Một phương pháp xử lý giá trị ngoại lai trong tập mẫu huấn luyện cây quyết định sử dụng đại số gia tử

A method for handling outliers in training data set to build a decision tree based on hedge algebra

Lê Văn Tường Lân, Nguyễn Mậu Hân, Nguyễn Công Hào

***Abstract**: In this paper we propose a method to handle the outliers of the fuzzy fileds in the sample training dataset that based on hedge algebra. Due to the value of the attribute domain may be value or linguistic so we need a method of approximate data in a simple way and effective to handle outliers of it.*

***Keyword**: Hedge algebra, decision tree, fuzzy decision tree, training data set.*

I. ĐẶT VẤN ĐỀ

Trong bài toán khai phá dữ liệu thì việc chọn đúng tập mẫu huấn luyện là một trong những giai đoạn rất quan trọng, nó quyết định kết quả của công việc khai phá. Bài toán xây dựng cây quyết định cũng là một bài toán của khai phá dữ liệu nên vấn đề chọn tập mẫu huấn luyện cây là vấn đề cần phải giải quyết.

Như chúng ta đã biết, cây quyết định được xây dựng dựa trên một tập dữ liệu huấn luyện bao gồm các đối tượng mẫu, tất cả các mẫu của tập đều có chung một cấu trúc, gồm những cặp <*Thuộc tính, Giá trị*>. Mỗi đối tượng được mô tả bởi một tập giá trị các thuộc tính và nhãn lớp. Để xây dựng cây quyết định, tại mỗi nút trong cần xác định một thuộc tính thích hợp để kiểm tra, phân chia dữ liệu thành các tập con. Trên tập mẫu huấn luyện *M*, về cơ bản, các thuật toán phân lớp phải thực hiện 2 bước sau:

Bước 1: Chọn thuộc tính A_i có các giá trị $a_{i1}, a_{i2}, \ldots, a_{in}$.

Bước 2: Với thuộc tính A_i được chọn, ta tạo một nút của cây và sau đó chia các mẫu ứng với nút này thành các tập mẫu tương ứng $M_1, M_2, \ldots, M_k$ và sau đó lại tiếp tục.

Đây là bước phân chia với kết quả nhận được từ *Bước 1*, điều này có nghĩa là chất lượng của cây kết quả phụ thuộc phần lớn vào cách chọn thuộc tính và cách phân chia các mẫu tại mỗi nút. Chính vì điều này, các thuật toán đều phải tính lượng thông tin nhận được trên các thuộc tính và chọn thuộc tính tương ứng có lượng thông tin tốt nhất để làm nút phân tách trên cây, nhằm để đạt được cây có ít nút nhưng có khả năng dự đoán cao [2,17].

Trong thế giới thực, dữ liệu nghiệp vụ rất đa dạng vì chúng được lưu trữ để phục vụ nhiều công việc khác nhau, nhiều thuộc tính đã được thuần nhất miền giá trị trước khi lưu trữ nhưng cũng tồn tại nhiều thuộc tính có miền trị chưa thuần nhất [5,7,20]. Khi các thuộc tính chưa thuần nhất này xuất hiện trong tập mẫu huấn luyện, các thuật toán học để xây dựng cây chưa thể tiến hành. Do đó, cần phải tiền xử lý dữ liệu để có được tập mẫu huấn luyện thuần nhất. Vấn đề đặt ra là ta phải xử lý như thế nào để có được kết quả là khả quan.

Ví dụ: Cho bảng dữ liệu DIEUTRA lưu trữ về tình hình mua máy tính xách tay của khách hàng tại một công ty như Bảng 1, ta cần xây dựng một cây quyết định cho việc dự đoán khách hàng mua hàng.

Lúc này, trong tập mẫu huấn luyện có thuộc tính *LươngTháng* chứa dữ liệu không nhất quán. Để có thể huấn luyện trên các tập mẫu này, chúng ta phải làm thuần nhất dữ liệu cho các thuộc tính này [8].

***Bảng 1**: Bảng dữ liệu điều tra*

Phiếu ĐT	Họ Tên	Số CMND	Chuyên Ngành	Lương Tháng	Máy Tính
M01045	Nguyễn An	193567450	Luật	48	Không
M01087	Lê Văn Bình	191568422	Luật	Thấp	Không
M02043	Hoàng Hà	196986568	CNTT	53	Có
M02081	Võ Văn Bình	191003117	LịchSử	20	Có
M02046	Trần Hương	196001278	LịchSử	Cao	Có
M03087	Nguyễn Lài	198235457	LịchSử	Cao	Không
M03025	Vũ Văn Hoa	198875584	CNTT	Rất cao	Có
M03017	Lê Bá Linh	191098234	Luật	35	Không
M04036	Võ Bạch Ân	196224003	Luật	100	Có
M04037	Lê Văn Bách	196678578	LịchSử	50	Có
M04042	Lê Vũ Bình	197543457	Luật	Quá cao	Có
M04083	Nguyễn Hoá	192267457	CNTT	Ít thấp	Có
M05041	Lê Thị Hoa	198234309	CNTT	55	Có
M05080	Trần Chung	196679345	LịchSử	50	Không
M05086	Trần Xuân Bi	196679995	Luật	QuáThấp	Không

Trong quá trình thuần nhất dữ liệu, việc định lượng cho các giá trị ngôn ngữ thông thường được dựa vào miền giá trị rõ của chính thuộc tính đó tức là ta có thể tìm thấy miền trị [ψ_{min}, ψ_{max}] từ miền giá trị rõ và sau đó sẽ định lượng cho các giá trị ngôn ngữ từ miền trị này [6,8]. Tuy vậy, việc tìm miền trị [ψ_{min}, ψ_{max}] không phải lúc nào cũng thuận lợi vì có thể xuất hiện các giá trị ngôn ngữ mà giá trị thật sự của nó nằm ngoài miền dữ liệu rõ đang có trong thuộc tính đang xét. Các giá trị này ta gọi là các giá trị ngoại lai.

Ví dụ như ở Bảng 1, với thuộc tính *LươngTháng*, ta có miền trị rõ của nó là [ψ_{min}, ψ_{max}] = [20, 100] chỉ đủ để định lượng cho các giá trị ngôn ngữ thông thường như: {*Thấp, Ít thấp, Cao, Rất cao*}. Các giá trị ngôn ngữ "*Quá cao*" hay "*Quá thấp*" trong thuộc tính *LươngTháng* không nằm trong miền trị này nên nó là các giá trị ngoại lai của thuộc tính đang xét. Việc sử dụng miền trị [ψ_{min}, ψ_{max}] = [20..100] để định lượng cho các giá trị ngôn ngữ này ta sẽ thu được các giá trị rõ, nhưng nó không phản ánh đúng giá trị ngôn ngữ này cần thể hiện.

Các phương pháp tiền xử lý dữ liệu truyền thống như sử dụng giá trị hằng toàn cục hay sử dụng giá trị trung bình của thuộc tính, phương pháp Binning, hồi quy,... [9,15,24] không thể sử dụng để xác định các giá trị ngoại lai này. Ta có thể bỏ qua các trường hợp ngoại lai này hoặc xem chúng cùng lớp tương đương với các giá trị ngôn ngữ khác, chẳng hạn ở trong mẫu trên ta sẽ đồng nhất ngữ nghĩa "*Quá cao*" với "*Rất cao*", nhưng việc làm này sẽ làm mất thông tin. Sử dụng ý kiến chuyên gia cho việc xác định giá trị rõ cho các giá trị ngoại lai này không phải lúc nào cũng cho kết quả như mong muốn vì còn phụ thuộc vào trình độ của chuyên gia.

Trong thời gian qua, đại số gia tử được nhiều tác giả nghiên cứu trong và ngoài nước nghiên cứu và đã có những kết quả đáng kể, đặc biệt trong lập luận xấp xỉ và trong một số bài toán điều khiển [1,6,12-19, 23]. Việc sử dụng đại số gia tử để xử lý các giá trị ngôn ngữ

trên miền dữ liệu chưa thuần nhất đã cho kết quả rất tích cực [6, 8]. Trong bài báo này, chúng tôi sẽ trình bày một cách để có thể giải quyết các giá trị ngoại lai của các thuộc tính có dữ liệu chưa thuần nhất theo tiếp cận đại số gia tử, nhằm giúp ta có thể xác định giá trị rõ cho các giá trị ngôn ngữ khi mà giá trị rõ thật sự của nó nằm ngoài các khoảng rõ đang có trong miền dữ liệu của tập mẫu huấn luyện

II. THUẦN NHẤT GIÁ TRỊ CHO CÁC THUỘC TÍNH DỰA TRÊN ĐẠI SỐ GIA TỬ

Với tập mẫu huấn luyện M có thuộc tính chưa thuần nhất A_i, ta gọi $Dom(A_i)$ là miền trị tương ứng với thuộc tính mờ A_i và xem như một đại số gia tử thì khi đó $Dom(A_i) = D_{Ai} \cup LD_{Ai}$, với D_{Ai} là tập các giá trị kinh điển của A_i và LD_{Ai} là tập các giá trị ngôn ngữ của A_i. Trước tiên, chúng ta phải xử lý để chuyển đổi thuộc tính chưa thuần nhất A_i nhằm đạt được thuần nhất trên tập giá trị kinh điển D_{Ai} hoặc trên tập giá trị ngôn ngữ LD_{Ai} [8].

II.1. Chuyển giá trị ngôn ngữ về giá trị số

Để chuyển đổi một giá trị trong ĐSGT (giá trị ngôn ngữ) thành một số trong [0,1] ta sử dụng hàm định lượng ngữ nghĩa v của X được xây dựng như sau với $x = h_{im} \ldots h_{i2}h_{i1}c$:

(1) x là phần tử sinh tức là $x = c^+$ hoặc $x = c^-$, lúc này:

$$v(c^-) = W - \alpha.fm(c^-) \text{ và } v(c^+) = W + \alpha.fm(c^+)$$

(2) x không phải là phần tử sinh:

- Với $1 \le j \le p$: $v(h_jx) = v(x) + Sign(h_jx) \times \left[\sum_{i=j}^{p} fm(h_ix) - \frac{1}{2}(1 - Sign(h_jx)Sign(h_1h_jx)(\beta - \alpha))\, fm(h_jx)\right]$
- Với $j > p$: $v(h_jx) = v(x) + Sign(h_jx) \times \left[\sum_{i=p+1}^{j} fm(h_ix) - \frac{1}{2}(1 - Sign(h_jx)Sign(h_1h_jx)(\beta - \alpha))\, fm(h_jx)\right]$

II.2. Chuyển giá trị số về giá trị ngôn ngữ

Trước tiên, chúng ta sử dụng hàm IC để chuyển một giá trị số về một giá trị thuộc [0,1], hàm IC: $Dom(A_i) \to [0,1]$ được xác định như sau:

Nếu $LD_{Ai} = \varnothing$ và $D_{Ai} \neq \varnothing$ thì $\forall \omega \in Dom(A_i)$ ta có:

$IC(\omega) = 1 - \dfrac{\psi_{\max} - \omega}{\psi_{\max} - \psi_{\min}}$, với $Dom(A_i) = [\psi_{min}, \psi_{max}]$ là miền trị kinh điển của A_i.

Nếu $D_{Ai} \neq \varnothing$, $LD_{Ai} \neq \varnothing$ thì $\forall \omega \in Dom(A_i)$ ta có $IC(\omega) = \{\omega \times v(\psi_{maxLV})\}/\psi_{max}$, với $LD_{Ai} = [\psi_{minLV}, \psi_{maxLV}]$ là miền trị ngôn ngữ của A_i.

Nếu chúng ta chọn các tham số W và độ đo tính mờ cho các gia tử sao cho $v(\psi_{maxLV}) \approx 1.0$ thì $\{\omega \times v(\psi_{maxLV})\}/\psi_{max} \approx 1 - \dfrac{\psi_{\max} - \omega}{\psi_{\max} - \psi_{\min}}$

Ví dụ: Cho miền trị cơ sở $U(T)$ = {0…100, … rất rất trẻ, … , rất rất già}.

D_T = {20, 25, 27, 30, 45, 60, 75, 66, 80}.

LD_T = {trẻ, rất trẻ, già, khá trẻ, khá già, ít già, rất già, rất rất trẻ}.

$Dom(T) = D_T \cup LD_T$.

Nếu $LD_T = \varnothing$ khi đó $Dom(T) = D_T$ = {20, 25, 27, 30, 45, 60, 75, 66, 80}. Do đó $\forall \omega \in D_T$, chuyển đổi giá trị về một số trong [0, 1] nhờ hàm $IC(\omega)$. Ta có D_T = {0.2, 0.25, 0.27, 0.3, 0.45, 0.6, 0.75, 0.66, 0.8}.

Nếu $D_T \neq \varnothing$ và $LD_T \neq \varnothing$ ta có: $Dom(T) = D_T \cup LD_T$ = {trẻ, rất trẻ, già, khá trẻ, khá già, ít giá, rất già, rất rất trẻ, 20, 25, 27, 30, 45, 60, 75, 66, 80}. Giả sử tính được $v(\psi_{maxLV})$ = v(rất rất già) = 0.98. Khi đó $\forall \omega \in D_T$. $IC(\omega) = \{\omega \times v(\psi_{maxLV})\}/\psi_{max} = (\omega \times 0.98)/100$, hay $\forall \omega \in D_T$ chuyển đổi giá trị về một số trong [0, 1] nhờ hàm $IC(\omega)$. Nên ta có: D_T = {0.196, 0.245, 0.264, 0.294, 0.441, 0.588, 0.735, 0.646, 0.784}.

Tiếp đến, hàm Φ_k được sử dụng để chuyển một giá trị trong [0, 1] thành một giá trị ngôn ngữ x tương ứng trong đại số gia tử X.

Φ_k: [0, 1] $\to X$ gọi là hàm ngược của hàm v theo mức k được xác định: $\forall a \in [0, 1]$, $\Phi_k(a) = x^k$ khi và chỉ khi $a \in I(x^k)$, với $x^k \in X^k$.

II.3. Định lý

Cho đại số gia tử $\underline{X} = (X, G, H, \leq)$, ν là hàm định lượng ngữ nghĩa của X, Φ_k là hàm ngược của ν, ta có:

(1) $\forall x^k \in X^k$, $\Phi_k(\nu(x^k)) = x^k$

(2) $\forall a \in I(x^k)$, $\forall b \in I(y^k)$, $x^k \neq_k y^k$, nếu $a < b$ thì $\Phi_k(a) <_k \Phi_k(b)$

Thật vậy:

(1). Đặt $a = \nu(x^k) \in [0, 1]$. Vì $\nu(x^k) \in I(x^k)$ nên $a \in I(x^k)$. Theo định nghĩa ta có $\Phi_k(\nu(x^k)) = x^k$.

(2) Vì $x^k \neq_k y^k$ nên theo định nghĩa ta có $x^k <_k y^k$ hoặc $y^k <_k x^k$, suy ra $\nu(x^k) < \nu(y^k)$ hoặc $\nu(y^k) < \nu(x^k)$. Mặt khác ta có $\nu(x^k) \in I(x^k)$ và $\nu(y^k) \in I(y^k)$, theo giả thiết $a < b$ do đó $x^k <_k y^k$. Hay $\Phi_k(a) <_k \Phi_k(b)$.

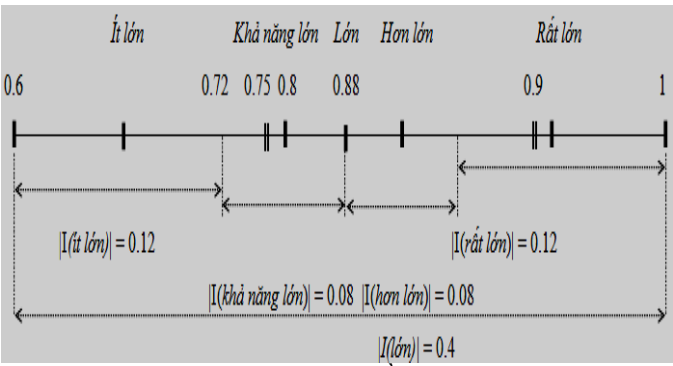
Ví dụ: Cho đại số gia tử $\underline{X} = (X, G, H, \leq)$, trong đó $H^+ = \{$*hơn, rất*$\}$ với *hơn* < *rất* và $H^- = \{$*ít, khả năng*$\}$ với *ít* > *khả năng*, $G = \{$*nhỏ, lớn*$\}$. Cho $W = 0.6$, μ(*hơn*) = 0.2, μ(*rất*) = 0.3, μ(*ít*) = 0.3, μ(*khả năng*) = 0.2.

Ta có $P^2 = \{I($*hơn lớn*$), I($*rất lớn*$), I($*ít lớn*$), I($*khả năng lớn*$), I($*hơn nhỏ*$), I($*rất nhỏ*$), I($*ít nhỏ*$), I($*khả năng nhỏ*$)\}$ là phân hoạch của [0, 1]. *fm*(*nhỏ*) = 0.6, *fm*(*lớn*) = 0.4, *fm*(*rất lớn*) = 0.12, *fm*(*khả năng lớn*) = 0.08. Ta có |*I*(*rất lớn*)| = *fm*(*rất lớn*) = 0.12, hay *I*(*rất lớn*) = [0.88, 1].

Do đó theo định nghĩa $\Phi_2(0.9)$ = *rất lớn* vì 0.9 $\in$ *I*(*rất lớn*), như Hình 1.

Tương tự ta có |*I*(*khả năng lớn*)| = *fm*(*khả năng lớn*) = 0.08, hay *I*(*khả năng lớn*) = [0.72, 0.8].

Do đó theo định nghĩa $\Phi_2(0.75)$ = *khả năng lớn* vì 0.75 $\in$ *I*(*khả năng lớn*).



***Hình 1**. Tính mờ của phần tử sinh lớn*

III. XỬ LÝ GIÁ TRỊ NGOẠI LAI THEO CÁCH TIẾP CẬN CỦA ĐẠI SỐ GIA TỬ

Như thế, với bất kỳ một thuộc tính không thuần nhất A_i, có $Dom(A_i) = D_{Ai} \cup LD_{Ai}$ ta sẽ chuyển thuần nhất về giá trị số thuộc đoạn [0,1] để từ đó chuyển về giá trị ngôn ngữ hay giá trị kinh điển tùy thuộc yêu cầu của việc sử dụng mẫu. Tuy vậy, trong quá trình xây dựng các ánh xạ chuyển, việc nhận ra các giá trị biên $[\psi_{min}, \psi_{max}]$ đối với miền trị kinh điển D_{Ai} của A_i hay $[\psi_{minLV}, \psi_{maxLV}]$ đối với miền trị ngôn ngữ của LD_{Ai} là thực sự cần thiết.

Trong hầu hết các trường hợp xảy ra, các giá trị biên này đã được lưu trữ sẵn trong tập mẫu và chúng được xác định dễ dàng thông qua việc duyệt mẫu và chọn. Lúc này, các giá trị ngôn ngữ trong LD_{Ai} của A_i khi được làm rõ sẽ có giá trị nằm trong $[\psi_{min}, \psi_{max}]$ của D_{Ai} và ngược lại.

Tuy vậy, đôi khi cũng xuất hiện các giá trị ngoại lai, các giá trị này có thể là các giá trị ngôn ngữ nhưng miền giá trị $[\psi_{min}, \psi_{max}]$ của D_{Ai} không thể hiện đúng và ngược lại. Cụ thể ta có thể gặp trường hợp $\nu(\psi_{minLV}) < IC(\psi_{min})$ hoặc $\nu(\psi_{maxLV}) > IC(\psi_{max})$. Khi gặp các trường hợp này, ta có thể chọn cách bỏ qua hoặc xem chúng cùng nằm trong cùng một lớp với các giá trị ψ_{min}, ψ_{max}. Tuy vậy, cách làm này sẽ làm mất thông tin và không thể hiện đúng bản chất sự việc. Ở đây, ta sẽ vận dụng phương pháp thuần nhất giá trị dựa đại số gia tử ở trên để đưa ra phương pháp xấp xỉ nhằm xác định giá trị rõ cho các giá trị ngôn ngữ ngoại lai này.

Theo phương pháp đã đề xuất ở mục II, ta thấy tính mờ của các giá trị trong đại số gia tử là một đoạn con của [0, 1] cho nên họ các đoạn con như vậy của các giá trị có cùng độ dài sẽ tạo thành phân hoạch của [0, 1]. Phân hoạch ứng với các giá trị có độ dài từ lớn hơn sẽ mịn hơn và khi độ dài lớn vô hạn thì độ dài của các đoạn trong phân hoạch giảm dần về 0. Hơn nữa, các giá trị ngôn ngữ là một tập sắp thứ tự tuyến tính nên ta sẽ chia các đoạn con tương ứng thành các phân hoạch nhỏ hơn nhằm xác định lại độ dài của các đoạn $[0, \nu(\psi_{GiáTrịNgoạiLai})]$ hay $[\nu(\psi_{GiáTrịNgoạiLai}), 1]$ để từ đó có xác định giá trị rõ cho các giá trị ngôn ngữ này.

Cách thức phân hoạch như sau :

1. Nếu $\psi_{GiáTrịNgoạiLai} < \psi_{MinLV}$ thì :

- Phân hoạch đoạn [0, $\nu(\psi_{MinLV})$] thành [0, $\nu(\psi_{GiáTrịNgoạiLai})$] và [$\nu(\psi_{GiáTrịNgoạiLai})$, $\nu(\psi_{MinLV})$]

- $fm(h_{GiáTrịNgoạiLai}) \sim fm(h_{MinLV}) \times I(\psi_{MinLV})$

- $fm(h_{MinLV}) = fm(h_{MinLV}) - fm(h_{GiáTrịNgoạiLai})$

2. Nếu $\psi_{GiáTrịNgoạiLai} > \psi_{MaxLV}$ thì:

- Phân hoạch đoạn [$\nu(\psi_{MaxLV})$, 1] thành [$\nu(\psi_{MaxLV})$, $\nu(\psi_{GiáTrịNgoạiLai})$] và [$\nu(\psi_{GiáTrịNgoạiLai})$, 1]

- $fm(h_{GiáTrịNgoạiLai}) \sim fm(h_{MaxLV}) \times I(\psi_{MaxLV})$

- $fm(h_{MaxLV}) = fm(h_{MaxLV}) - fm(h_{GiáTrịNgoạiLai})$

Như vậy, ta có thuật toán tổng quát như sau:

Bước 1: Tách riêng các giá trị ngoại lai này ra khỏi A_i, được A'_i ;

Bước 2: Thực hiện việc thuần nhất các giá trị cho A'_i theo cách đã đề cập ở mục 3;

Bước 3: So sánh các $\psi_{GiáTrịNgoạiLai}$ với ψ_{Max} và ψ_{Min} của A'_i. Thực hiện lại các phân hoạch trên đoạn [0, 1] ;

Bước 4: Dựa vào $IC(\omega)$ của A'_i, tính lại $IC(\omega)$ cho A_i ;

Bước 5: Thuần nhất giá trị cho A_i .

Tính đúng của giải thuật: Do tất cả các phân hoạch trên không vượt ra khỏi đoạn đang xét là $|fm(h_{minLV})|$ hay $|fm(h_{maxLV}|$ nên không làm phá vỡ các phân hoạch đang có của đoạn [0, 1]. Do $I(\psi_{MinLV}) > 0$ và $I(\psi_{MaxLV}) < 1$, nên cách phân hoạch trên là phù hợp với phương pháp thuần nhất đã nêu ở Mục III.

Vì vậy, thuật toán này xác định được giá trị rõ nằm ngoài miền giá trị đang có cho giá trị ngôn ngữ ngoại lai trong tập mẫu.

IV. ỨNG DỤNG XỬ LÝ GIÁ TRỊ NGOẠI LAI CHO DỮ LIỆU BÁN HÀNG

Cho tập mẫu huấn luyện như Bảng 1, hãy xây dựng cây quyết định hỗ trợ dự đoán việc mua máy. Do tập mẫu huấn luyện có thuộc tính *LươngTháng* là chưa thuần nhất nên ta phải thuần nhất các giá trị cho *LươngTháng.* Ta có:

$Dom(LươngTháng) = D_{LươngTháng} \cup LD_{LươngTháng}$

$D_{LươngTháng} = \{20,35, 48, 50, 53, 55, 100\}$;

$\psi_{min} = 20$; $\psi_{max} = 100$

$LD_{LươngTháng}$ = *{Quá thấp, Rất thấp, Thấp, Ít thấp, Cao, Rất cao, Quá cao}*.

Trong $LD_{LươngTháng}$, các giá trị ngôn ngữ: *Rất thấp, Thấp, Ít thấp, Cao, Rất cao* là các giá trị thông thường có miền trị nằm trong miền [ψ_{min}, ψ_{max}] đã biết còn các giá trị ngôn ngữ *Quá thấp, Quá cao* không nằm trong miền trị này, chúng là các giá trị ngoại lai của *LươngTháng*. Lúc này, chúng ta phải định lượng cho toàn bộ các giá trị ngôn ngữ của $LD_{LươngTháng}$ nhằm đạt tính thuần nhất của tập mẫu huấn luyện nhưng không có ý kiến của chuyên gia để xác định lại miền trị [ψ_{min}, ψ_{max}].

Đầu tiên, ta thuần nhất các giá trị cho *LươngTháng* nhưng không xét các giá trị ngoại lai trong mô hình. Lúc này:

$D_{LươngTháng} = \{20,35, 48, 50, 53, 55, 100\}$;

$LD_{LươngTháng}$ = *{Thấp, Ít thấp, Cao, Rất cao}*

Xây dựng 1 ĐSGT để tính cho thuộc tính không thuần nhất *LươngTháng* như sau:

$\underline{X}_{LươngTháng} = (X_{LươngTháng}, G_{LươngTháng}, H_{LươngTháng}, \leq)$, với $G_{LươngTháng}$ = {*cao, thấp*},

$H^{+}_{LươngTháng}$ = {hơn, *rất*},

$H^{-}_{LươngTháng}$ = {*khả năng, ít*}.

Quan hệ ngữ nghĩa: *rất* > *hơn* và *ít* >*khả năng*.

$W_{LươngTháng} = 0.4$, $fm(thấp) = 0.4$, $fm(cao) = 0.6$,

$\mu(rất) = 0.35$, $\mu(hơn) = 0.25$,

$\mu(khả\ năng) = 0.20$, $\mu(ít) = 0.20$.

Lúc này ta có :

$fm(rất\ thấp) = 0.35 \times 0.4 = 0.14$,

$fm(hơn\ thấp) = 0.25 \times 0.4 = 0.10$,

$fm(ít\ thấp) = 0.2 \times 0.4 = 0.08$,

$fm(khả\ năng\ thấp) = 0.2 \times 0.4 = 0.08$.

Vì *rất thấp* < *hơn thấp* < *thấp* < *khả năng thấp* < *ít thấp* nên:

I(*rất thấp*) = [0, 0.14], *I*(*hơn thấp*) = [0.14, 0.24],

I(*khả năng thấp*) = [0.24, 0.32],

I(*ít thấp*) = [0.32, 0.4].

Ta lại có: *fm*(*rất cao*) = 0.35 × 0.6 = 0.21,

fm(*hơn cao*) = 0.25 × 0.6 = 0.15,

fm(*ít cao*) = 0.2 × 0.6 = 0.12,

fm(*khả năng cao*) = 0.2 × 0.6 = 0.12.

Vì *ít cao* < *khả năng cao* < *cao* < *hơn cao* < *rất cao* nên :

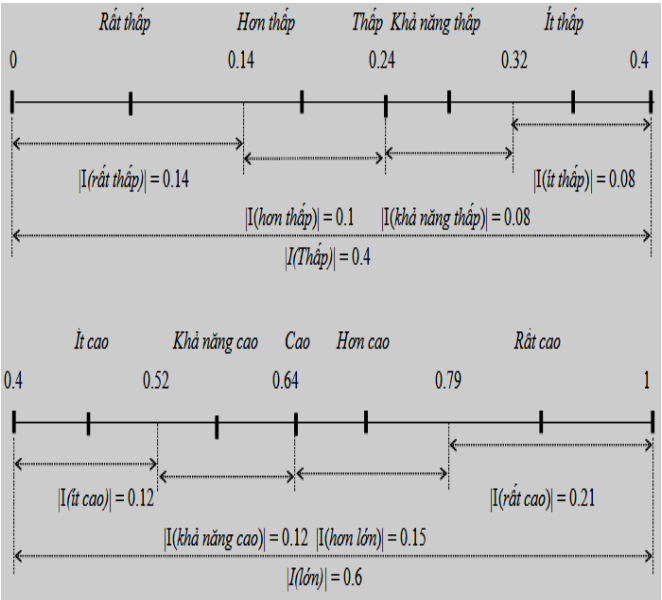
I(*ít cao*) = [0.4, 0.52],

I(*khả năng cao*) = [0.52, 0.64],

I(*hơn cao*) = [0.64, 0.79],

I(*rất cao*) = [0.79, 1].

Ta có hình ảnh của sự phân bố tính mờ của thuộc tính *LươngTháng* được mô tả tóm tắt như ở Hình 2.



Hình 2. *Tính mờ của LươngTháng.*

Như vậy, với thuộc tính không thuần nhất *LươngTháng* đang xét, ta có

$A_{LươngTháng}$= {48, Thấp, 53, 20, Cao, Cao, Rất cao, 35,100, 50, Ít thấp, 55, 50},

Chọn ψ_1 = 100 $\in X_{LươngTháng}$ khi đó $\forall\omega \in$ *Num*(*LươngTháng*),

IC(ω) = {0.35, 0.24, 0.41, 0, 0.64, 0.64, 1, 0.19, 1, 0.38, 0.4, 0.44, 0.38}.

Tiếp theo, ta phải tính cho các giá trị ngoại lai đã xác định ở trên. Ta sẽ chọn các phân hoạch thích hợp của các khoảng mờ để chèn các giá trị ngoại lai vào các khoảng này.

Do giá trị ngoại lai *quá cao* > *rất cao* nên ta sẽ phân hoạch đoạn [0.79, 1] tương ứng của |*I*(*lớn*)|.

Như vậy ta có:

fm(quá *cao*) ~ *fm*(*rất cao*) × *I*(*rất cao*) = 0.21 × 0.79 = 0.17

Nên *I*(*rất cao*) = [0.79, 0.96],

I(*quá cao*) = [0.96, 1].

Quá thấp < *rất thấp* nên ta sẽ phân hoạch đoạn [0, 0.14] tương ứng của |*I*(*thấp*)|.

Như vậy: *fm*(*quá thấp*) ~ *fm*(*rất thấp*) × *I*(*rất thấp*) = 0.14 × 0.14 = 0.02

Nên *I*(*rất thấp*) = [0.02, 0.14],

I(*quá cao*) = [0, 0.02].

Như vậy, thuộc tính *LươngTháng* nhận được như sau:

$A_{LươngTháng}$= *{48, Thấp, 53, 20, Cao, Cao, Rất cao, 35, 100, 50, Quá cao, Ít thấp, 55, 50, Quá thấp},*

IC(ω) = {0.35, 0.24, 0.41, 0.02, 0.64, 0.64, 0.79, 0.19, 0.79, 0.38, 1, 0.4, 0.44, 0.38, 0}.

Do đó, Φ_2(0.35) = *ít thấp* vì 0.35 $\in$ *I*(*ít thấp*).

Tương tự cho các giá trị còn lại, ta có thuộc tính *LươngTháng* theo ngữ nghĩa sẽ như sau:

{Ít thấp, Thấp, Ít cao, Rất thấp, Cao, Cao, Rất cao, Hơn thấp, Rất cao, Ít thấp, Quá cao, Ít thấp, Ít cao, Ít thấp, Quá thấp}.

Lúc này, thuộc tính *LươngTháng* sẽ được làm thuần nhất theo giá trị là:

{48, 24, 53, 20, 64, 64, 100, 35, 100, 50, 130, 50, 40, 55, 50, 12}.

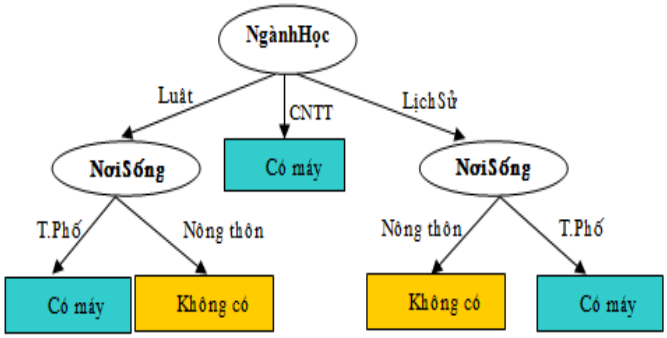
V. NHẬN XÉT VÀ ĐÁNH GIÁ

Với tập mẫu huấn luyện DIEUTRA đã cho ở Bảng 1, chứa thuộc tính chưa thuần nhất *LươngTháng*, ta có:

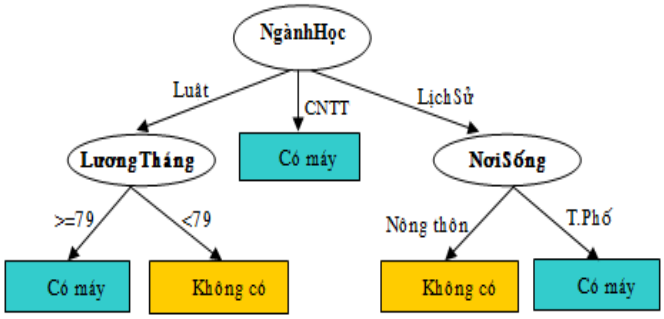
1. Không làm thuần nhất dữ liệu mà loại bỏ các giá trị ngôn ngữ, xem giá trị ngôn ngữ là giá trị "lỗi", ta thu được cây như ở Hình 3.

Như vậy, quá trình loại các mẫu "lỗi" đã làm cho thuộc tính *LươngTháng* trong mẫu không đủ hữu hiệu để xất hiện trên cây kết quả, điều này không phản ánh hết thông tin thực tế.

2. Không xét các giá trị ngoại lai như đã đề cập ở mục 4 mà chỉ thực hiện việc thuần nhất dữ liệu các giá trị còn lại. Kết quả sau khi thuần nhất giá trị, ta có tập giá trị của thuộc tính LươngTháng là: {45, 24, 52, 34, 64, 79, 35, 100, 50, 79, 40, 55, 50}. Tiến hành xây dựng cây, thu được cây kết quả ở Hình 4.

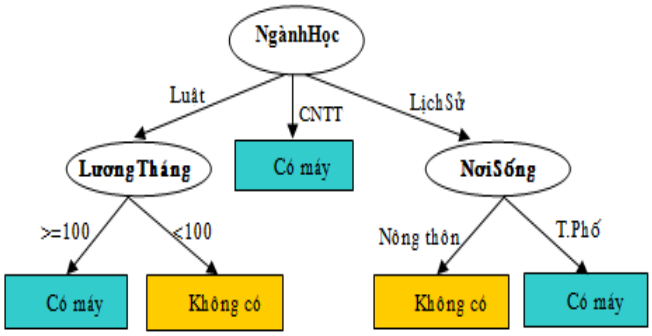


***Hình 3**. Cây quyết định được tạo khi bỏ qua các giá trị ngôn ngữ trong tập mẫu huấn luyện.*



***Hình 4.** Cây quyết định được tạo khi bỏ qua các giá trị ngoại lai trong tập mẫu huấn luyện.*

3. Xét riêng các giá trị ngoại lai như đã đề cập ở mục IV, kết quả sau khi thuần nhất giá trị, ta có tập giá trị của thuộc tính LươngTháng là: {48, 24, 53, 20, 64, 64, 100, 35, 100, 50, 130, 50, 40, 55, 50, 12}. Tiến hành xây dựng cây, thu được cây kết quả ở Hình 5.



***Hình 5**. Cây quyết định được tạo khi ta có xử lý các giá trị ngoại lai.*

Đối sánh 2 cây thu được ở Hình 4, Hình 5 và các giá trị rõ ở tập mẫu ban đầu, ta thấy khi bỏ qua các giá trị ngoại lai sẽ có thể dẫn đến các sai khác ở các vị trí biên trên cây kết quả. Như vậy, việc xử lý các giá trị ngoại lai là thực sự cần thiết.

VI. KẾT LUẬN

Bài báo đã đánh giá tính phức tạp của tập mẫu huấn luyện khi trích chọn từ dữ liệu nghiệp vụ thông qua việc phân tích tính đa dạng của miền trị thuộc tính. Bài báo cũng đã chỉ ra tính phức tạp khi định lượng giá trị ngôn ngữ đặc biệt là các giá trị ngôn ngữ ngoại lai trong tập mẫu huấn luyện. Trên cơ sở của đại số gia tử, bằng việc xem xét tính hiệu quả khi làm thuần nhất giá trị cho các thuộc tính chưa thuần nhất trong mẫu theo giá trị ngôn ngữ hay theo giá trị kinh điển. Bài báo đã chỉ ra một cách thức để có thể xác định được giá trị rõ cho các giá trị ngôn ngữ ngoại lai trong tập huấn luyện, để từ đó ta có thể huấn luyện được cây quyết định phù hợp với thực tế.

TÀI LIỆU THAM KHẢO

[1] DƯƠNG THĂNG LONG: *Phương pháp xây dựng hệ mờ dạng luật với ngữ nghĩa dựa trên đại số gia tử và ứng dụng trong bài toán phân lớp*, Luận án Tiến sĩ Toán học, Viện Công nghệ Thông tin, 2010.

[2] ĐOÀN VĂN BAN, LÊ MẠNH THẠNH, LÊ VĂN TƯỞNG LÂN: *Một cách chọn mẫu huấn luyện và thuật toán học để xây dựng cây quyết định trong khai phá dữ liệu*, Tạp chí Tin học và Điều khiển học, T23, S4, 2007.

[3] NGUYỄN CÁT HỒ: *Lý thuyết tập mờ và Công nghệ tính toán mềm*, Tuyển tập các bài giảng về Trường thu hệ mờ và ứng dụng, 2006.

[4] NGUYỄN CÁT HỒ: *Cơ sở dữ liệu mờ với ngữ nghĩa đại số gia tử*, Bài giảng trường Thu - Hệ mờ và ứng dụng, Viện Toán học Việt Nam, 2008.

[5] NGUYỄN CÔNG HÀO, NGUYỄN CÁT HỒ: *Một cách tiếp cận xấp xỉ dữ liệu trong cơ sở dữ liệu mờ*, Tạp chí Tin học và Điều khiển học, 2006.

[6] NGUYỄN CÔNG HÀO: *Cơ sở dữ liệu mờ với thao tác dữ liệu dựa trên đại số gia tử*, Luận án Tiến sĩ Toán học, Viện Công nghệ Thông tin, 2008.

[7] LÊ VĂN TƯỜNG LÂN: *Phụ thuộc dữ liệu và tác động của nó đối với bài toán phân lớp của khai phá dữ liệu*, Tạp chí khoa học Đại học Huế, Tập:19, Số:53, 2009.

[8] LÊ VĂN TƯỜNG LÂN: *Một cách tiếp cận chọn tập mẫu huấn luyện cây quyết định dựa trên đại số gia tử, Hội nghị Quốc gia lần thứ VI về nghiên cứu cơ bản và ứng dụng Công nghệ Thông tin (FAIR),* XNB Khoa học tự nhiên và công nghệ, 2013.

[9] PHẠM HẠ THỦY: *Xác định phần tử ngoại lai trong cơ sở dữ liệu quan hệ*, Tạp chí Tin học và Điều khiển học, T21, S4, 2005.

[10] A.K. BIKAS, E. M. VOUMVOULAKIS AND N. D. HATZIARGYRIOU, *Neuro-Fuzzy Decision Trees for Dynamic Security Control of Power Systems*, Department of Electrical and Computer Engineering, Greece, 2008.

[11] CHIDA A, *Enhanced Encoding with Improved Fuzzy Decision Tree Testing Using CASP Templates*, Computational Intelligence Magazine, IEEE, 2012.

[12] CHANG, ROBIN L. P. PAVLIDIS, *Fuzzy Decision Tree Algorithms, Man and Cybernetics*, IEEE, 2007.

[13] DORIAN P, *Data Preparation for Data Mining, Morgan Kaufmann*, 1999.

[14] DAVEEDU R. A., JAYA SUMA. G, LAVANYA DEVI G. *Construction of Fuzzy Decision Tree using Expectation Maximization Algorithm*, International Journal of Computer Science and Management Research, 2012.

[15] E. M. KNORR, Outliers and data mining: *finding exceptions in data*, Doctor' thesis, Dept. of Computer science, University of British Columbia, 2002.

[16] FERNANDEZ A., CALDERON M, BARRENECHEA E., *Enhancing Fuzzy Rule Based Systems in Multi-Classication Using Pairwise Coupling with Preference Relations*, EUROFUSE Workshop Preference Modelling and Decision Analysis, Public University of Navarra, Pamplona, Spain, 2009.

[17] FA. CHAO LI, JUAN SUN, XI-ZHAO WANG, *Analysis on the fuzzy filter in fuzzy decision trees*, Proceedings of the Second International Conference on Machine Learxung and Cybernetics, 2003.

[18] KAVITA SACHDEVA, MADASU HANMANDLU, AMIOY KUMAR, *Real Life Applications of Fuzzy Decision Tree*, International Journal of Computer Applications, 2012.

[19] HESHAM A. HEFNY, AHMED S. GHIDUK, ASHRAF ABDEL WAHAB, *Effective Method for Extracting Rules from Fuzzy Decision Trees based on Ambiguity and Classifiability*, Universal Journal of Computer Science and Engineering Technology, Cairo University, Egypt, 2010.

[20] HO TU BAO, *Introduction to knowledge discovery and data mining*, Institute of Information Technology National Center for Natural Science, 2000.

HO N. C. AND NAM H. V, *An algebraic approach to linguistic hedges in Zadeh's fuzzy logic, Fuzzy Sets and Systems*, vol.129, pp.229-254, 2002.

[21] MOUSTAKIDIS, S. MALLINIS, G. KOUTSIAS, N. THEOCHARIS, J.B., PETRIDIS, V, SVM-*Based Fuzzy Decision Trees for Classification of High Spatial Resolution Remote Sensing Images*, Geoscience and Remote Sensing, IEEE, 2012.

[22] OLEKSANDR DOROKHOV, VLADIMIR CHERNOV, *Application of the fuzzy decision trees for the tasks of alternative choices, Transport and Telecommunication Institute, Lomonosova*, Latvia, Vol.12, No 2, 2011.

[23] V. BARNETT, T. LEWIS, *Outliers in Statistical Data, John Wiley*, 3rd edition, 1994.

Ngày nhận bài: 08/04/2015

SƠ LƯỢC VỀ TÁC GIẢ

LÊ VĂN TƯỜNG LÂN

Sinh năm 1974 tại TP. Huế.

Nhận bằng thạc sỹ Tin học, chuyên ngành CNTT tại Trường ĐH Bách khoa Hà Nội, năm 2002. Đang là NCS tại trường ĐH Khoa học – ĐH Huế, chuyên ngành Khoa học Máy tính.

Hiện công tác tại khoa CNTT, Trường ĐH Khoa học, ĐH Huế.

Lĩnh vực nghiên cứu: Khai phá dữ liệu, công nghệ phần mềm.

Điện thoại liên hệ: 0905151357.

Địa chỉ email: nmhan2009@gmail.com

NGUYỄN MẬU HÂN

Sinh năm 1957 tại Thừa thiên Huế.

Nhận bằng tiến sĩ tại Viện CNTT. Được phong hàm Phó Giáo sư năm 2013.

Hiện là giảng viên chính tại khoa CNTT, Trường ĐH Khoa học, ĐH Huế.

Lĩnh vực nghiên cứu: Xử lý song song và phân tán, tính toán lưới và điện toán đám mây.

Email: nchao@hueuni.edu.vn

NGUYỄN CÔNG HÀO

Sinh năm 1976 tại Thừa thiên Huế.

Nhận bằng tiến sĩ tại Viện CNTT năm 2008.

Hiện là Giám đốc Trung tâm CNTT, ĐH Huế.

Lĩnh vực nghiên cứu: Cơ sở dữ liệu mờ, các phương pháp tính toán mềm, các phương pháp lập luận xấp xỉ.

Địa chỉ email: nchao@hueuni.edu.vn