảng con của [0,1], được gọi là khoảng tính mờ mức k của x và nó được xây dựng quy nạp theo k như sau:

1) Với $k = 1$, xây dựng các khoảng tính mờ $\Im_1(c^-), \Im_1(W), \Im_1(c^+)$ với $|\Im_1(x)| = fm(x)$, sao cho chúng có thứ tự tương đồng với thứ tự của các hạng từ c^-, W, c^+.

2) Với $k > 1$, và $x \notin C$, ta xây dựng các khoảng tính mờ $\Im_k(x)$ sao cho:

(i) Nếu $|x| < k - 1$ thì $|\Im_k(x)| = |\Im_{k-1}(x)|$.

(ii) Nếu $|x| = k - 1$ thì $|\Im_k(x)| = \mu(h_0) fm(x)$.

(iii) Nếu $|x| = k$ thì $|\Im_k(x)| = fm(x)$.

(iv) Thứ tự của các khoảng tính mờ tương đồng với thứ tự của các hạng từ x.

Ta giả sử các khoảng tính mờ là đóng trái, chỉ có hằng có giá trị 1 là đóng cả hai đầu.

Ánh xạ định lượng ngữ nghĩa biểu diễn ngữ nghĩa lõi của các giá trị ngôn ngữ. Cho độ đo tính mờ fm của ĐSGT $\mathcal{AX}^*$. Khi đó ánh xạ định lượng khoảng f được định nghĩa như sau $f(x) = \Im_{l(x)+1}(h_0 x) \in PI[0,1],\ \forall x \in X^*$ và $l(x)$ là độ dài của x. Lưu ý rằng nếu $x = h_0 z$, thì $f(x) = \Im_{l(x)+1}(h_0 x) = \Im_{l(x)}(h_0 z)$.

Vì ngữ nghĩa lõi của các từ ngôn ngữ được xác định bởi các giá trị khoảng của ánh xạ định lượng khoảng của ĐSGT mở rộng được sử dụng trong mục sau để sinh ngữ nghĩa tập mờ hình thang của các từ ngôn ngữ của biến ngôn ngữ, nên ĐSGT mở rộng có thể được áp dụng trong việc giải quyết các bài toán ứng dụng khác nhau. Bài báo này sẽ chứng minh khả năng ứng dụng hiệu quả của chúng trong việc giải các bài toán phân lớp.

3. THIẾT KẾ HPLM DẠNG LUẬT VỚI NGỮ NGHĨA DỰA TRÊN ĐẠI SỐ GIA TỬ

Tri thức của HPLM dạng luật trong bài báo này là một tập các luật mờ có trọng số có dạng

Luật R_q : IF $\mathcal{X}_1$ is $A_{q,1}$ AND ... AND $\mathcal{X}_n$ is $A_{q,n}$ THEN C_q with CF_q, với $q = 1..N$ (3.1)
trong đó $\mathcal{X} = \{X_j, j = 1, ..., n\}$ là tập n biến ngôn ngữ (thuộc tính) và $A_{q,j}(j = 1, ..., n)$ là các nhãn ngôn ngữ của các điều kiện mờ trong tiền đề, C_q là tên lớp kết luận của R_q và N là số luật mờ, CF_q là trọng số hay độ tin cậy của luật thứ q. Thông thường, một bài toán phân lớp $\mathcal{P}$ được cho bởi một tập dữ liệu mẫu phân lớp $\boldsymbol{D}$ gồm m mẫu dữ liệu được gán nhãn $\boldsymbol{d}_p = [d_{p,1}, d_{p,2}, ..., d_{p,n}, C_p], p = 1, ..., m$, với n thuộc tính, M lớp kết luận (hay M nhãn) trong tập $\boldsymbol{C} = \{C_q : q = 1, ..., M\}$. Luật R_q dạng (3.1) có thể được viết gọn lại thành $\boldsymbol{A}_q \Rightarrow C_q$ with CF_q, trong đó A_q là các tiền đề của luật thứ q và C_q là một nhãn lớp.

Giải bài toán thiết kế HPLM dạng luật là xây dựng phương pháp trích rút một hệ luật mờ từ tập $\boldsymbol{D}$ cho HPLM sao cho đạt hiệu quả phân lớp cao đồng thời hệ luật thu được lại dễ hiểu đối với người dùng. Phương pháp thiết kế HPLM dạng luật với ngữ nghĩa ĐSGT bao gồm hai bước:
(1) Thiết kế tự động các hạng từ ngôn ngữ tối ưu và ngữ nghĩa dựa trên tập mờ của chúng đối với từng thuộc tính của tập dữ liệu mẫu với việc áp dụng giải thuật tiến hóa đa mục tiêu sao cho việc thiết kế là kết quả của sự tương tác giữa ngữ nghĩa của các hạng từ và dữ liệu.
(2) Trích rút các cơ sở luật mờ từ dữ liệu mẫu sao cho cơ sở luật mờ thu được có sự cân bằng hợp lý giữa hiệu quả phân lớp và tính dễ hiểu. Trên cơ sở tính đa dạng và tính phù hợp của các hạng từ thu được từ bước (1), mục tiêu của bước này là tìm giải pháp đạt được sự cân bằng (tradeoff) tối ưu giữa hiệu quả phân lớp và tính ít phức tạp và dễ hiểu của hệ luật trên cơ sở cân bằng giữa tính khái quát và tính riêng biệt của các hạng từ khi tương tác với dữ liệu mẫu.

Đối với bài toán phân lớp cụ thể, để thiết kế các hạng từ cho đặc tính F_j, các dữ kiện sau cần được chỉ ra để sinh tập các hạng từ $X_{j,(k_j)}$ và để xác định các ngữ nghĩa tập mờ của chúng:
• Các hạng từ sinh c_j^- và c_j^+, tập các gia tử âm $H^- = \{h_{j,i} : -q_j \le i \le -1\}$, tập các gia tử dương $H^+ = \{h_{j,i} : 1 \le i \le p_j\}$, gia tử h_0. Trong nghiên cứu này chỉ sử dụng một gia tử âm và một gia tử dương nên $p_j = q_j = 1$.
• Các ràng buộc của các tham số mờ để duy trì sự phù hợp ngữ nghĩa của các hạng từ

$$a_j \le fm_j(c^-) \le a'_j,\ b_j \le fm_j(W) \le b'_j,\ fm_j(c^-) + fm_j(W) + fm_j(c^+) = 1,$$

$$e_j \le \mu(h_{j,i}) \le e'_j,\ \Sigma_{h_{j,i} \in H_j} \mu(h_{j,i}) = 1$$

và độ dài tối đa của các hạng từ $k_j \le K$, với K là một số nguyên dương.

3.1. Thiết kế các hạng từ ngôn ngữ

Các nghiên cứu [3, 4, 11] cho thấy cú pháp và ngữ nghĩa của các hạng từ đóng vai trò rất quan trọng. Tuy nhiên, các cách tiếp cận trong phạm vi lý thuyết tập mờ, cú pháp và ngữ nghĩa định tính của các từ ngôn ngữ chưa được đề cập. Vì vậy, mục này trình bày việc thiết kế tối ưu các hạng từ mang ngữ nghĩa định tính cùng với ngữ nghĩa mờ của chúng cho bài toán phân lớp mờ và nó là cơ sở để tạo ra các luật mờ với ngữ nghĩa dễ hiểu đối với người sử dụng.

Phương pháp lượng hóa ĐSGT mở rộng là cơ sở hình thức hóa cho phép ngữ nghĩa định tính xác định hàm định lượng ngữ nghĩa khoảng [10] và dùng h_0x để sinh khoảng ngữ nghĩa lõi của hạng từ x, đồng thời, nó xác định các khoảng tính mờ của các từ của thuộc tính. Với ngữ nghĩa định lượng khoảng và khoảng tính mờ của các từ ta có thể xây dựng ngữ nghĩa tập mờ. Trong nghiên cứu này chúng là các tập mờ hình thang. Đây là đặc trưng làm cho cách tiếp cận ĐSGT khác biệt với cách tiếp cận dựa trên tập mờ.

Phương pháp luận trên được áp dụng vào việc thiết kế các hạng từ ngôn ngữ với ngữ nghĩa dựa trên tập mờ có độ dài tối đa là k_j cho bài toán phân lớp mờ như sau:

+ Đối với mỗi thuộc tính j của tập dữ liệu mẫu được liên kết với một ĐSGT $\mathcal{AX}_j$, tiến hành xây dựng tập các hạng từ $X_{j,(k_j)}$ có độ dài nhỏ hơn hoặc bằng k_j và có thứ tự theo ngữ nghĩa định tính của chúng.

+ Với mỗi $0 < k \leq k_j$, giả sử $X_{j,k} = \{x_{j,i} : i = 1, ..., N_{jk}\}$ với thứ tự $x_{j,1} \leq ... \leq x_{j,N_{jk}}$ và ký hiệu $\mathcal{J}_j$ là tập các tham số mờ của $\mathcal{AX}_j$. Với các giá trị cụ thể của các tham số mờ trong $\mathcal{J}_j$, xây dựng các khoảng tính mờ $\Im_k(x_{j,i})$ mức k và các giá trị định lượng ngữ nghĩa $f(x_{j,i})$ đối với các hạng từ trong $X_{j,k}$ [10] và các đại lượng này hoàn toàn được xác định. Các khoảng tính mờ $\Im_k(x_{j,i})$ thỏa mãn thứ tự $\Im_k(x_{j,1}) \leq ... \leq \Im_k(x_{j,N_{j_k}})$ và tạo thành một phân hoạch của [0, 1]. Ví dụ, Hình 3.1 biểu diễn hệ khoảng tính mờ với $k_j = 2$ của một ĐSGT có 2 gia tử, trong đó $L \in H^-$ và $V \in H^+$.

a. Hệ khoảng tính mờ của các hạng từ ngôn ngữ mức $k = 1$

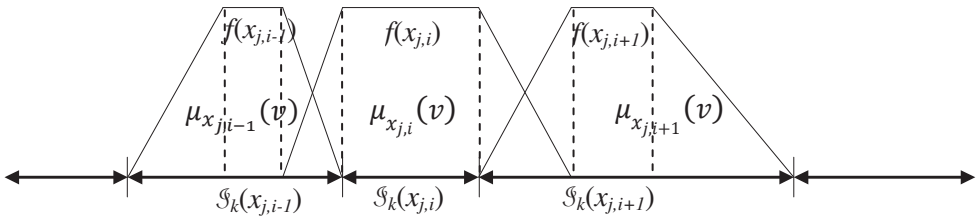
b. Hệ khoảng tính mờ của các hạng từ ngôn ngữ mức $k = 2$

Hình 3.1. Hệ khoảng tính mờ của các hạng từ ngôn ngữ với $k_j = 2$

Theo [10] thì $f(x_{j,i}) \subseteq \Im(x_{j,i})$ xác định giá trị định lượng ngữ nghĩa khoảng của hạng từ $x_{j,i}$ biểu thị khoảng định lượng ngữ nghĩa lõi của $x_{j,i}$. Ta có thể dùng nó làm đáy nhỏ của tập mờ hình thang của từ $x_{j,i}$ do các giá trị trong khoảng đó phù hợp với ngữ nghĩa định tính của từ nhất. Đối với các hạng từ $x_{j,i}$ có độ dài nhỏ hơn k thì $f(x_{j,i}) = \Im_k(x_{j,i})$, có độ dài bằng k thì $f(x_{j,i}) = \Im_{k+1}(h_0x_{j,i})$.

Để lập luận phân lớp cho các mẫu dữ liệu dựa trên hệ luật mờ dạng (3.1) bằng phương pháp lập luận Single Winner Rule [7, 8 , 11], mỗi giá trị ngôn ngữ được gán một hàm định lượng ngữ nghĩa. Với ngữ nghĩa định lượng khoảng và phương pháp phân hoạch trên miền thuộc tính theo các khoảng tính mờ được xác định ở trên, hàm định lượng ngữ nghĩa gán cho mỗi giá trị $x_{j,i}$ được xác định theo dạng hình thang $\mu_{x_{j,i}}(\nu)$ (công thức 3.2) sao cho giá trị hàm càng gần vị trí khoảng định lượng ngữ nghĩa lõi của $x_{j,i}$ là $f(x_{j,i})$ thì càng lớn và nhận giá trị 1 trong khoảng $f(x_{j,i})$, nó sẽ nhận giá trị 0 nếu chạm đầu mút phải của $f(x_{j,i-1})$ và đầu mút trái của $f(x_{j,i+1})$ (Hình 3.2).

Ký kiệu $\mathcal{L}(\bullet)$ và $\mathcal{R}(\bullet)$ lần lượt là điểm mút trái và mút phải của một khoảng bất kỳ. Trong Hình 3.2, hạng từ $x_{j,i}$ có độ dài nhỏ hơn k nên $f(x_{j,i}) = \Im_k(x_{j,i})$. Trong khi đó các hạng từ $x_{j,i-1}$ và $x_{j,i+1}$ có độ dài bằng k nên $f(x_{j,i-1}) \subseteq \Im_k(x_{j,i-1})$ và $f(x_{j,i+1}) \subseteq \Im_k(x_{j,i+1})$. Như vậy, các tập mờ hình thang được xây dựng có miền tin cậy (đáy nhỏ) là $f(x_{j,i})$ hay khoảng



Hình 3.2. Hàm định lượng dạng hình thang của các hạng từ

$[\mathcal{L}(f(x_{j,i})), \mathcal{R}(f(x_{j,i}))]$ và miền xác định (đáy lớn) là khoảng $[\mathcal{R}(f(x_{j,i-1})), \mathcal{L}(f(x_{j,i+1})]$. Giả sử đặt $a = \mathcal{R}(f(x_{j,i-1}))$, $b = \mathcal{L}(f(x_{j,i}))$, $c = \mathcal{R}(f(x_{j,i}))$, $d = \mathcal{L}(f(x_{j,i+1}))$, ta có công thức tính giá trị hàm định lượng ngữ nghĩa dạng hình thang công thức (3.2).

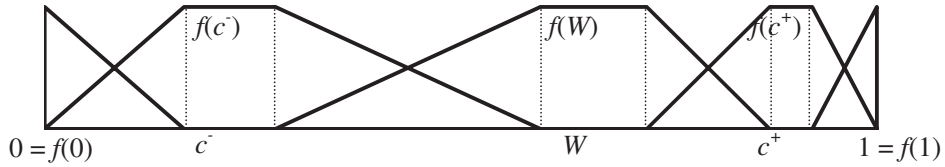
$$\mu_{x_{j,i}}(\nu) = \begin{cases} 0 \text{ khi } \nu < a \text{ hoặc } \nu > d \\ \dfrac{\nu - a}{b - a} \text{ với } a \le \nu \le b \\ \dfrac{d - \nu}{d - c} \text{ với } c < \nu \le d \end{cases} \tag{3.2}$$

Như vậy, khi cho các giá trị cụ thể của các tham số ngữ nghĩa (xem Mục 3.3) thì các khoảng tính mờ tích hợp với các hạng từ, các giá trị định lượng khoảng biểu thị khoảng định lượng ngữ nghĩa lõi của các hạng từ được sản sinh ra bằng một thuật toán. Với cách xây dựng tập mờ hình thang dựa trên các khoảng tính mờ và ánh xạ định lượng khoảng, được sinh ra cùng với các hạng từ như trong Hình 3.2, các tọa độ của mỗi đỉnh hình thang được xác định. Nghĩa là tập mờ hình thang được sinh ra bằng một thuật toán. Để làm giảm độ phức tạp của thuật toán sinh luật được trình bày trong phần tiếp theo, việc sinh các tập mờ hình thang được thực hiện đồng thời với việc tính các khoảng tính mờ và ánh xạ định lượng khoảng của các hạng từ ngôn ngữ.

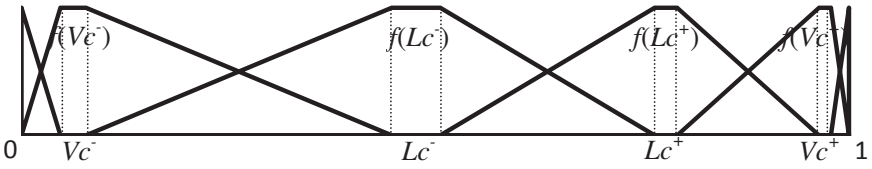
Các nghiên cứu [9, 11] đã phân tích đến sự cân bằng giữa tính phổ quát và tính riêng của các từ trong một hệ luật. Luật chứa càng nhiều từ phổ quát thì độ phức tạp của luật càng giảm và số mẫu được đoán nhận bởi luật càng lớn, trong khi luật có nhiều từ riêng có thể phân lớp đúng cho các mẫu có tính đặc thù. Vì vậy, để tăng tính phổ quát của các hạng từ trong $X_{j,(kj)}$, các tập mờ đã được thiết kế ở dạng đa thể hạt (multiple granularity) để tạo ra sự đa dạng của tính phổ quát và tính riêng của ngữ nghĩa các từ cho phép tìm kiếm sự cân bằng giữa tính phổ quát và tính riêng của việc thiết kế các từ ngôn ngữ cho bài toán $\mathcal{P}$. Mỗi thể hạt chỉ thiết kế tập mờ cho các hạng từ có độ dài bằng nhau và hai hạng từ 0 và 1. Đầu mút trái và phải của đáy lớn tập mờ hình thang tương ứng là đầu mút trái của ánh xạ định lượng khoảng của hạng từ bên phải và đầu mút phải của ánh xạ định lượng khoảng của hạng từ bên trái, có cùng độ dài liền kề nó. Hai tập mờ của hai hạng từ 0 và 1 có dạng hình tam giác. Hình 3.3 biểu diễn ba phân hoạch mờ, biểu thị cho ba thể hạt đối với thuộc tính thứ 29 của tập dữ liệu mẫu wdbc được cung cấp bởi KEEL-Dataset repository (http://sci2s.ugr.es/keel/datasets.php), có 569 mẫu dữ liệu, 30 thuộc tính và 2 nhãn lớp.

3.2. Thủ tục xây dựng tập luật khởi sinh từ tập dữ liệu mẫu

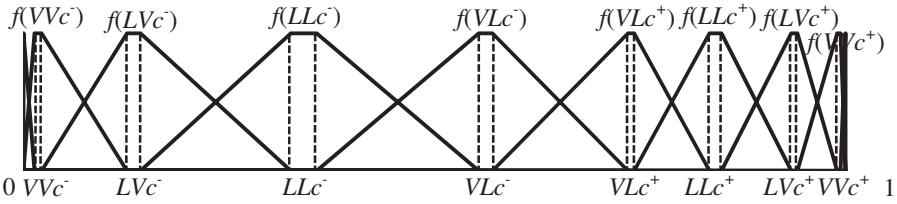
Để nghiên cứu vai trò của ngữ nghĩa dạng hình thang của từ ngôn ngữ, ta áp dụng phương pháp xây dựng HPLM với ngữ nghĩa tập mờ tam giác được nghiên cứu trong [11] để so sánh.



a. Các tập mờ hình thang của các hạng từ có độ dài 1



b. Các tập mờ hình thang của các hạng từ có độ dài 2



c. Các tập mờ hình thang của các hạng từ có độ dài 3

Hình 3.3. Các tập mờ được thiết kế dưới dạng đa thể hạt

Theo phương pháp này, đầu tiên ta xây dựng tập luật khởi sinh để làm đầu vào cho giải thuật tìm kiếm tập luật tối ưu cho HPLM cần xây dựng cho bài toán phân lớp $\mathcal{P}$.

Trong tiếp cận lý thuyết tập mờ được nghiên cứu trong [7–9], việc xây dựng tập luật khởi sinh từ tập dữ liệu mẫu dựa trên các phân hoạch mờ của không gian các thuộc tính được xác định bởi các tập mờ tam giác đã được lựa chọn dựa vào độ hỗ trợ và độ tin cậy của khai phá dữ liệu để trích rút luật mờ. Theo đó, mọi kết hợp có thể có giữa các tập mờ của các thuộc tính đều có thể là tiền đề của một luật mờ vì có thể có những mẫu trong $\boldsymbol{D}$ hỗ trợ cho chúng và, do đó, số luật cần xem xét để xây dựng tập luật khởi sinh là hàm mũ của số thuộc tính.

Với DSGT mở rộng, việc phân hoạch mờ trên mỗi miền thuộc tính j dựa trên các khoảng tính mờ mức k_j của các từ trong $X_{j,(kj)}$, tức là các khoảng rõ. Các phân hoạch này sinh ra một phân hoạch rõ $X_j,(k_j)$ của không gian n chiều của $\boldsymbol{D}$. Các tập của $\boldsymbol{\mathcal{H}}_S$ là các siêu hộp có dạng $\Im_{k_1}(x_{1,i1}) \times \Im_{k_2}(x_{2,i2}) \times ... \times \Im_{k_n}(x_{n,in})$, với $\Im_{k_j}(x_{j,i_j})$ là khoảng tính mờ mức k_j của từ $x_{j,i_j} \in X_{j,(k_j)},\ i_j = 1, ..., |X_{j,(k_j)}|$. Khi đó, mỗi mẫu $p_i \in \boldsymbol{D}$ xác định duy nhất một siêu hộp $HB_i \in \boldsymbol{\mathcal{H}}_S$ chứa nó. Siêu hộp này xác định một luật ngôn ngữ cơ sở có dạng IF $\mathcal{X}_1$ is $A_{q,1}$ AND ... AND $\mathcal{X}_n$ is $A_{q,n}$ THEN C_q, với $A_{q,j} = x_{j,i_j},\ j = 1, ..., n$, và C_q là tên lớp của mẫu p_i. Trên quan điểm luật sinh ra từ dữ liệu, ta chỉ xem xét việc sinh các luật cơ sở từ những siêu hộp có chứa mẫu dữ liệu. Do đó số luật cơ sở tối đa được sinh là N.

Từ tập luật cơ sở ta sinh ra các luật có độ dài ngắn hơn bằng cách bỏ đi một số điều kiện tiền đề của luật có độ dài n. Khi đó số luật tối đa phải xem xét là $m \times \sum_{l=1}^{lmax} C_n^l$ luật, trong đó m là số mẫu dữ liệu, $lmax$ là chiều dài tối đa của luật.

Đặt $\boldsymbol{\Pi} = \cup\{\mathcal{J}_j \cup \{k_j\} | j = 1, ..., n\}$ và gọi chung các giá trị trong Π là các tham số ngữ nghĩa. Với tham số đầu vào Π, ta sẽ thiết kế một thủ tục xây dựng một tập luật mờ, gọi là

tập khởi sinh, từ N_0 mẫu dữ liệu (tập dữ liệu huấn luyện) lấy từ tập dữ liệu mẫu $\boldsymbol{D}$ của bài toán $\mathcal{P}$. Ký hiệu thủ tục này là $\mathcal{V}(\boldsymbol{\Pi}, \boldsymbol{D}, N_0)$.

Thủ tục xây dựng tập luật khởi sinh bao gồm các bước sau:

Bước 1. Xây dựng tập các hạng từ, tập khoảng tính mờ, tập ánh xạ định lượng khoảng và các tập mờ hình thang của các hạng từ đối với mọi thuộc tính của tập dữ liệu mẫu.

Mỗi thuộc tính j của tập dữ liệu mẫu được liên kết với một ĐSGT mở rộng $\mathcal{AX}_j$. Với các từ nguyên thủy và tập các gia tử âm, gia tử dương và gia tử h_0 được xác định trước, tạo sinh các từ ngôn ngữ của $\mathcal{AX}_j$. Từ tập các giá trị cụ thể của các tham số ngữ nghĩa cho trong $\boldsymbol{\Pi}$ và tập các hạng từ $X_{j,k}$ mức k đã được sinh ra, tính các khoảng tính mờ $\Im_k(x_{j,i})$ mức k với $x_{j,i} \in X_{j,k}$ đối với mọi $k \leq k_j$. Tính các giá trị ánh xạ định lượng khoảng $f(x_{j,i})$ và xây dựng các tập mờ hình thang ứng với các hạng từ trong $X_{j,(k_j)}$.

Bước 2. Chọn lọc các luật khởi sinh từ các mẫu dữ liệu.

Khi các hạng từ tích hợp với các tập mờ được xác định, nhiệm vụ tiếp theo là sinh tất cả các luật ngôn ngữ mờ trực tiếp từ N_0 mẫu dữ liệu của $\boldsymbol{D}$. Mỗi mẫu dữ liệu $\boldsymbol{p}_l = (d_l, C_l),\ d_l \in \boldsymbol{D}$ sinh ra một luật mờ bởi các khoảng tính mờ $\Im_{kj}(x_{j,i})$ mức k_j của $X_{j,(k_j)}$ như sau:

Vì các khoảng tính mờ $\Im_{k_j}(x_{j,i})$ mức k_j của $X_{j,(k_j)}$ tạo thành một phân hoạch nhị phân của không gian được chuẩn hóa của thuộc tính thứ j là [0, 1] nên chỉ có duy nhất một khoảng tính mờ $\Im_{k_j}(x_{j,i(j)})$ ứng với hạng từ $x_{j,i(j)}$ mức k_j chứa thành phần dữ liệu $d_{j,l}$ với $j = 1, ..., n$ của d_l. Tập các khoảng tính mờ chứa thành phần dữ liệu $d_{j,l}$ xác định một khối hộp $\mathcal{H}_l$ chứa mẫu dữ liệu d_l. Khối hộp $\mathcal{H}_l$ cùng với lớp kết luận C_l của p_l xác định luật mờ cơ sở độ dài n có dạng sau:

$$\text{IF } \mathcal{X}_1 \text{ is } x_{1,i(1)} \text{ AND ... AND } \mathcal{X}_n \text{ is } x_{n,i(n)} \text{ THEN } C_l \qquad (\mathrm{R_b})$$

Để cho gọn, luật trên được kí hiệu là $\boldsymbol{A}_l \Rightarrow C_l$, trong đó $\boldsymbol{A}_l = (A_{l,1}, ..., A_{l,n})$ với $A_{l,j} =$ "$\mathcal{X}_j$ is $x_{j,i(j)}$", $j = 1, ..., n$. Từ các luật cơ sở có độ dài n, các luật ứng viên có độ dài nhỏ hơn n được xây dựng bằng cách bỏ đi một số điều kiện tiền đề $A_{l,j}$ của luật cơ sở và có được luật dạng sau:

$$(A_{q,j_1}, ..., A_{q,j_{i(q)}}) \Longrightarrow C_q = \boldsymbol{A}_q \Rightarrow C_q \qquad (\mathrm{R_{cnd}})$$

trong đó $1 \leq j_1 \leq ... \leq j_{i(q)} \leq n$, còn phần kết luận của luật là lớp C_q được chọn từ các nhãn lớp có độ tin cậy của luật $\boldsymbol{A}_q \Rightarrow C_q$ là lớn nhất, nghĩa là C_q được tính theo công thức sau

$$C_q = argmax\{c(\boldsymbol{A}_q \Rightarrow C_h) | h = 1, ..., M\}, \qquad (3.3)$$

trong đó $c(\boldsymbol{A}_q \Rightarrow C_h)$ là độ tin cậy của luật $\boldsymbol{A}_q \Rightarrow C_h$, tức là ta có [7, 8, 11]

$$c(\boldsymbol{A}_q \Rightarrow C_h) = \sum_{d_p \in C_h} \mu_{A_q}(d_p) \Big/ \sum_{p=1}^{m} \mu_{\boldsymbol{A}_q}(d_p), \qquad (3.4)$$

với $\mu_{\boldsymbol{A}_q}(d_p)$ là độ đốt cháy của mỗi mẫu dữ liệu d_p đối với luật R_q, tức là được tính bằng biểu thức toán tử nhân (product operator) sau

$$\mu_{\boldsymbol{A}_q}(d_p) = \Pi_{i=1}^{i(q)} \mu_{A_{q,j_i},j}(d_{p,j_i}). \qquad (3.5)$$

Bước 3. Chọn lọc tập luật khởi sinh $\boldsymbol{S}_0$ từ tập luật ứng viên sử dụng một tiêu chuẩn sàng. Sau Bước 2 ta thu được tập luật $\boldsymbol{S}_0$ và chỉ giữ lại NR_0 luật từ tập luật này với NB_0 là tham số điều chỉnh. Tập luật $\boldsymbol{S}_0$ được phân nhóm thành M nhóm theo các nhãn lớp ở phần kết luận của luật. Các luật được lựa chọn theo tiêu chuẩn sàng, tức là sắp xếp các luật giảm dần

trong mỗi nhóm theo tiêu chuẩn sàng và chọn ra NB_0 luật trong mỗi nhóm từ trên xuống dưới. Tiêu chuẩn sàng được sử dụng phổ biến là tích $(c.s)$ của độ tin cậy (c) và độ hỗ trợ (s). Độ tin cậy được tính theo công thức (3.4), độ hỗ trợ được tính theo công thức sau [7, 8, 11]

$$s(\boldsymbol{A}_q \Rightarrow C_h) = \sum_{d_p \in C_h} \mu_{A_q}(d_p)/n. \tag{3.6}$$

Như vậy, sau Bước 3 ta thu được hệ luật khởi sinh $\boldsymbol{S}_0$ có $NR_0 = NB_0 * M$ luật. Trong quá trình lập luận, các luật được gán một trọng số (rule weight). Theo [7, 8, 11], trọng số luật được tính bằng công thức

$$CF(\boldsymbol{A}_q \Rightarrow C_q) = c_q - c_{q,2nd}, \tag{3.7}$$

trong đó, c_q là độ tin cậy của luật R_q và $c_{q,2nd}$ là độ tin cậy lớn nhất của các luật có cùng tiền đề điều kiện $\boldsymbol{A}_q$nhưng kết luận là lớp khác với C_q

$$c_{q,2nd} = \max\{c(\boldsymbol{A}_q \Rightarrow C_h) | h = 1, ..., M; C_h \neq C_q\}.$$

So với thủ tục xây dựng tập luật khởi sinh trong [11], tại Bước 1 các ánh xạ định lượng điểm truyền thống được thay thế bằng các ánh xạ định lượng khoảng, các tập mờ hình tam giác được thay thế bằng các tập mờ thang. Tại Bước 2, các khoảng tính mờ mức k_j được sử dụng làm phân hoạch trên miền các thuộc tính thay cho các khoảng tương tự mức k_j.

Độ phức tạp của thủ tục xây dựng các luật khởi sinh là đa thức đối với kích thước và số thuộc tính của tập dữ liệu mẫu $\boldsymbol{D}$ như đã được chứng minh trong [11].

3.3. Tối ưu hóa tham số ngữ nghĩa của các hạng từ ngôn ngữ và tối ưu hóa hệ luật

Kết quả của bất kỳ phương pháp tiếp cận theo hệ mờ dựa trên luật nào cũng phụ thuộc vào các tham số mờ. Với phương pháp thiết kế HLPM dựa trên hệ luật mờ theo tiếp cận lý thuyết tập mờ [6–9], các tác giả sử dụng các chiến lược tìm kiếm tối ưu các tham số mờ, cụ thể là hiệu chỉnh các tham số của tập mờ dạng tam giác bằng giải thuật di truyền. Với phương pháp thiết kế HLPM dạng luật theo tiếp cận ĐSGT [11], các tác giả đã phân tích sự phụ thuộc của kết quả phân lớp của hệ luật vào các tham số mờ gia tử (tham số ngữ nghĩa) và sử dung giải thuật di truyền lai SGA với việc đánh trọng số các hàm mục tiêu để hiệu chỉnh các tham số này cho từng tập dữ liệu mẫu cụ thể. Sau khi có được bộ tham số mờ gần tối ưu, tiến hành tạo sinh hệ luật khởi sinh làm đầu vào cho thủ tục tối ưu hóa hệ luật mờ do các tiêu chuẩn sàng được áp dụng không cho kết quả tốt. Với phương pháp lượng hóa ĐSGT mở rộng, bài báo sử dụng giải thuật tối ưu hóa bầy đàn đa mục tiêu PSO (Particle Swarm Optimization) với hàm thích nghi chia sẻ [12], định hướng việc tìm kiếm các cá thể trong tối ưu toàn cục (mặt *Pareto* hay *Pareto front*), để tìm kiếm tối ưu các tham số ngữ nghĩa và tối ưu hóa hệ luật. Nhờ đó các mục tiêu tối ưu được chia sẻ bình đẳng thông qua hệ số chia sẻ và kết quả là một tập các phương án.

Tập các tham số cần được hiệu chỉnh thích nghi đối với bài toán thiết hệ phân lớp mờ $\mathcal{P}$ dựa trên phương pháp lượng hóa ĐSGT mở rộng là các tham số ngữ nghĩa trong $\mathcal{P}$. Chúng bao gồm các độ đo tính mờ $fm_j(c^-)$ của hạng từ sinh c^- và của hằng $fm_j(W)$; các độ đo tính mờ của các gia tử của thuộc tính j và tham số k_j nguyên dương hạn chế độ dài của các từ ngôn ngữ được thiết kế của thuộc tính j. Có thể thấy so với phương pháp tối ưu hóa trong

[11], phương pháp được đề nghị ở đây có nhiều hơn hai tham số hiệu chỉnh thích nghi cho mỗi thuộc tính j: tham số độ đo tính mờ của W_j và tham số độ đo tính mờ của gia tử $h_{j,0}$.

Để hiệu chỉnh thích nghi các tham số ngữ nghĩa được nêu ở trên cho phù hợp với từng tập dữ liệu mẫu, bài toán tiến hóa tối ưu hóa đa mục tiêu thiết kế các từ ngôn ngữ tối ưu cho bài toán phân lớp $\mathcal{P}$ được đặt ra là [11] với $\mathcal{V}(\mathbf{\Pi}, \boldsymbol{D}, N_0)$ là thủ thục xây dựng hệ luật khởi sinh và với các ràng buộc về các tham số ngữ nghĩa đã được nêu ở trên. Khi đó, mục tiêu của bài toán đặt ra là tối đa hóa hiệu quả phân lớp và tối thiểu hóa độ dài trung bình các luật của $\mathcal{P}$.

Như đã được đề cập ở trên, giải thuật PSO được sử dụng để tìm kiếm giá trị tối ưu của các tham số ngữ nghĩa cho bài toán phân lớp mờ cụ thể $\mathcal{P}$ với cơ sở luật $\boldsymbol{S}$. Thủ tục tối ưu hóa các tham số nghĩa được đặt tên là **MPSO_SPO**. Các mục tiêu của bài toán tối ưu hóa là

$$\text{maximize } perf(\boldsymbol{S}) \text{ và maximize avg}(\boldsymbol{S})^{-1} \text{ với ràng buộc } \boldsymbol{S} \subset \boldsymbol{S}_0. \tag{3.8}$$

trong đó $perf(\boldsymbol{S})$ là tỷ lệ phân lớp đúng của hệ $\boldsymbol{S}$ trên tập mẫu huấn luyện, avg$(S)^{-1}$ là nghịch đảo của độ dài luật trung bình của hệ $\boldsymbol{S}$. Số luật dùng để tối ưu hóa các tham số ngữ nghĩa được cố định trước theo từng tập dữ liệu mẫu cụ thể nên không nằm trong các mục tiêu cần tối ưu hóa.

Sau quá trình tối ưu hóa tham số các tham số ngữ nghĩa bằng giải thuật **MPSO_SPO** ta thu được các tham số gần tối ưu $\mathbf{\Pi}_{opt}$. Sử dụng thủ tục xây dựng luật khởi sinh để sinh tập luật khởi sinh $\boldsymbol{S}_0$ với N R_0 luật sử dụng các tham số $\mathbf{\Pi}_{opt}$. Có thể sử dụng các tiêu chuẩn sàng để lựa chọn một tập luật cho HPLM dạng luật từ tập luật $\boldsymbol{S}_0$. Tuy nhiên, việc sử dụng các tiêu chuẩn sàng có thể cho tập luật không đạt được các mục tiêu (3.8) tốt. Bài toán đặt ra là phải chọn ra một tập luật con của S_0 cho HPLM sao cho đạt các mục tiêu sau

$$\begin{gathered}\text{maximize } perf(\boldsymbol{S}),\ \text{maximize } NR(\boldsymbol{S})^{-1} \text{ và maximize avg}(\boldsymbol{S})^{-1} \text{ với ràng buộc} \\ \boldsymbol{S} \subset \boldsymbol{S}_0,\ NR(\boldsymbol{S}) \leq N_{\max}.\end{gathered} \tag{3.9}$$

trong đó $NR(\boldsymbol{S})^{-1}$ là nghịch đảo của số luật trung bình và $N_{\max}$ là số luật chọn tối đa và được cho trước. Ta gọi bài toán này là bài toán tối ưu hóa hệ luật. Tiếp tục sử dụng giải thuật tiến hóa đa mục tiêu PSO cho bài toán tối ưu hóa hệ luật với ba hàm mục tiêu cụ thể trong (3.9). Thủ tục tối ưu hóa hệ luật được đặt tên là **MPSO_RBO**.

Để lựa chọn các tập luật con từ $\boldsymbol{S}_0$ cho việc sinh các cá thể cho giải thuật **MPSO_RBO**, phương pháp mã hóa số thực được sử dụng. Mỗi cá thể ứng với mỗi lời giải là một tập luật $\boldsymbol{S}$ được chọn từ $\boldsymbol{S}_0$ và được biểu diễn bởi một chuỗi số thực $\boldsymbol{r}_i = (p_1, ..., p_{N_{\max}}),\ p_j \in [0, 1]$. Giá trị p_j xác định chỉ số của luật trong $\boldsymbol{S}_0$ được chọn cho $\boldsymbol{S}$ có giá trị là $p_j \times |\boldsymbol{S}_0|$, ta có $0 \leq p_j \times |\boldsymbol{S}_0| < |\boldsymbol{S}_0|$.

$$\boldsymbol{S} = \{R_i \in \boldsymbol{S}_0 | i = \lfloor p_j \times |\boldsymbol{S}_0| \rfloor,\ i \geq 0\} \tag{3.10}$$

trong đó $\lfloor \bullet \rfloor$ là phép lấy phần nguyên.

Do giải thuật tối ưu hóa hệ luật cho một tập dữ liệu mẫu cụ thể có áp dụng phương pháp tìm kiếm tối ưu Pareto nên cho kết quả là một tập các phương án. Từ tập các phương án tìm được ta chọn ra một phương án với tập luật cho kết quả phân lớp trên tập huấn luyện cao nhất. Nếu có nhiều phương án giống nhau thì chọn ngẫu nhiên một phương án.

4. KẾT QUẢ THỰC NGHIỆM VÀ ĐÁNH GIÁ

Mục này trình bày các kết quả thực nghiệm của hệ phân lớp đối với một số tập dữ liệu mẫu chuẩn được cung cấp bởi KEEL-Dataset repository (http://sci2s.ugr.es/keel/datasets.php).

Mỗi tập dữ liệu mẫu được chia thành 10 phần bằng nhau sẵn có từ liên kết trên. Tiến hành lấy lần lượt từng phần để kiểm tra (tập kiểm tra), toàn bộ 9 phần còn lại được dùng để sinh luật (tập huấn luyện). Mỗi tập dữ liệu mẫu được chạy thử nghiệm 3 lần 10-folds theo cách như trên, ta có kết quả của 30 lần chạy. Kết quả cuối cùng của các lần thử nghiệm sau khi tối ưu hóa hệ luật được tính trung bình đối với số luật #R, độ phức tạp của hệ luật #C, tỷ lệ phân lớp đúng trên tập huấn luyện P_{tr} và trên tập kiểm tra P_{te}. Độ phức tạp của hệ luật được tính theo công thức $\#C = \#R \times Avg$, trong đó Avg là độ dài trung bình của hệ luật.

Trong bài toán thực nghiệm, số gia tử âm và gia tử dương đều được lấy là 1. Giả sử gia tử âm là L và gia tử dương là V. Các ràng buộc của các tham số ngữ nghĩa bao gồm: giới hạn độ dài của các hạng từ $k_j \leq 3; 0,2 \leq fm_j(c^-) \leq 0,7; 0,0001 \leq fm_j(W) \leq 0,2; fm_j(c^-) + fm_j(W) + fm_j(c^+) = 1; 0,2 \leq \mu_j(L) \leq 0,7; 0,0001 \leq \mu_j(h_0) \leq 0,5$. Giá trị của các tham số này được chọn là kết quả của nhiều lần thực nghiệm với các giá trị khác nhau của các tham số đối với một số tập dữ liệu mẫu trên cơ sở đánh giá sự cân bằng giữa hiệu quả phân lớp và sự cân đối giữa miền giá trị của $fm_j(c^-)$ và $fm_j(c^+)$.

Các tham số cho thuật toán **MPSO_SPO** gồm: số thế hệ tối đa: 250; số cá thể mỗi thế hệ: 600; Hệ số Inertia: 0,4; hệ số nhận thức cá nhân 0,2; hệ nhận thức xã hội: 0,2; Số luật khởi sinh bằng số thuộc tính; giới hạn độ dài của luật là 1; Thực hiện toán tử đột biến khi tỷ lệ di chuyển của các cá thể nhỏ hơn 70%.

Các tham số cho thuật toán **MPSO_RBO** như sau: Số thế hệ tối đa: 1000; số cá thể mỗi thế hệ: 600; Hệ số Inertia: 0,4; hệ số nhận thức cá nhân 0,1; hệ số nhận thức xã hội: 0,1; Thực hiện toán tử đột biến khi tỷ lệ di chuyển của các cá thể nhỏ hơn 70%; Hệ số chia sẻ các hàm thích nghi được tính tự động; Số luật khởi đầu $|\boldsymbol{S}_0| = 300\times$ số lớp; giới hạn độ dài của luật đối với tập dữ liệu mẫu có số thuộc tính lớn hơn 30 là 2, ngược lại là 3.

Phương pháp lập luận được sử dụng là *Single Winner Rule* [8].

Kết quả chạy thực nghiệm của phương pháp được đề xuất và so sánh với kết quả của phương pháp lượng hóa truyền thống [11] được thể hiện trong Bảng 1.

Bảng 1. Kết quả chạy $3\times10-folds$ trên 10 tập mẫu được thử nghiệm

Tập dữ liệu mẫu		**Lượng hóa sử dụng ngữ nghĩa hình thang**				**Lượng hóa sử dụng ngữ nghĩa tam giác**				**So sánh**	
Stt	Tên	$\#R$	$\#R\times\#C$	P_{tr}	P_{te}	$\#R$	$\#R\times\#C$	P_{tr}	P_{te}	$\neq P_{te}$	$\neq\#R\times\#C$
1	Bands	7,00	**78,17**	76,28	**72,10**	6,00	**83,40**	75,57	**70,63**	*1,47*	*-5,23*
2	Bupa	8,97	**170,65**	77,54	**69,41**	8,97	**196,37**	77,40	**67,71**	*1,7*	*-25,72*
3	Dermatology	10,87	**189,45**	96,88	**95,52**	10,93	**194,61**	98,82	**95,52**	*0*	*-5,16*
4	Haberman	4,00	**20,00**	77,67	**77,43**	3,00	**13,30**	76,78	**75,11**	*2,32*	*6,7*
5	Pima	5,97	**50,32**	78,53	**76,66**	5,00	**51,17**	79,03	**75,70**	*0,96*	*-0,85*
6	Sonar	5.97	**53,89**	86,84	**77,29**	7,00	**84,00**	88,59	**76,73**	*0,56*	*-30.11*
7	Vehicle	11,03	**216,26**	71,64	**68,12**	11,93	**324,98**	70,59	**67,46**	*0,66*	*-108,72*
8	Wdbc	4,97	**41,56**	97,40	**95,85**	4,97	**45,86**	96,51	**94,90**	*0,95*	*-4,3*
9	Wine	5,87	**42,06**	100,0	**98,52**	5,73	**65,17**	99,79	**98,30**	*0,22*	*-23,11*
10	Wisconsin	6,93	**57,55**	96,74	**96,45**	5,97	67,42	98,38	**96,72**	-0,27	*-9,87*

Trong Bảng 1, cột gộp "So sánh" gồm cột "$\neq P_{te}$" thể hiện phần trăm tăng hiệu quả phân lớp của phương pháp được đề xuất so với phương pháp lượng hóa truyền thống đối với từng tập dữ liệu mẫu, cột "$\neq \#R \times \#C$" thể hiện sự chênh lệch độ phức tạp của hệ luật. Kết quả cho thấy, phương pháp được đề xuất cho kết quả phân lớp cao hơn đối với hầu hết các tập

dữ liệu mẫu được thực nghiệm, đồng thời có độ phức tạp thấp hơn rất nhiều so với phương pháp lượng hóa truyền thống [11]. Cụ thể, hiệu quả phân lớp tăng 8.57% và độ phức tạp của hệ luật giảm 18,32%.

Bảng 2. So sánh hiệu suất phân lớp của HPLM sử dụng Wilcoxon với mức $\alpha = 0,05$

VS	R^+	R^-	Exact *P*-value	Asymp. *P*-value	Confidence interval	Exact confidence
Lượng hóa sử dụng ngữ nghĩa tam giác	43,0	2,0	0,011718	0,012851	[0,33 , 1,47]	0,95118

Bảng 3. So sánh độ phức tạp của hệ luật sử dụng Wilcoxon với mức $\alpha = 0,05$

VS	R^+	R^-	Exact *P*-value	Asymp. *P*-value	Confidence interval	Exact confidence
Lượng hóa sử dụng ngữ nghĩa tam giác	50,0	5,0	0,019532	0,019059	[-54,785 , -3,005]	0,95118

Thực hiện phương pháp kiểm tra Wilcoxon Signed Rank [13, 14] sử dụng dữ liệu trong Bảng 1 để so sánh kết quả của hai phương pháp. Các kết quả kiểm tra hiệu suất phân lớp và độ phức tạp của hệ luật lần lượt được thể hiện trong Bảng 2 và Bảng 3. Với giá trị R^- là tổng các xếp hạng ứng với các hiệu quả phân lớp của phương pháp lượng hóa ĐSGT mở rộng nhỏ hơn giá trị găng (critical value) ứng với số tập dữ liệu mẫu $N_{ds} = 10$ và $p = 0,05$ bằng 8 (có thể tham khảo trong bảng phân phối T Wilcoxon (bảng B.12 trong [14])) nên ta có thể khẳng định rằng phương pháp thiết kế HPLM dạng luật với phương pháp lượng hóa ĐSGT mở rộng không những cho hiệu quả tốt phân lớp hơn mà còn có độ phức tạp của hệ luật thu được giảm đáng kể so với phương pháp lượng hóa ĐSGT truyền thống.

5. KẾT LUẬN

Bài báo đề xuất và phát triển phương pháp thiết kế HPLM dạng luật dựa trên ĐSGT mở rộng được đề xuất trong [10] và tiến hành nghiên cứu thử nghiệm phương pháp thiết kế đối với một số tập dữ liệu mẫu chuẩn được cung cấp bởi KEEL-Dataset repository. Phương pháp lượng hóa ĐSGT mở rộng cho phép xây dựng các phân hoạch trên miền các thuộc tính dựa trên chính các khoảng tính mờ mức k và cho phép định lượng ngữ nghĩa lõi của các từ ngôn ngữ dưới dạng khoảng. Đây chính là cơ sở cho phép xây dựng các tập mờ hình thang với đáy nhỏ là giá trị định lượng khoảng của các từ ngôn ngữ. So với tập mờ hình tam giác thì tập mờ hình thang có miền tin cậy rộng hơn và hai cạnh bên có độ dốc lớn hơn nên cho tỷ lệ mất mát thông tin ít hơn. Kết quả thử nghiệm trên 10 tập dữ liệu mẫu cho hiệu quả phân lớp của HPLM được đề xuất cao hơn với độ phức tạp nhỏ hơn so với HPLM sử dụng phương pháp lượng hóa truyền thống. Điều đó chứng tỏ rằng phương pháp được đề xuất không chỉ có một cơ sở lí thuyết chặt chẽ mà còn hứa hẹn tạo ra một khả năng ứng dụng tiềm năng.

TÀI LIỆU THAM KHẢO

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

[2] Nguyen Cat Ho, Tran Thai Son, Tran Dinh Khang, Le Xuan Viet, Fuzziness measure, quantified semantic mapping and interpolative method of approximate reasoning in medical expert systems, *Journal of Computer Science and Cybernetics* **18** (3) (2002) 237–252.

[3] Nguyễn Cát Hồ, Trần Thái Sơn, Dương Thăng Long, Tiếp cận đại số gia tử cho phân lớp mờ, *Tạp chí Tin học và Điều khiển học* **25** (1) 2009 53–68.

[4] Dương Thăng Long, Một phương pháp xây dựng hệ luật mờ có trọng số để phân lớp dựa trên đại số gia tử, *Tạp chí Tin học và Điều khiển học* **26** (1) (2010) 55–72.

[5] A. Fernandez, M. Calderón, E. Barrenechea, H. Bustince F. Herrera, Enhancing fuzzy rule based systems in multi-classification using pairwise coupling with preference relations, *EUROFUSE09 Workshop on Preference Modelling and Decision Analysis Pamplona*, Spain, September 16-18, 2009 (39–46).

[6] Chen Ji-lin, Hou Yuan-long, Xing Zong-y, Jia Li-min, Tong Zhong-zhi, A multi-objective genetic-based method for design fuzzy classification systems, *IJCSNS International Journal of Computer Science and Network Security* **6** (8A) (August 2006) 110–118.

[7] H. Ishibuchi and T. Yamamoto, Rule weight specification in fuzzy rule-based classification systems, *IEEE Trans. on Fuzzy Systems* **13** (4) (2005) 428–435.

[8] H. Ishibuchi, T. Yamamoto, Fuzzy rule selection by multi-objective genetic local search algorithms and rule evaluation measures in data mining, *Fuzzy Sets and Systems* **141** (1) (2004) 59–88.

[9] Rafael Alcalá, Yusuke Nojima, Francisco Herrera, Hisao Ishibuchi, Multi-objective genetic fuzzy rule selection of single granularity-based fuzzy classication rules and its interaction with the lateral tuning of membership functions, *Journal Soft Computing* **15** (12) (December 2011) 2303–2318.

[10] Nguyễn Cát Hồ, Trần Thái Sơn, Phạm Đình Phong, Định lượng ngữ nghĩa khoảng của Đại số gia tử với việc bổ sung một gia tử đặc biệt, *Tạp chí Tin học và Điều khiển học* **28** (4) (2012) 346–358.

[11] Cat Ho Nguyen, Witold Pedrycz, Thang Long Duong, Thai Son Tran, A genetic design of linguistic terms for fuzzy rule based classifiers, *International Journal of Approximate Reasoning, Elsevier Science Inc* **54** (1) (January 2013) 1–21.

[12] Maximino Salazar Lechuga, "Multi-Objective Optimisation using Sharing in Swarm Optimisation Algorithms", Doctor thesis, School of Computer Science, The University of Birmingham, 2006.

[13] Janez Demsar, Statistical comparisons of classifiers over multiple data sets, *Journal of Machine Learning Research* **7** (2006) 1–30.

[14] J. Zar, *Biostatistical Analysis*, Prentice-Hall, Upper Saddle River, NJ, 1999.

Ngày nhận bài 21 - 3 - 2013
Nhận lại sau sửa ngày 27 - 11 - 2013