Tạp chí Khoa học Đại học Huế: Kỹ thuật và Công nghệ; pISSN 2588-1175 | eISSN 2615-9732

Tập 134, Số 2A, 2025, Tr. 75–84; DOI: 10.26459/hueunijtt.v134i2A.7816

MỘT CÁCH TIẾP CẬN KHAI PHÁ LUẬT KẾT HỢP MỜ SỬ DỤNG ĐẠI SỐ GIA TỬ

Nguyễn Công Hào*

Ban Kiểm tra và Pháp chế Đại học Huế, Huế, Việt Nam

Tóm tắt. Luật kết hợp mờ theo cách tiếp cận lý thuyết tập mờ đã được nhiều tác giả quan tâm nghiên cứu theo và đã công bố nhiều kết quả đáng kể, tạo sự mềm dẽo và linh hoạt trong khai phá dữ liệu với thông tin không chắc chắn. Tuy nhiên, cách tiếp cận sử dụng lý thuyết mờ cho bài toán khai phá luật kết hợp mờ vẫn còn một số hạn chế nhất định. Trong bài báo này, chúng tôi đề xuất một phương pháp khai phá luật kết hợp mờ sử dụng Đại số gia tử (ĐSGT) với mỗi giá trị ngôn ngữ được biểu bởi một khoảng lân cận. Với ưu điểm sử dụng các hàm đo và định lượng ngữ nghĩa của ĐSGT, vấn đề khai phá các luật kết hợp mờ và tính toán khá đơn giản và trực quan. Kết quả thu được sau khi trích rút các luật kết hợp mờ trên bộ dữ liệu khảo sát của sinh viên năm thứ nhất trong việc lựa chọn trường đại học là một kênh cung cấp thông tin quan trọng cho lãnh đạo các trường đại học trong việc định hướng truyền thông công tác tuyển sinh.

Từ khóa: Trường đại học, đại số gia tử, khai phá dữ liệu mờ, luật kết hợp mờ, truyền thông marketing

An approach for mining fuzzy association rules based on hedge algebra

Nguyen Cong Hao*

Department of Inspection and Legal Affairs, Hue University, Hue, Vietnam

Abstract. Fuzzy association rules based on fuzzy set theory have been researched by many authors and have published many significant results and flexibility in data mining with uncertain information. However, the approach using fuzzy set theory for the problem of mining fuzzy association rules still has certain limitations. In this paper, we propose a method for mining fuzzy association rules based on hedge algebra with each linguistic value represented by their neighborhood. The hedge algebra has advantage of using the measure functions and semantic quantifier functions, the problem of mining fuzzy association rules and calculating them is quite simple and intuitive. The results obtained after extracting fuzzy association rules on the survey dataset of freshman in choosing a university are a channel providing useful information for university managers in enrollment communication.

** Liên hệ: nchao@hueuni.edu.vn*

Ngày nhận bài: 10–04–2025; Ngày nhận đăng: 02–06–2025

Keywords: university, hedge algebra, fuzzy data mining, fuzzy association rule, marketing communications

1 Giới thiệu

Khai phá dữ liệu là một quá trình quan trọng, sử dụng các kỹ thuật để tự động tìm kiếm các mẫu, mối quan hệ và thông tin hữu ích từ các bộ dữ liệu lớn. Nó biến đổi dữ liệu thô thành kiến thức giá trị, hỗ trợ đưa ra quyết định tối ưu hơn trong nhiều lĩnh vực. Mục tiêu chính là khám phá các thông tin ẩn giấu, dự đoán xu hướng và cải thiện hiệu quả hoạt động. Có nhiều kỹ thuật khai phá dữ liệu như: phân cụm (clustering), phân loại (classification), luật kết hợp (association rules), dự đoán (prediction), phân tích chuỗi thời gian (time series analysis), khám phá dữ liệu văn bản (text mining). Trong đó, kỹ thuật khai phá luật kết hợp là một kỹ thuật khá đơn giản và có nhiều ứng dụng. Tuy nhiên, khi xử lý dữ liệu không chắc chắn, dữ liệu mờ (fuzzy data) cần có kỹ thuật mở rộng luật kết hợp để đáp ứng quá trình khai phá.

Luật kết hợp mờ (fuzzy association rules) là một sự mở rộng của luật kết hợp, áp dụng lý thuyết tập mờ để xử lý dữ liệu không chắc chắn, không rõ ràng. Trong khi luật kết hợp chỉ sử dụng giá trị nhị phân (có hoặc không hay 1 hoặc 0), luật kết hợp mờ cho phép thao tác với các giá trị thuộc về nhiều mức độ khác nhau (độ thuộc là một giá trị mờ thuộc [0,1]), từ đó mô tả tốt hơn các mối quan hệ phức tạp trong dữ liệu. Các tác giả đã đề xuất thuật toán luật kết hợp mờ để dự đoán bệnh nhân ung thư vú. Kết quả thực nghiệm cho thấy phương pháp này hiệu quả hơn các phương pháp thông thường [1]. Một phương pháp khác khai phá luật kết hợp mờ để xác định mối liên kết giữa số lượng bán hàng và sản phẩm thương mại điện tử với hình dáng hàm thuộc là tam giác [2]. Một cách tiếp cận khai phá luật kết hợp mờ với việc tối ưu hóa thời gian và lưu trữ với hàm thuộc dạng tam giác và hình thang [5]. Khai phá luật kết hợp mờ cho hiệu suất nhanh và hiệu quả trong tập dữ liệu lớn và phát hiện mô hình tội phạm cộng đồng [3,4]. Khai phát luật kết hợp mờ dựa vào tối ưu hóa vùng mờ và khai phá tập mục mờ song song [7,8].

Mặc dù đã có nhiều kết quả nghiên cứu khai phá luật kết hợp mờ và ứng dụng khá rộng rãi trong thực tiễn, nhưng vẫn còn một số hạn chế do việc xây dựng các hàm thuộc của các tập mờ, phương pháp mờ hóa, khử mờ…làm ảnh hưởng không nhỏ và khả năng sai số kết quả. Để khắc phục những hạn chế này, cách tiếp cận khai phát luật kết hợp mờ sử dụng ĐSGT được các tác giả nghiên cứu và bước đầu công bố một số kết quả. Sử dụng luật kết hợp mờ từ giai đoạn mở rộng cây phân lớp với mỗi phần tử của ĐSGT là một vùng mờ [9]. Phương pháp khai phá luật kết hợp mờ sử dụng ĐSGT trong việc hỗ trợ sinh viên lập kế hoạch học tập, các cấu trúc tập mờ của các thuộc tính mờ được xây dựng dựa trên ĐSGT và có dạng đa hạt thể, cách tiếp cận này đơn giản hơn so với cách tiếp cận lý thuyết tập mờ [11].

Đối với vấn đề tuyển sinh nói chung và quảng bá tuyển sinh nói riêng là những nội dung rất quan trọng của một cơ sở giáo dục đại học nhằm thu hút người học, tạo nguồn thu chính cho

đơn vị phát triển. Để nâng cao hiệu quả truyền thông tuyển sinh đại học, các cơ sở giáo dục đại học cần kết hợp đồng bộ nhiều giải pháp. Điều này bao gồm việc đổi mới nội dung truyền thông, sử dụng hiệu quả các phương tiện truyền thông, và đặc biệt là đẩy mạnh truyền thông trên mạng xã hội để tiếp cận đối tượng mục tiêu. Các nghiên cứu khảo sát theo nhiều cách tiếp cận khác nhau giúp hiểu rõ hơn về nhu cầu và mong muốn của người học.

Bài toán đặt ra là sử dụng kết quả khảo sát của sinh viên năm thứ nhất theo cách tiếp cận việc ảnh hưởng của truyền thông marketing đến việc lựa chọn cơ sở giáo dục đại học khi còn học sinh phổ thông. Từ đó, khai phá và rút trích ra các thông tin có ích để hỗ trợ cho người làm công tác truyền thông tuyển sinh cũng như nhà quản lý các cơ sở giáo dục đại học trong việc xây dựng và định hướng công tác tuyền thông tuyển sinh một cách hiệu quả.

Vì vậy, trong bài báo này, chúng tôi đề xuất cách tiếp cận mới khai phá dữ liệu sử dụng ĐSGT với mỗi giá trị ngôn ngữ được biểu diễn bởi một khoảng lân cận mức k nào đó ($k \in Z$). Với cách tiếp cận này, mỗi tập mờ (giá trị ngữ) được xem là một phần tử của ĐSGT và việc thao tác, tính toán được thực hiện trực tiếp trên ngôn ngữ. Cấu trúc bài báo được trình bày trong 4 phần, ngoài phần giới thiệu, phần 2 trình bày một số kiến thức cơ sở về đại số gia tử, luật kết hợp mờ, phần 3 trình bày cách tiếp cận luật kết hợp mờ sử dụng đại số gia tử, phần 4 trình bày kết quả thực nghiệm, kết luận và hướng nghiên cứu tiếp theo.

2 Một số kiến thức cơ sở

2.1 Đại số gia tử

Xét miền ngôn ngữ của biến chân lý *NHIỆT ĐỘ* gồm các giá trị ngôn ngữ như sau: *Dom*(*NHIỆT ĐỘ*) = {*nóng, lạnh, rất nóng, rất lạnh, ít_nhiều nóng, ít_nhiều lạnh, khả năng nóng, khả năng lạnh, ít nóng, ít lạnh, rất khả năng nóng, rất khả năng lạnh*...}, trong đó *nóng, lạnh* là các từ nguyên thuỷ, các từ nhấn như *rất*, *ít_nhiều*, *khả năng*, *ít* gọi là các gia tử (hedges). Khi đó, miền ngôn ngữ T = *Dom*(*NHIỆT ĐỘ*) có thể biểu thị như một đại số $\underline{X} = (\boldsymbol{X}, G, H, \leq)$, trong đó G là tập các từ nguyên thuỷ được xem là các phần tử sinh. $H = H^- \cup H^+$ với H^+ và H^- tương ứng là tập các gia tử dương, âm và được xem như là các phép toán một ngôi, quan hệ $\leq$ trên các giá trị ngôn ngữ (các khái niệm mờ) là quan hệ sắp thứ tự tuyến tính trên $\boldsymbol{X}$ cảm sinh từ ngữ nghĩa của ngôn ngữ. Ví dụ dựa trên ngữ nghĩa, các quan hệ thứ tự sau là đúng: *lạnh* $\leq$ *nóng*, *hơn nóng* $\leq$ *rất nóng* nhưng *rất lạnh* $\leq$ *hơn lạnh*, *khả năng nóng* $\leq$ *nóng* nhưng *lạnh* $\leq$ *khả năng lạnh*... Tập $\boldsymbol{X}$ được sinh ra từ G bởi các phép toán trong H. Như vậy, mỗi phần tử của $\boldsymbol{X}$ sẽ có dạng biểu diễn $x = h_n h_{n-1} \ldots h_1 c$, trong đó, $h_n, h_{n-1} \ldots h_1 \in H$ và $c \in G$.

Ký hiệu $H(x)$ là tập tất cả các phần tử được sinh ra từ một phần tử x. Nếu G có đúng hai từ *nguyên thuỷ mờ*, thì một được gọi là *phần tử sinh dương* ký hiệu là c^+, một gọi là *phần tử sinh âm*

ký hiệu là c^- và ta có $c^- < c^+$. Trong ví dụ nêu trên, *nóng* là phần tử sinh dương, còn *lạnh* là phần tử sinh âm.

Như vậy, một cách tổng quát, cho $\underline{X} = (\boldsymbol{X}, G, H, \leq)$ là một đại số gia tử với $G = \{0, c^-, W, c^+, 1\}$, $H = H^- \cup H^+$ với giả thiết $H^+ = \{h_1, h_2, .., h_p\}$, $H^- = \{h_{-q}, ..., h_{-1}\}$, $h_1 < h_2 < ... < h_p$ và $h_{-q} > h_{-q+1} ... > h_{-1}$ là dãy các gia tử. Trong đại số gia tử tuyến tính, chúng ta bổ sung thêm vào hai phép tính Σ và Φ với ngữ nghĩa là cận trên đúng và cận dưới đúng của tập $H(x)$, tức là $\Sigma x = supremum\ H(x)$ and $\Phi x = infimum\ H(x)$. Khi đó, đại số gia tử tuyến tính được gọi là đại số gia tử tuyến tính đầy đủ và được ký hiệu $\underline{X} = (\boldsymbol{X}, G, H, \Sigma, \Phi, \leq)$ [10].

Các hàm định lượng ngữ nghĩa (υ), hàm độ đo tính mờ (fm), hàm dấu (Sgn) và các tính chất của đại số gia tử có thể tham khảo trong công trình đã công bố [9-10].

2.2 Lân cận mức k

Khi định nghĩa lân cận mức k chúng ta mong muốn các giá trị đại diện như vậy phải là điểm trong của lân cận mức k. Giả thiết mỗi tập H^- và H^+ chứa ít nhất 2 gia tử, ta định nghĩa độ tương tự mức k như sau: Xét X_k là tập tất cả các phần tử độ dài k, dựa trên các khoảng mờ mức k và các khoảng mờ mức $k + 1$ chúng ta mô tả việc xây dựng một phân hoạch của miền [0, 1]:

Với $k = 1$, các khoảng mờ mức 1 gồm $I(c^-)$ và $I(c^+)$. Các khoảng mờ mức 2 trên khoảng $I(c^-)$ là $I(h_p c^-) = I(h_{p-1} c^-) = \ldots = I(h_2 c^-) = I(h_1 c^-) = \upsilon_A(c^-) = I(h_{-1} c^-) = I(h_{-2} c^-) = \ldots = I(h_{-q+1} c^-) = I(h_{-q} c^-)$. Khi đó, ta xây dựng phân hoạch về độ tương tự mức 1 gồm các lớp tương đương sau: $S(0) = I(h_p c^-)$, $S(c^-) = I(c^-) \setminus [I(h_{-q} c^-) \cup I(h_p c^-)]$; $S(W) = I(h_{-q} c^-) \cup I(h_{-q} c^+)$; tương tự ta có $S(c^+) = I(c^+) \setminus [I(h_{-q} c^+) \cup I(h_p c^+)]$ và $S(1) = I(h_p c^+)$.

Tương tự, với $k = 2$, ta có thể xây dựng phân hoạch các lớp tương tự mức 2. Bằng cách tương tự như vậy, ta có thể xây dựng các phân hoạch các lớp tương tự mức k bất kỳ. Tuy nhiên, trong thực tế ứng dụng, chúng ta có thể giới hạn các gia tử tác động liên tiếp lên các phần tử nguyên thủy c^- và c^+ là một số nguyên k nào đó. Các giá trị rõ và các giá trị mờ gọi là có độ tương tự mức k nếu các giá trị đại diện của chúng cùng nằm trong một lớp tương tự mức k.

Giả sử phân hoạch các lớp tương tự mức k là các khoảng $S(x_1)$, $S(x_2)$, …, $S(x_m)$. Khi đó, mỗi giá trị ngôn ngữ a chỉ và chỉ thuộc về một lớp tương tự, chẳng hạn đó là $S(x_i)$ và nó gọi là lân cận mức k của a và ký hiệu là $\Omega_k(a)$. Dựa trên khái niệm độ tương tự, quan hệ bằng nhau được định nghĩa như sau:

Định nghĩa 2.1. Cho $R = \{A_1, A_2, \ldots, A_n\}$ là lược đồ quan hệ gồm n thuộc tính, $r(R)$ là quan hệ xác định trên R, giả sử t và s là hai bộ dữ liệu thuộc quan hệ $r(R)$. Ta ký hiệu $t[A_i] =_k s[A_i]$ và được gọi là bằng nhau mức k, nếu một trong các điều kiện sau xảy ra:

a) Nếu $t[A_i]$ và $s[A_i]$ là giá trị rõ, thì $t[A_i] = s[A_i]$.

b) Nếu $t[A_i]$ là giá trị rõ, $s[A_i]$ là giá trị mờ, thì $t[A_i] \in \Omega_k(s[A_i])$.

c) Nếu $t[A_i]$ và $s[A_i]$ là giá trị mờ, thì $\Omega_k(t[A_i]) = \Omega_k(s[A_i])$.

Ví dụ 2.1. Xét lược đồ quan hệ R = {*MAGV, HOTEN, BAIBAOWoS, DETAI*} với ý nghĩa: Mã giảng viên (*MAGV*), Họ và tên giảng viên (*HOTEN*), Số bài báo đăng trên tạp chí quốc tế có uy tín (*BAIBAOWoS*), Số đề tài đã chủ trì (*DETAI*) là 2 thuộc tính nhận giá trị mờ. Trong đó $D_{BAIBAOWoS}$ = [0, 50] và D_{DETAI} = [0, 20]. $LD_{BAIBAOWoS}$ và LD_{DETAI} có cùng tập các giá trị ngôn ngữ với tập các phần tử sinh là G = {0, *thấp*, *W*, *cao*, 1} và tập các gia tử H^- = {*ít*, *khả năng*}, H^+ = {*hơn*, *rất*}. Mặc dù các thuộc tính mờ đang xét có cùng tập các giá trị ngôn ngữ, nhưng ngữ nghĩa định lượng của chúng khác nhau do $D_{SOBAIBAOWoS}$ *và* D_{DETAI} *khác nhau.*

a) ***Đối với thuộc tính BAIBAOWoS***

Cho $W = 0.7$, $\alpha = 0.4$ và $\beta = 0.6$. Ta có: $fm(cao) = 0.3$, $fm(thấp) = 0.7$, $\mu(khả\ năng) = 0.2$, $\mu(ít) = 0.2$, $\mu(hơn) = 0.3$ và $\mu(rất) = 0.3$. Ta phân hoạch đoạn [0, 50] thành 5 khoảng tương tự mức 1 là: $S(0)$, $S(thấp)$, $S(W)$, $S(cao)$ và $S(1)$. Ta có: $fm(rất\ cao) \times 50 = 0.3 \times 0.3 \times 50 = 4.5$. Vậy $S(1) \times 50 = (45.5, 50]$; $(fm(khả\ năng\ cao) + fm(hơn\ cao)) \times 50 = (0.2 \times 0.3 + 0.3 \times 0.3) \times 50 = 7.5$ và $S(cao) \times 50 = (38, 45.5]$; $(fm(ít\ thấp) + fm(ít\ cao)) \times 50 = (0.2 \times 0.7 + 0.2 \times 0.3) \times 50 = 10$ và $S(W) \times 50 = (28, 38]$; $(fm(khả\ năng\ thấp) + fm(hơn\ thấp)) \times 50 = (0.2 \times 0.7 + 0.3 \times 0.7) \times 50 = 17.5$ và $S(thấp) \times 50 = (10.5, 28]$, $S(0) \times 50 = [0, 10.5]$. Khi đó, ta có $30 =_1 cao$, vì $30 \in \Omega_1(cao) = S(cao)$ và $10 =_1 thấp$, vì $10 \in \Omega_1(thấp) = S(thấp)$.

b) ***Đối với thuộc tính DETAI***

Cho $W = 0.6$, $\alpha = 0.6$ và $\beta = 0.4$. Ta có: $fm(cao) = 0.4$, $fm(thấp) = 0.6$, $\mu(khả\ năng) = 0.3$, $\mu(ít) = 0.3$, $\mu(hơn) = 0.2$ và $\mu(rất) = 0.2$. Ta phân hoạch đoạn [0, 20] thành 5 khoảng tương tự mức 1 là: $S(0)$, $S(thấp)$, $S(W)$, $S(cao)$ và $S(1)$. Ta có: $fm(rất\ cao) \times 20 = 0.2 \times 0.4 \times 20 = 1.6$. Vậy $S(1) \times 20 = (18.4, 20]$; $(fm(khả\ năng\ cao) + fm(hơn\ cao)) \times 20 = (0.3 \times 0.4 + 0.2 \times 0.4) \times 20 = 4$ và $S(cao) \times 20 = (14, 18.4]$; $(fm(ít\ thấp) + fm(ít\ cao)) \times 20 = (0.3 \times 0.6 + 0.3 \times 0.4) \times 20 = 6$ và $S(W) \times 20 = (8, 14]$; $(fm(khả\ năng\ thấp) + fm(hơn\ thấp)) \times 20 = (0.3 \times 0.6 + 0.2 \times 0.6) \times 20 = 6$ và $S(thấp) \times 20 = (2, 8]$, $S(0) \times 20 = [0, 2]$. Khi đó, ta có $9 =_1 cao$, vì $15 \in \Omega_1(cao) = S(cao)$ và $4 =_1 thấp$, vì $4 \in \Omega_1(thấp) = S(thấp)$.

2.3 Luật kết hợp mờ

Cho D là một cơ sở dữ liệu giao dịch, $I = \{I1, I2, \ldots, In\}$ là tập n thuộc tính (tập mục), trong đó mỗi bản ghi T là một giao dịch và chứa các mục, $T \subseteq I$, một số tập mục nhận giá trị mờ.

Định nghĩa 2.2: Cho $X, Y \neq \varnothing$ và $X \cap Y = \varnothing$. Một luật kết hợp mờ là một quan hệ có dạng $X \rightarrow Y$, trong đó $X, Y \subset I$ là các tập mục nhận giá trị mờ.

Định nghĩa 2.3: *Độ hỗ trợ* (support) của luật kết hợp mờ $X \rightarrow Y$ là tỷ lệ phần trăm các giao dịch chứa cả X và Y với tổng số các giao dịch có trong cơ sở dữ liệu.

Định nghĩa 2.4: *Độ tin cậy* (confidence) của luật kết hợp mờ $X \rightarrow Y$ là tỷ lệ phần trăm của số giao dịch có chứa chứa cả X và Y với số giao dịch có chứa X trong cơ sở dữ liệu.

3 Luật kết hợp mờ sử dụng đại số gia tử

Thuật toán khai phá luật kết hợp mờ sử dụng ĐSGT được mở rộng trên cơ sở thuật toán khai phá luật kết hợp và khắc phục việc tính toán, xây dựng các hàm thuộc trong khai phá luật kết hợp mờ sử dụng lý thuyết tập mờ. Mỗi phần tử của ĐSGT được biểu diễn bằng một khoảng tương tự mức *k* và việc đối sánh giữa các giá trị thực hiện như trong mục 2.2.

Vào:

CSDL D với tập thuộc tính I và tập giao dịch T, ngưỡng minsupp, minconf.

Cấu trúc ĐSGT cho các thuộc tính mờ trong I.

Ra: *Tập các luật kết hợp mờ thỏa mãn ngưỡng minsupp, minconf*

Phương pháp:

(1) **begin**

(2) $(D_F, I_F, T_F) = \text{FuzzyConvert}(D, I, T)$

(3) $F_1 = \text{Counting}(D_F, I_F, T_F, minsupp)$

(4) $m = 2$

(5) **while** $(F_{m-1} \neq \varnothing)$

(6) {

(7) $C_m = \text{Join}(F_{m-1})$

(8) $C_m = \text{Prune}(C_m)$

(9) $F_m = \text{Cheking}(C_m, D_F, minsupp)$

(10) $F = F \cup F_m$

(11) $m = m + 1$

(12) }

(13) Fuzzy_Generate_Rules(F, *minconf*)

(14) **end**

Độ phức tạp thuật toán là đa thức đối với số giao dịch *T* nhưng hàm mũ theo số thuộc tính của *I*. Do đó, để tối ưu hóa thuật toán, chúng tôi giới hạn tối đa số thuộc tính mờ trong luật mờ sinh ra, tăng ngưỡng độ hỗ trợ tối thiểu và sử dụng các khoảng lân cận trong đại số gia tử đối sánh để giảm số lượng tổ hợp.

Hàm FuzzyConvert(D, I, T): Thực hiện chuyển đổi từ CSDL D ban đầu sang CSDL D_F với mỗi miền trị của mỗi thuộc tính mờ là một ĐSGT. Việc xây dựng các khoảng tương tự được thực hiện thông qua các hàm đo trong ĐSGT. Chẳng hạn miền trị thuộc tính **Số lượng sinh viên** được mờ hóa là một ĐSGT có hai phần tử sinh "*thấp*", "*cao*" nên biểu diễn 5 khoảng tương tự ở mức 1 là S(0), S(*thấp*), S(W), S(*cao), S(1)*. Cho các gia tử H^- ={*ít*, *khả năng*}, H^+ ={*hơn*, *rất*} và các giá trị cụ thể của W, α và β sẽ tính được các khoảng tương tự mức 1, các khoảng này chứa trong [0,1] và nhân với độ dài của miền giá trị thực tương ứng của miền ngôn ngữ sẽ trở thành giá trị trong thực tế (chẳng hạn S(0)×|$D_{Soluongsv}$|…). Tiếp tục cho các gia tử tác động vào các phần tử sinh âm và dương sẽ xây dựng được các khoảng tương tự mức 2, mức 3… Về lý thuyết có xây dựng các khoảng tương tự mức k bất kỳ. Tuy nhiên, chúng ta chọn giá trị k phù hợp khi thực hiện.

Hàm Counting(D_F, I_F, T_F, *minsupp*): Tạo ra F_1 là tất cả các tập phổ biến có độ dài bằng 1 và có độ hỗ trợ lớn hơn hoặc bằng *minsupp*.

Hàm Join(F_{m-1}): Thực hiện kết nối các cặp các thuộc tính mờ từ tập các thuộc tính mờ phổ biến F_{m-1} phần tử.

Hàm Prune(C_m): Sử dụng tính chất "*mọi tập con khác rỗng của tập phổ biến cũng là tập phổ biến" và "mọi tập chứa tập không phổ biến đều là tập không phổ biến*", để cắt tỉa những thuộc tính mờ nào trong C_m có tập con lực lượng $m - 1$ không thuộc tập các tập thuộc tính mờ phổ biến F_{m-1}.

Hàm Checking(C_m, D_F, *minsupp*): Duyệt qua CSDL D_F để cập nhật độ hỗ trợ cho các tập thuộc tính mờ trong C_m. Checking(C_m, D_F, *minsupp*) sẽ chọn những tập mục phổ biến thỏa mãn *minsupp* để đưa vào trong F_m.

Hàm Fuzzy_Generate_Rules(F, *minconf*): Sinh luật kết hợp mờ từ tập các tập phổ biến F có độ tin cậy thỏa mãn *minconf*.

4 Kết quả thực nghiệm

Dữ liệu thực nghiệm được khảo sát từ 200 sinh viên năm thứ nhất. Bảng khảo sát được thiết kế gồm 8 nhóm: **Yếu tố thông điệp truyền thông marketing (TT), Yếu tố kênh truyền thông marketing (KT), thái độ (TĐ), chuẩn chủ quan (CQ), nhận thức kiểm soát hành vi (NT), yếu tố đặc điểm trường học (ĐĐ), ý định lựa chọn trường (YĐ), quyết định lựa chọn trường (QĐ**) với 40 nhận định về "*Tác động của truyền thông marketing đến quyết định lựa chọn cơ sở giáo dục đại học của sinh viên*". Mỗi nhận định được chọn 1 đến 5 thể hiện mức độ đồng ý của sinh viên về một nhận định (1. Hoàn toàn không đồng ý, 2. không đồng ý, 3. Không có ý kiến, 4. đồng ý, 5. Hoàn toàn đồng ý). Mã hóa các nhóm yếu tố và nhận định tác động truyền thông marketing được thể hiện trong Bảng 1.

Trên cơ sở số liệu khảo sát, không mất tính tổng quát, chúng tôi xem trọng số của mỗi nhận định trong cùng một nhóm là như nhau, sử dụng phương pháp tích hợp ĐSGT để xây dựng cơ sở dữ liệu cần khai phá luật kết hợp mờ.

Bảng 1. Mã hóa các nhóm yếu tố và nhận định tác động truyền thông marketing

STT	Nhóm	Nhận định						
1	**TT**	TT1	TT2	TT3	TT4	TT5	TT6	TT7
2	**KT**	KT1	KT2	KT3	KT4	KT5	KT6	
3	**TĐ**	TĐ1	TĐ2	TĐ3	TĐ4	TĐ5		
4	**CQ**	CQ1	CQ2	CQ3	CQ4	CQ5		
5	**NT**	NT1	NT2	NT3				
6	**ĐĐ**	ĐĐ1	ĐĐ2	ĐĐ3	ĐĐ4	ĐĐ5	ĐĐ6	ĐĐ7
7	**YĐ**	YĐ1	YĐ2	YĐ3				
8	**QĐ**	QĐ1	QĐ2	QĐ3	QĐ4			

Xem miền trị của 8 thuộc tính nhóm là ĐSGT, khi đó ta chọn cấu trúc ĐSGT như sau:

$\underline{X} = (\boldsymbol{X}, G, H, \preceq)$, với $\boldsymbol{X}$ là tên của thuộc tính nhóm, G = {0, "*thấp*", W, "*cao*",1}, H^- = {*ít*, *khả năng*}, H^+ = {*hơn*, *rất*}, Cho $W = 0.6$, $\alpha = 0.6$ và $\beta = 0.4$. Ta có fm(*thấp*) = 0.6, fm(*cao*) = 0.4, μ(*khả năng*) = 0.4, μ(*ít*) = 0.2, μ(*hơn*) = 0.25 và μ(*rất*) = 0.15

Phân hoạch đoạn [0, 200] thành 5 khoảng tương tự mức 1 là: S(0), S(*thấp*), S(*W*), S(*cao*) và S(1). Ta có: fm(*rất cao*) × 200 = 0.15 × 0.4 × 200 = 12. Vậy S(1) × 200 = (188, 200]; (fm(*khả năng cao*) + fm(*hơn cao*)) × 200 = (0.4 × 0.4 + 0.25 × 0.4) × 200 = 52 và S(*cao*) × 200 = (136, 188];

(fm(*ít thấp*) + fm(*ít cao*)) × 200 = (0.2 × 0.6 + 0.2 × 0.4) × 200 = 40 và S(*W*) × 200 = (96, 136]; (fm(*khả năng thấp*) + fm(*hơn thấp*)) × 200 = (0.4 × 0.6 + 0.25 × 0.6) × 200 = 78 và S(*thấp*) × 200 = (18, 96], S(0) × 200 = [0, 18].

Tương tự, phân hoạch đoạn [0, 200] ta có 8 khoảng tương tự mức 2 là: S(*rất thấp*) = [0, 18], S(*hơn thấp*) = (18, 48], S(*khả năng thấp*) = (48, 96], S(*ít thấp*) = (96, 120], S(*ít cao*) = (120, 136], S(*khả năng cao*) = (136, 168], S(*hơn cao*) = (168, 188] và S(*rất cao*) = (188, 200].

i) Kết quả thu được một số luật kết hợp mờ với độ hỗ trợ 25% độ tin cậy 60% như sau: Không có luật kết hợp mờ sinh ra ở mức độ 1 và 2. Mức độ 3 và mức độ 4 có số luật kết hợp mờ sinh ra là tương tự nhau. Do đó, trong trường hợp này ý nghĩa cho việc định hướng trong tuyển sinh không cao. Đối với mức độ 5, Nếu TT là *"hơn cao"* và TD là *"hơn cao"* và CQ là *"rất cao"* thì QĐ "trường đại học ưu tiên chọn lựa". Điều này có nghĩa sinh viên rất quan tâm đến các yếu tố nhận định liên quan đến TT, TD và CQ.

ii) Kết quả thu được một số luật kết hợp mờ với độ hỗ trợ 29% độ tin cậy 76% như sau:

Đối với mức độ 3, Nếu KT là *"khả năng cao"* và CQ là *"rất cao"* và NT *là "rất cao"* và YD là *"hơn cao"* thì QĐ "trường đại học ưu tiên chọn lựa". Điều này có nghĩa sinh viên chỉ quan tâm đến các yếu tố nhận định liên quan đến KT, CQ, NT và YD.

Đối với mức độ 4, số luật kết hợp mờ sinh ra từ tất cả các nhóm nhận định. Do đó, trong trường hợp này các nhóm yếu tố nhận định ảnh hưởng đến việc lựa chọn trường đại học của sinh viên là như nhau.

Đối với mức độ 5, Nếu TD là *"hơn cao"* và CQ là *"rất cao"* thì QĐ "trường đại học ưu tiên chọn lựa". Điều này có nghĩa sinh viên rất quan tâm đến các yếu tố nhận định liên quan đến TD và CQ.

iii) Kết quả thu được một số luật kết hợp mờ với độ hỗ trợ 36% độ tin cậy 92% như sau:

Đối với mức độ 3, Nếu NT là *"rất cao"* và YD là *"hơn cao"* thì QĐ "trường đại học X ưu tiên chọn lựa". Điều này có nghĩa sinh viên rất quan quan tâm đến các yếu tố nhận định liên quan đến NT và YD.

Đối với mức độ 4, Nếu KT là *"khả năng cao"* và TĐ là *"hơn cao"* và ĐĐ là *"hơn cao"* thì QĐ "giới thiệu cho bạn bè lựa chọn trường đại học X". Điều này có nghĩa sinh viên quan tâm đến các yếu tố nhận định liên quan đến KT, TĐ và ĐĐ.

iv) Kết quả thu được một số luật kết hợp mờ với độ hỗ trợ 37% độ tin cậy 94% như sau:

Đối với mức độ 3, Nếu NT là *"rất cao"* và YD là *"hơn cao"* thì QĐ "trường đại học X ưu tiên chọn lựa". Điều này có nghĩa sinh viên rất quan tâm đến các yếu tố nhận định liên quan đến NT và YD.

Đối với mức độ 4, Nếu TĐ là *"hơn cao"* và ĐĐ là *"hơn cao"* thì QĐ "giới thiệu cho bạn bè lựa chọn trường đại học X". Điều này có nghĩa sinh viên quan tâm đến các yếu tố nhận định liên quan đến TĐ và ĐĐ.

5 Kết luận

Bài báo trình bày một cách tiếp cận khai phá luật kết hợp mờ sử dụng ĐSGT và ứng dụng để khai phá dữ liệu khảo sát của sinh viên năm thứ nhất trong việc chọn trường đại học. Chúng tôi xem miền trị của mỗi nhóm nhận định là một cấu trúc ĐSGT, mỗi giá trị ngôn ngữ được xây dựng bằng một khoảng lân cận. Bước đầu thử nghiệm trên bộ dữ liệu khảo sát từ 200 sinh viên với 8 nhóm và 39 nhận định. Kết quả này là một trong những phương pháp giúp cho lãnh đạo các trường đại học cũng như các bộ phận làm công tác tuyển sinh trong trường đại học trong việc xây dựng chiến lược truyền thông tuyển sinh một cách hiệu quả. Việc thu thập mẫu dữ liệu lớn

và tối ưu các tham số trong ĐGST nhằm khai phá luật kết hợp mờ tốt hơn sẽ được chúng nghiên cứu và trình bày trong những công trình tiếp theo.

Tài liệu tham khảo

1. Le kha, A., C.V. Srikrishna and Viji Vinod, *"Fuzzy Assciation Rule Mining"*, Journal of Computer Science, 11(1) (2015), 71-74.
2. Onur Dogan, Furkan Can Kem, Basar Oztaysi, *"Fuzzy assciation rule mining approach to Indentify e-comerce product association considering sales amount"*, Complex and Intelligent Systems, 8 (2022), 1551-1560.
3. Ashish Mangalampalli, Vikram Pudi, "*Fuzzy Association Rule Mining Algorithm for Fast and Efficient Performance on Very Large Datasets*", IEEE International Conference on Fuzzy Systems (2009)
4. Anna L. Buczak, Christopher M. Gifford, "*Fuzzy Association Rule Mining for Community Crime Pattern Discovery*", ISI-KDD (2010)
5. Gamal F. Elhady, Haitham Elwahsh, Maazen Alsabaan, Mohamed I. Ibrahem, Ebtesam Shemis, "*A Formal Fuzzy Concept-Based Approach for Association Rule Discovery with Optimized Time and Storage*", *Mathematics* (2024), *12*(22), 2-17.
6. Régis Pierrard, Jean-Philippe Poli, Céline Hudelot, "*A Fuzzy Close Algorithm for Mining Fuzzy Association Rules*", 17th International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems (IPMU 2018), 1-11.
7. Tran, T. T., Nguyen, T. N., Nguyen, T. T., Nguyen, G. L., & Truong, C. N., "*A Fuzzy Association Rules Mining Algorithm with Fuzzy Partitioning Optimization for Intelligent Decision Systems*". International Journal of Fuzzy Systems, 2022, 1-14
8. Tran, T. T., Nguyen, T. T., Nguyen, G. L., & Truong, C. N. "*Parallel Fuzzy Frequen Itemset Mining Using Cellular Automata*". Journal of Computer Science and Cybernetics, 38(4), 2022, 293-310.
9. Nguyen Công Hào, Nguyễn Công Đoàn, "Luật kết hợp mờ dựa trên ngữ nghĩa đại số gia tử", Tạp chí Khoa học Đại học Huế, Vol 75A (2012), 39-52.
10. Nguyễn Cát Hồ, Lê Xuân Vinh, Nguyễn Công Hào, *"Thống nhất dữ liệu và xây dựng quan hệ tương tự trong cơ sở dữ liệu ngôn ngữ bằng đại số gia tử"*, Tạp chí Tin học và Điều khiển học, T.25, S.4 (2009), 314-332.
11. Nguyễn Tuấn Anh, Trịnh Thúy Hà, *"Ứng dụng khai phá luật kết hợp mờ hỗ trợ sinh viên lập kế hoạch học tập"*, Tạp chí Khoa học và Công nghệ Đại học Thái Nguyên, T.226, S.02 (2021), 35-41.