MỘT PHƯƠNG PHÁP ĐỊNH LƯỢNG GIÁ TRỊ NGÔN NGỮ TRONG TẬP MẪU HUẤN LUYỆN ĐỂ XÂY DỰNG CÂY QUYẾT ĐỊNH MỜ

Lê Văn Tường Lân[1], Nguyễn Mậu Hân[1], Nguyễn Công Hào[2]

[1] Khoa Công nghệ Thông tin, Đại học Khoa học, Đại học Huế

[2] Trung tâm Công nghệ Thông tin, Đại học Huế

lvtlan@yahoo.com, nmhan2005@yahoo.com, nchao@hueuni.edu.vn

Tóm tắt. Tập mẫu đóng vai trò quan trọng trong quá trình huấn luyện. Khi miền trị của các thuộc tính trong tập mẫu huấn luyện là chưa thuần nhất, việc làm thuần nhất tập huấn luyện là bắt buộc. Đại số gia tử là một công cụ hữu ích để làm thuần nhất tập huấn luyện, bằng cách chuyển miền dữ liệu của thuộc tính chưa thuần nhất thành miền dữ liệu chứa các giá trị ngôn ngữ hay định lượng các giá trị ngôn ngữ về các giá trị kinh điển. Trong quá trình thuần nhất, ta cần phải biết các giá trị ψ_{min}, ψ_{max} *của* miền trị kinh điển. Tuy nhiên, trong thực tế, nhiều lúc ta chưa biết cụ thể giá trị ψ_{min}, ψ_{max} của thuộc tính đang xét. Trong bài báo này, chúng ta xây dựng một cách thức để có thể các định lượng cho các giá trị ngôn ngữ khi không biết miền giá trị [ψ_{min}, ψ_{max}] mà chỉ biết đoạn con [ψ_1, ψ_2] của chúng.

Từ khoá: *Tập mẫu huấn luyện, Giá trị ngôn ngữ, Đại số gia tử, Cây quyết định mờ.*

I. Đặc vấn đề

Cho một tập huấn luyện, tất cả các mẫu của tập đều có chung một cấu trúc, gồm những cặp <*Thuộc tính, Giá trị*>, một trong những thuộc tính này đại diện cho lớp và ta gọi là thuộc tính dự đoán hay thuộc tính phân lớp. Bài toán phân lớp là bài toán tìm quy tắc xếp các đối tượng vào một trong các lớp đã cho dựa trên tập mẫu huấn luyện. Có nhiều phương pháp tiếp cận bài toán phân lớp: Hàm phân biệt tuyến tính Fisher, Naïve Bayes, Logistic, Mạng nơ-ron, Cây quyết định, … trong đó phương pháp cây quyết định là phương pháp phổ biến do tính trực quan, dễ hiểu và hiệu quả của nó [2, 18]. 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 mẫu huấn luyện M, về cơ bản, các thuật toán phân lớp phải thực hiện 2 bước sau:

Bước 1: Chọn thuộc tính A_J có các giá trị $a_1, a_2,\ldots, a_n$;

Bước 2: Với thuộc tính A_J được chọn, ta tạo một nút của cây và sau đó chia các mẫu ứng với nút này thành các tập tương ứng $M_1, M_2, \ldots, M_n$; Sau đó lại tiếp tục [17]. 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][12][18].

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, 8, 12]. 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.

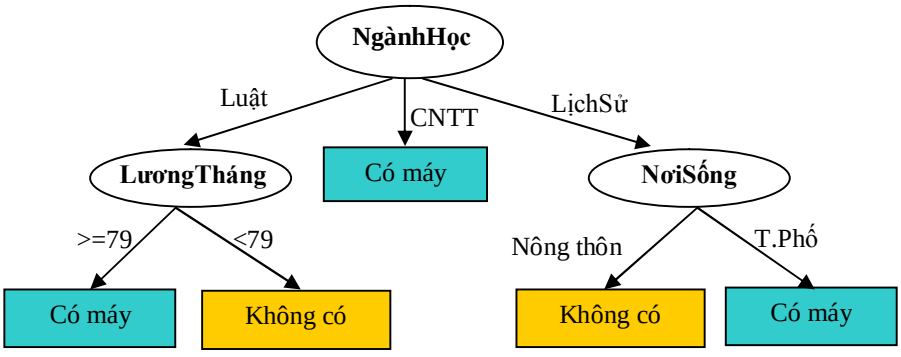
Ví dụ 1: Cho bảng dữ liệu DIEUTRA lưu trữ về tình hình mua máy tính xách tay của khách hàng tại một công ty như bảng 1, cần chọn mẫu huấn luyện để xây dựng cây quyết định cho việc dự đoán khách hàng mua máy hay không.

Bảng 1: Tập mẫu có thuộc tính với dữ liệu không nhất quán (***LươngTháng***)

PhiếuĐT	HọVàTên	NơiSống	NgànhHọc	KinhTếGĐ	LươngTháng	MáyTính
M01045	Nguyễn Văn An	T.Phố	Luật	Chưa tốt	45	Không
M01087	Lê Văn Bình	NôngThôn	Luật	Chưa tốt	Thấp	Không
M02043	Nguyễn Thị Hoa	T.Phố	CNTT	Chưa tốt	52	Có
M02081	Trần Bình Thám	T.Phố	LịchSử	Trung bình	20	Có
M02046	Trần Thị Hương	T.Phố	LịchSử	Khá	Cao	Có
M03087	Nguyễn Thị Lài	NôngThôn	LịchSử	Khá	Cao	Không
M03025	Vũ Tuấn Hoa	NôngThôn	CNTT	Khá	Rất cao	Có
M03017	Lê Bá Linh	T.Phố	Luật	Trung bình	35	Không
M04036	Bạch Ân Lai	T.Phố	Luật	Khá	100	Có
M04037	Lý Thị Hoa	T.Phố	LịchSử	Trung bình	50	Có
M04042	Vũ Quang Bình	NôngThôn	Luật	Trung bình	Rất cao	Có
M04083	Nguyễn Thị Hoa	NôngThôn	CNTT	Trung bình	Ít thấp	Có
M05041	Lê Xuân Hoan	T.Phố	CNTT	Chưa tốt	55	Có
M05080	Trần Quế Chung	NôngThôn	LịchSử	Trung bình	50	Không

Trong thời gian qua, đại số gia tử được nhiều nhóm tác giả 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, 11-17, 21]. 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 ví dụ 1, miền trị của thuộc tính *LươngTháng* trong Bảng 1 được thuần nhất theo giá trị ngôn ngữ là: *{Ít cao, Thấp, Khả năng cao, Ít thấp, Cao, Cao, Rất cao, Ít thấp, Rất cao, Khả năng cao, Rất cao, Ít thấp, Khả năng cao, Khả năng cao}* hay miền trị sau khi được định lượng giá trị là: {45, 24, 52, 34, 64, 64, 79, 35,100, 50, 79, 40, 55, 50} với miền trị kinh điển của thuộc tính *LươngTháng* trong tập mẫu được xác định là *Dom*(*LươngTháng*) = $[\psi_{min}, \psi_{max}]$ =[20,100]. Cây quyết định sau khi huấn luyện như hình 1.



Hình 1. Cây quyết định được tạo sau khi làm thuần nhất giá trị cho thuộc tính *LươngTháng* ở bảng 3 dựa theo ĐSGT.

Tuy vậy, khi định lượng giá trị ngôn ngữ, không phải lúc nào ta cũng tìm được các giá trị ψ_{min}, ψ_{max} trong tập dữ liệu. Với việc không thể tìm được miền giá trị kinh điển $[\psi_{min}, \psi_{max}]$ trong thuộc tính đang xét của tập mẫu huấn, ta phải nhờ ý kiến của chuyên gia để xác định chúng và sau đó tiếp tục công việc, như tập mẫu huấn luyện ở Bảng 2, ta nhờ chuyên gia để xác định $[\psi_{min}, \psi_{max}]$ =[20,100] và sau đó tiếp tục.

Bảng 2: Tập mẫu có thuộc tính không tìm được miền [ψ_{min}, ψ_{max}] (***LươngTháng***)

NơiSống	NgànhHọc	KinhTếGĐ	LươngTháng	MáyTính
T.Phố	Luật	Chưa tốt	Ít cao	Không
NôngThôn	Luật	Chưa tốt	Thấp	Không
T.Phố	CNTT	Chưa tốt	Khả năng cao	Có
T.Phố	LịchSử	Trung bình	Rất thấp	Có
T.Phố	LịchSử	Khá	Cao	Có
NôngThôn	LịchSử	Khá	65	Không
NôngThôn	CNTT	Khá	Rất cao	Có
T.Phố	Luật	Trung bình	30	Không
T.Phố	Luật	Khá	Rất cao	Có
T.Phố	LịchSử	Trung bình	Khả năng cao	Có
NôngThôn	Luật	Trung bình	Rất cao	Có
NôngThôn	CNTT	Trung bình	Ít thấp	Có
T.Phố	CNTT	Chưa tốt	Khả năng cao	Có
NôngThôn	LịchSử	Trung bình	Khả năng cao	Không

Việc nhờ ý kiến của chuyên gia không phải lúc nào cũng thực hiện được và hơn nữa ta không thể tận dụng hết các thông tin đã lưu trữ trong tập mẫu huấn luyện. Trong bài báo này, chúng tôi sẽ trình bày một cách để có thể định lượng cho các giá trị ngôn ngữ khi không tìm thấy miền trị kinh điển [ψ_{min}, ψ_{max}] trong thuộc tính đang xét của tập huấn luyện dựa vào đại số gia tử.

II. Đại số gia tử

Với tập mẫu huấn luyện *M*, ta xét miền trị của biến ngôn ngữ của thuộc tính chưa thuần nhất có trong tập mẫu *M*. Với *LươngTháng* là thuộc tính chưa thuần nhất, *ta có Dom*(*LươngTháng*) = {*cao, thấp, rất cao, rất thấp, ít cao, ít thấp, cao hơn, thấp hơn, khả năng cao, khả năng thấp,...}* trong đó *cao*, *thấp* là các từ nguyên thuỷ, các từ nhấn *rất, hơn, ít, khả năng* gọi là các gia tử. Khi đó miền ngôn ngữ *T = Dom(LươngTháng)* có thể biểu thị như một đại số $\underline{X}$=*(**X**, G, H, ≤)*, trong đó *G* là tập các từ nguyên thuỷ *{thấp, cao}* được xem là các phần tử sinh. *H* là tập các gia tử được xem như là các phép toán một ngôi, quan hệ "≤ " trên các từ là quan hệ thứ tự được "cảm sinh" từ ngữ nghĩa tự nhiên.

Tập ***X*** được sinh ra từ *G* bởi các phép tính trong *H*. Như vậy mỗi phần tử của ***X*** sẽ có dạng biểu diễn $x = h_n h_{n-1}.....h_{1x}$, $x \in G$. Tập tất cả các phần tử được sinh ra từ một phần tử *x* được ký hiệu là *H(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 $c^- < c^+$. Ở đây, *cao* là dương còn *thấp* là âm.

Cho đại số gia tử $\underline{X}$ = *(**X**, G, H, ≤)*, với $G = \{c^+, c^-\}$, trong đó c^+ và c^- tương ứng là phần tử sinh dương và âm, ***X*** là tập nền. $H = H^+ \cup H^-$ với $H^- = \{h_1, h_2, .., h_p\}$ và $H^+ = \{h_{p+1}, ..., h_{p+q}\}$, $h_1 > h_2 > ... > h_p$ và $h_{p+1} < ... < h_{p+q}$.

1. Hàm định lượng ngữ nghĩa [3, 5]: $f: X \to [0,1]$ gọi là hàm định lượng ngữ nghĩa của *X* nếu $\forall h, k \in H^+$ hoặc $\forall h, k \in H^-$ và $\forall x, y \in X$, ta có : $\left|\frac{f(hx)-f(x)}{f(kx)-f(x)}\right| = \left|\frac{f(hy)-f(y)}{f(ky)-f(y)}\right|$

Với đại số gia tử và hàm định lượng ngữ nghĩa ta có thể định nghĩa *tính mờ* của một khái niệm mờ. Cho trước hàm định lượng ngữ nghĩa *f* của *X*. Xét bất kỳ $x \in X$. Tính mờ của x khi đó được đo bằng đường kính của tập $f(H(x)) \subseteq [0,1]$

2. Độ đo tính mờ [3, 5]: Hàm $fm: X \to [0,1]$ được gọi là *độ đo tính mờ* trên ***X*** nếu thoả mãn các điều kiện sau:

(1) $fm(c^-) = W > 0$ và $fm(c^+) = 1 - W > 0$

(2) Với $c \in \{c^-, c^+\}$ thì $\sum_{i=1}^{p+q} fm(h_i c) = fm(c)$.

(3) Với $\forall x,y \in \boldsymbol{X}$, $\forall h \in H$, $\frac{fm(hx)}{fm(x)} = \frac{fm(hy)}{fm(y)} = \frac{fm(hc)}{fm(c)}$, với $c \in \{c^-, c^+\}$,

nghĩa là tỉ số này không phụ thuộc vào *x* và *y*, được kí hiệu là $\mu(h)$ gọi là độ đo tính mờ của gia tử *h*.

3. Một số tính chất của độ đo tính mờ. [3-6]

(i) $fm(hx) = \mu(h) fm(x)$, với $\forall x \in X$

(ii) $\sum_{i=1}^{p+q} fm(h_i c) = fm(c)$, trong đó $c \in \{c-, c^+\}$

(iii) $\sum_{i=1}^{p+q} fm(h_i x) = fm(x)$, với $\forall x \in X$

(iv) $\sum_{i=1}^{p} \mu(h_i) = \alpha$ và $\sum_{i=p+1}^{p+q} \mu(h_i) = \beta$, với $\alpha, \beta > 0$ và $\alpha + \beta = 1$.

4. Chuyển giá trị ngôn ngữ về giá trị số [8]: Để chuyển đổi một giá trị ngôn ngữ trong ĐSGT thành một số trong [0,1] ta sử dụng hàm định lượng ngữ nghĩa ν của *X* được xây dựng như sau với $x = h_{im}...h_{i2}h_{i1}c$:

(1) $\nu(c^-) = W - \alpha.fm(c^-)$ và $\nu(c^+) = W + \alpha.fm(c^+)$

(2) $\nu(h_j x) = \nu(x) + Sign(h_j x) \times \left[\sum_{i=j}^{p} fm(h_i x) - \frac{1}{2}(1 - Sign(h_j x) Sign(h_1 h_j x)(\beta - \alpha)) fm(h_j x) \right]$ với $1 \le j \le p$, và

$\nu(h_j x) = \nu(x) + Sign(h_j x) \times \left[\sum_{i=p+1}^{j} fm(h_i x) - \frac{1}{2}(1 - Sign(h_j x) Sign(h_1 h_j x)(\beta - \alpha)) fm(h_j x) \right]$ với $j > p$

5. Chuyển giá trị số về giá trị ngôn ngữ [8]: Để chuyển một giá trị số về một giá trị thuộc [0,1], ta có hàm $IC: Dom(A_i) \to [0,1]$ được xác định như sau:

- Nếu $LD_{Ai} = \emptyset$ và $D_{Ai} \neq \emptyset$ thì $\forall \omega \in Dom(A_i)$ ta có: $IC(\omega) = 1 - \frac{\psi_{\max} - \omega}{\psi_{\max} - \psi_{\min}}$, với $Dom(A_i) = [\psi_{\min}, \psi_{\max}]$ là miền trị kinh điển của A_i.

- Nếu $D_{Ai} \neq \emptyset$, $LD_{Ai} \neq \emptyset$ thì $\forall \omega \in Dom(A_i)$ ta có $IC(\omega) = \{\omega * \nu(\psi_{maxLV})\}/\psi_{\max}$, 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 $\nu(\psi_{maxLV}) \approx 1.0$ thì $(\{\omega * \nu(\psi_{maxLV})\}/\psi_{\max}) \approx 1 - \frac{\psi_{\max} - \omega}{\psi_{\max} - \psi_{\min}}$

6. Hàm ngược của hàm định lượng ngữ nghĩa [8]: Cho đại số gia tử $\underline{X} = (X, G, H, \le)$, ν là hàm định lượng ngữ nghĩa của *X*. Φ_k: $[0,1] \to X$ gọi là hàm ngược của hàm ν 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 $x^k \in X^k$.

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 $\Phi_k(a) <_k \Phi_k(b)$.

III. Định lượng giá trị ngôn ngữ khi không tìm được miền giá trị kinh điển *Min, Max*

Như thế, với bất kỳ một thuộc tính không thuần nhất A, ta sẽ chuyển về giá trị ngôn ngữ để rồi có thể chuyển về giá trị số thuần nhất. Trong tập mẫu đã cho ở bảng 1, ta sẽ 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*$\}$ với quan hệ ngữ nghĩa: *rất* > *hơn* và *ít* >*khả năng*. $W_{LươngTháng} = 0.6$, $fm(thấp) = 0.4$, $fm(cao) = 0.6$, $fm(rất) = 0.35$, $fm(hơn) = 0.25$, $fm(khả\ năng) = 0.20$, $fm(ít) = 0.20$.

Lúc này ta có : $fm(rất\ thấp) = 0.35x0.4=0.14$, $fm(hơn\ thấp) =025x0.4=0.10$, $fm(ít\ thấp) = 0.2x0.4=0.08$, $fm(khả\ năng\ thấp) =0.2x0.4= 0.08$. 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: $fm(rất\ cao) = 0.35x0.6=0.21$, $fm(hơn\ cao) =025x0.6=0.15$, $fm(ít\ cao) = 0.2x0.6=0.12$, $fm(khả\ năng\ cao) =0.2x0.6= 0.12$. Vì *ít cao* < *khả năng cao* < *cao* < *hơn cao* < *rất cao* nên : $I(ít\ cao) = [0.4,0.52]$, $I(khả\ năng\ cao) = [0.52,0.64]$, $I(hơn\ cao) = [0.64,0.79]$, $I(rất\ cao) = [0.79,1]$.

Vậy, với $U_{LươngTháng}$= {45, Thấp, 52, 34, Cao, Cao, Rất cao, 35, 100, 50, Rất cao, Ít thấp, 55, 50}, $[\psi_{min}, \psi_{max}] = [20,100]$, ta tìm được $IC(\omega)$ = {0.45, 0.24, 0.52, 0.34, 0.64, 0.64, 0.79, 0.35, 1, 0.50, 0.79, 0.4, 0.55, 0.50}. Giá trị mờ của thuộc tính *LươngTháng* là *{Ít cao, Thấp, Khả năng cao, Ít thấp, Cao, Cao, Rất cao, Ít thấp, Rất cao, Khả năng cao, Rất cao, Ít thấp, Khả năng cao, Khả năng cao*} nên sau khi định lượng giá trị cho thuộc tính *LươngTháng* sẽ được các giá trị rõ là: {45, 24, 52, 34, 64, 64, 79, 35,100, 50, 79, 40, 55, 50}.

Tuy vậy, quá trình định lượng cho các giá trị ngôn ngữ ở trên chỉ thực hiện được khi chúng ta có thể tìm được miền trị kinh điển $[\psi_{min}, \psi_{max}]$ của thuộc tính đang xét, ở đây là [*20, 100*]. Trong trường hợp không tìm thấy miền trị này thì giải thuật trên không thể áp dụng.

1. Định lượng giá trị ngôn ngữ khi biết một đoạn con của $[\psi_{min}, \psi_{max}]$ và toàn bộ $IC(\omega)$

Cho thuộc tính không thuần nhất A_i, lúc này ta có $Dom(A_i) = D_{Ai} \cup LD_{Ai}$ nhưng giá trị biên $[\psi_{min}, \psi_{max}]$ đối với miền trị kinh điển D_{Ai} của A_i không được xác định, mà ta chỉ biết một đoạn con $[\psi_1, \psi_2]$ tương ứng giá trị ngôn ngữ $[\psi_{LV1}, \psi_{LV2}]$ của LD_{Ai} *và* tất cả các giá trị định lượng mờ $IC(\omega)$ của chúng.

Ví dụ như thuộc tính LươngTháng ở Bảng 2, giá trị mờ của thuộc tính LươngTháng là {Ít cao, Thấp, Khả năng cao, Ít thấp, Cao, Cao, Rất cao, Ít thấp, Rất cao, Khả năng cao, Rất cao, Ít thấp, Khả năng cao, Khả năng cao}. IC(ω) = {0.45, 0.24, 0.52, 0.34, 0.64, 0.64, 0.79, 0.35, 1, 0.50, 0.79, 0.4, 0.55, 0.50}. Ở đây, ta không biết $[\psi_{min}, \psi_{max}]$ tương ứng với giá trị ngôn ngữ $[\psi_{minLV}, \psi_{maxLV}]$ = [Rất thấp, Rất cao] mà chỉ biết đoạn con có miền trị là $[\psi_1, \psi_2]$ = [30, 65] tương ứng với miền trị của ngôn ngữ là $[\psi_{LV1}, \psi_{LV2}]$ = [Ít thấp, Hơn cao].

Lúc này, do $IC(\omega) = 1 - \frac{\psi_{max} - \omega}{\psi_{max} - \psi_{min}}$ nên tất cả các ω nằm giữa $[\psi_1, \psi_2]$ sẽ đúng với quy tắc này.

Hơn nữa, do độ lớn của các ω sẽ tỷ lệ với bán kính $f(H(x)) \subseteq [0,1]$ tức là $\omega_1 > \omega_2$ lớn khi $IC(\omega_1) > IC(\omega_2)$ và $\frac{\omega_1}{IC(w_1)} = \frac{\omega_2}{IC(w_2)}$ khi tất cả các $IC(\omega_1)$, $IC(\omega_2)$ về cùng một phía với W. Do vậy, giá trị định lượng cho các giá trị ngôn ngữ này được tính theo giải thuật như sau:

B1: Với ω mà giá trị ngôn ngữ tương ứng trong đoạn $[\psi_{LV1}, \psi_{LV2}]$, ta có: $\omega = IC(w)(\psi_2 - \psi_1) + \psi_1$

B2: Với ω mà giá trị ngôn ngữ tương ứng trong đoạn $[\psi_{LV2}, \psi_{maxLV}]$, ta tính tuần tự tăng theo đoạn $\psi_{LV2}..\psi_{maxLV}$, với $\omega_i = \psi_2 \frac{IC(w_2)}{IC(w_i)}$ và dịch chuyển vị trí ψ_{LV2} đến vị trí i vừa tìm được.

B3: Với ω mà giá trị ngôn ngữ tương ứng trong đoạn $[\psi_{minLV}, \psi_{LV1}]$, ta tính tuần tự giảm theo đoạn $\psi_{LV1}..\psi_{minLV}$, với $\omega_i = \psi_1 \frac{IC(w_1)}{IC(w_i)}$ và dịch chuyển vị trí ψ_{LV1} lùi về vị trí i vừa tìm được.

Ví dụ 1: Cho 1 ĐSGT để mô tả thuộc tính không thuần nhất *LươngTháng* trong Bảng 2 như sau: $\underline{X}_{LươngTháng}$ = ($X_{LươngTháng}$, $G_{LươngTháng}$, $H_{LươngTháng}$, ≤), với $G_{LươngTháng}$ = {*cao, thấp*}, $H^+_{LươngTháng}$ = {hơn, *rất*}, $H^-_{LươngTháng}$ = {*khả năng*, *ít*} với quan hệ ngữ nghĩa: *rất* > *hơn* và *ít* >*khả năng*. $W_{LươngTháng}$ = 0.6, *fm*(*thấp*) = 0.4, *fm*(*cao*) = 0.6, *fm*(*rất*) = 0.35, *fm*(*hơn*) = 0.25, *fm*(*khả năng*) = 0.20, *fm*(*ít*) = 0.20. Miền trị ngôn ngữ là {*Ít cao, Thấp, Khả năng cao, Ít thấp, Cao, Cao, Rất cao, Ít thấp, Rất cao, Khả năng cao, Rất cao, Ít thấp, Khả năng cao, Khả năng cao*}. $IC(\omega)$ = {0.45, 0.24, 0.52, 0.34, 0.64, 0.64, 0.79, 0.35, 1, 0.50, 0.79, 0.4, 0.55, 0.50}. Biết đoạn con có miền trị là $[\psi_1, \psi_2]$ = [30, 65] tương ứng với miền trị của ngôn ngữ là $[\psi_{LV1}, \psi_{LV2}]$ = [*Ít thấp, Hơn cao*]. Hãy định lượng các giá trị ngôn ngữ cho *LươngTháng*.

Ta có: *fm*(*rất thấp*) = 0.35x0.4=0.14, *fm*(*hơn thấp*) =025x0.4=0.10, *fm*(*ít thấp*) = 0.2x0.4=0.08, *fm*(*khả năng thấp*) =0.2x0.4= 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]. *fm*(*rất cao*) = 0.35x0.6=0.21, *fm*(*hơn cao*) =025x0.6=0.15, *fm*(*ít cao*) = 0.2x0.6=0.12, *fm*(*khả năng cao*) =0.2x0.6= 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].

B1: Tính các ω có giá trị ngôn ngữ trong đoạn [*Ít thấp, Hơn cao*]

$\omega_{Ít\ thấp} = IC(\omega_{Ít\ cao})(\psi_2-\psi_1)+\psi_1 = 0.4(65-30)+30=44$

$\omega_{Ít\ cao} = IC(\omega_{Ít\ cao})(\psi_2-\psi_1)+\psi_1 = 0.52(65-30)+30=48$

$\omega_{Khả\ năng\ cao} = IC(\omega_{Khả\ năng\ cao})(\psi_2-\psi_1)+\psi_1 = 0.64(65-30)+30=52$

B2: Tính các ω có giá trị ngôn ngữ trong đoạn [*Hơn cao, Rất cao*]

$\omega_{Hơn\ cao} = \psi_2 * IC(\omega_{Khả\ năng\ cao})/IC(\omega_{Hơn\ cao}) = 65 * 0.64 / 0.52 = 80$

$\omega_{Rất\ cao} = \psi_2 * IC(\omega_{Hơn\ cao})/IC(\omega_{Rất\ cao}) = 80 * 0.79 / 0.64 = 99$

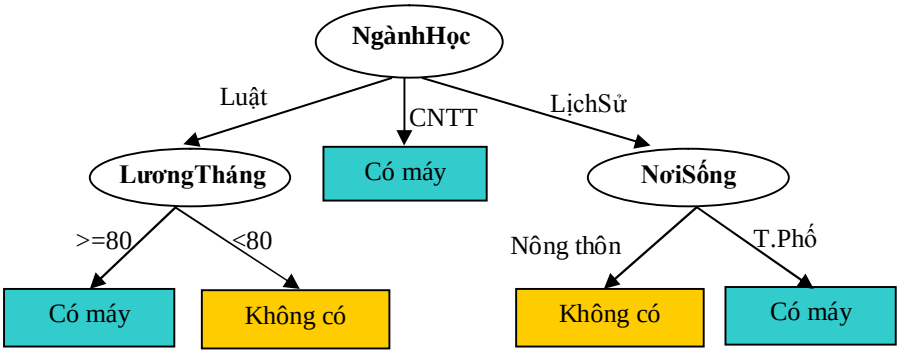
B3: Tính các ω có giá trị ngôn ngữ trong đoạn [*Rất thấp, Ít thấp*]

$\omega_{Khả\ năng\ thấp} = \psi_1 * IC(\omega_{Ít\ thấp})/IC(\omega_{Khả\ năng\ thấp}) = 30 * 0.32 / 0.4 = 24$

$\omega_{Hơn\ thấp} = \psi_1 * IC(\omega_{Khả\ năng\ thấp})/IC(\omega_{Hơn\ thấp}) = 24 * 0.24 / 0.32 = 18$

$\omega_{Rất\ thấp} = \psi_1 * IC(\omega_{Hơn\ thấp})/IC(\omega_{Rất\ thấp}) = 18 * 0.14 / 0.24 = 10$

Vậy miền trị sau khi được định lượng giá trị là: **{48, 18, 52, 30, 80, 80, 99, 30, 99, 52, 99, 30, 52, 52}**. Cây quyết định sau khi huấn luyện như hình 2.



Hình 2. Cây quyết định được tạo sau khi định lượng thuộc tính nhờ biết đoạn con của $[\psi_{min}, \psi_{max}]$ và toàn bộ $IC(\omega)$

2. Định lượng giá trị ngôn ngữ khi chỉ biết một đoạn con của $[\psi_{min}, \psi_{max}]$ nhưng chưa xác định được toàn bộ $IC(\omega)$

Cho thuộc tính không thuần nhất A_i, lúc này ta có $Dom(A_i) = D_{Ai} \cup LD_{Ai}$ nhưng giá trị biên $[\psi_{min}, \psi_{max}]$ đối với miền trị kinh điển D_{Ai} của A_i không được xác định, mà ta cũng chỉ tìm được một đoạn con $[\psi_1, \psi_2]$ của nó tương ứng giá trị ngôn ngữ $[\psi_{LV1}, \psi_{LV2}]$ của LD_{Ai} tức là $v(\psi_{LV1}) = IC(\psi_1)$ và $v(\psi_{LV2}) = IC(\psi_2)$. Lúc này ta phải tìm các giá trị $IC(\omega_i)$ còn lại tức các $IC(\omega_i)$ thỏa $IC(\psi_i) < IC(\psi_1)$ hoặc $IC(\psi_i) > IC(\psi_2)$

Do $IC(\omega) = 1 - \frac{\psi_{max} - \omega}{\psi_{max} - \psi_{min}}$ nên tất cả các ω nằm giữa $[\psi_1, \psi_2]$ sẽ đúng với quy tắc này, tức là $IC(\omega) = 1 - \frac{\psi_2 - \omega}{\psi_2 - \psi_1}$ với $\omega \in [\psi_2 - \psi_1]$. Do vậy có thể xây dựng một ĐSGT để định lượng giá trị cho chúng.

Theo phương pháp xây dựng ĐSGT đã nêu 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 $[0, v(\psi_i)]$ hay $[v(\psi_i), 1]$ để từ đó có xác định giá trị rõ cho các giá trị ngôn ngữ này. Đây chính là điểm để tính các $IC(\omega)$ không nằm trong đoạn $[\psi_1, \psi_2]$ bằng cách phân chia liên tiếp các đoạn con này để xác định các $IC(\omega_i)$ tương ứng. Vậy ta có giải thuật như sau

B1: Xây dựng 1 ĐSGT trong miền $[\psi_1, \psi_2]$ để tính các $IC(\omega)$ tương ứng cho các giá trị trong đoạn $[\psi_1, \psi_2]$ này.

B2: Tính lại các phân hoạch cho các $IC(\omega)$ như sau :

1. Nếu $\psi_i < \psi_1$ thì :

   - Phân hoạch đoạn $[0, v(\psi_1)]$ thành $[0, v(\psi_i)]$ và $[v(\psi_i), v(\psi_1)]$

   - Tính $fm(h_i) \sim fm(h_1)$ x $I(\psi_1)$ và $fm(h_1) = fm(h_1) - fm(h_i)$

2. Nếu $\psi_i > \psi_2$ thì :

   - Phân hoạch đoạn $[v(\psi_2), 1]$ thành $[v(\psi_2), v(\psi_i)]$ và $[v(\psi_i), 1]$

   - Tính $fm(h_i) \sim fm(h_2)$ x $I(\psi_2)$ và $fm(h_2) = fm(h_2) - fm(h_i)$

3. Tính giá trị $IC(\omega_i)$ và ψ_i tại vị trí i. Gán vị trí i đang có thành vị trí 1 và tiếp tục tính lùi với các giá trị còn lại với $\psi_i < \psi_1$ hay gán vị trí i đang có thành vị trí 2 và tiếp tục tính tiến với các giá trị còn lại với $\psi_i > \psi_2$

B3: Thực hiện định lượng các giá trị ngôn ngữ với cách tính ở mục 1 khi đã biết toàn bộ $IC(\omega)$.

Tính đúng của giải thuật: Do tất cả các phân hoạch trên không vượt ra khỏi đoạn đang xét là $|fm(h_1)|$ hay $|fm(h_2|$ nên không làm phá vỡ các phân hoạch đang có của đoạn [0,1], do $I(\psi_1)>0$ và $I(\psi_2)<1$, nên cách phân hoạch trên là phù hợp với phương pháp thuần nhất đã nêu ở mục II.

Ví dụ 2: Cho tập mẫu huấn luyện như ở Bảng 3. Hãy định lượng cho các giá trị ngôn ngữ ở thuộc tính *LươngTháng*.

Bảng 3: Tập mẫu có thuộc tính với dữ liệu không nhất quán, không tìm được miền $[\psi_{min}, \psi_{max}]$ (***LươngTháng***)

NơiSống	NgànhHọc	KinhTếGĐ	LươngTháng	MáyTính
T.Phố	Luật	Chưa tốt	48	Không
NôngThôn	Luật	Chưa tốt	Thấp	Không
T.Phố	CNTT	Chưa tốt	53	Có
T.Phố	LịchSử	Trung bình	Rất thấp	Có
T.Phố	LịchSử	Khá	Cao	Có
NôngThôn	LịchSử	Khá	80	Không
NôngThôn	CNTT	Khá	Rất cao	Có
T.Phố	Luật	Trung bình	30	Không
T.Phố	Luật	Khá	80	Có
T.Phố	LịchSử	Trung bình	50	Có
NôngThôn	Luật	Trung bình	Rất cao	Có
NôngThôn	CNTT	Trung bình	Ít thấp	Có
T.Phố	CNTT	Chưa tốt	55	Có
NôngThôn	LịchSử	Trung bình	50	Không

Tập mẫu có thuộc tính *LươngTháng* là chưa thuần nhất nên ta phải thuần nhất các giá trị cho *LươngTháng*. Ta *có:* $Dom(LươngTháng) = D_{LươngTháng} \cup LD_{LươngTháng}$. $D_{LươngTháng}$={30, 48, 50, 53, 55, 80}*;* ψ_1=30; ψ_2=80. $LD_{LươngTháng}$={Rất thấp, Thấp, Ít thấp, Cao, Rất cao}. Các giá trị ngôn ngữ có giá trị kinh điển nằm ngoài $[\psi_1, \psi_2]$: {*Rất thấp, Rất cao*}.

B1: Tính các giá trị *IC*(ω) trong *LươngTháng* tương ứng trong đoạn $[\psi_1, \psi_2]$ = [30, 80]. Lúc này: $D_{LươngTháng}$={30, 48, 50, 53, 55, 80}*;*$LD_{LươngTháng}$*={Thấp, Ít thấp, Cao}.* Xây dựng 1 ĐSGT để tính cho thuộc tính không thuần nhất *LươngTháng* như sau:

$\underline{X}_{LươngTháng}$ = ($X_{LươngTháng}$, $G_{LươngTháng}$, $H_{LươngTháng}$, ≤), 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, μ(*rất*) = 0.35, μ(*hơn*) = 0.25, μ(*khả năng*) = 0.20, μ(*ít*) = 0.20.

Lúc này ta có: *fm*(*rất thấp*) = 0.35 *x* 0.4=0.14, *fm*(*hơn thấp*) =0.25 *x* 0.4 = 0.10, *fm*(*ít thấp*) = 0.2 *x* 0.4=0.08, *fm*(*khả năng thấp*) =0.2 *x* 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]. *fm*(*rất cao*) = 0.35 *x* 0.6 = 0.21, *fm*(*hơn cao*) = 025 *x* 0.6 = 0.15, *fm*(*ít cao*) = 0.2 *x* 0.6 = 0.12, *fm*(*khả năng cao*) = 0.2 *x* 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].

DOM(LươngTháng)= {48, Thấp, 53, Rất thấp, Cao, 80, Rất cao, 30, 80, 50, Rất cao, Ít thấp, 55, 50},

Chọn $\psi_1 = 80 \in X_{LươngTháng}$ khi đó $\forall \omega \in Num(LươngTháng)$,

$IC(\omega)$ = {0.36, 0.24, 0.46, _, 0.64, 1, _, 0, 1, 0.40, _, 0.32, 0.50, 0.40}.

B2: Tính cho các giá trị ngoài khoảng bằng cách tìm các phân hoạch thích hợp của các khoảng mờ để chèn các giá trị ngoại lai vào các khoảng này.

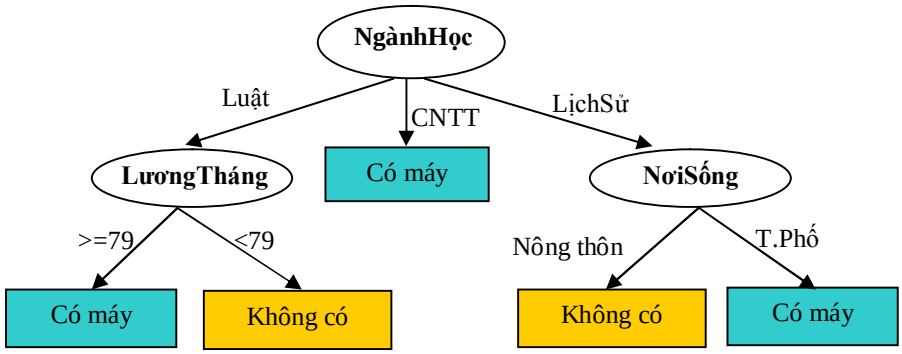
Do giá trị *Rất cao* > *Hơn cao* nên ta sẽ phân hoạch đoạn [0.79,1] tương ứng của $|I(lớn)|$.

Như vậy ta có: *fm*(Rất *cao*) ~ *fm(Hơn cao) x I*(*Hơn cao*) = 0.21 *x* 0.79 = 0.17. Nên *I*(*Hơn cao*) = [0.79, 0.96], *I*(*Rất cao*) = [0.96, 1]. Do đó $\psi_{Rất\ cao} = 97$.

Rất thấp < *Hơn thấp* nên ta sẽ phân hoạch đoạn [0, 0.14] tương ứng của $|I(thấp)|$. *fm*(Rất *thấp*) ~ *fm(Hơn thấp)x I*(*Hơn thấp*) = 0.14 *x* 0.14=0.02. Nên *I*(*Hơn thấp*) = [0.02, 0.14], *I*(*Rất thấp*) = [0, 0.02]. Do đó $\psi_{Rất\ thấp} = 4$.

B3: Tính lại $IC(\omega)$ với $[\psi_1, \psi_2] = [4, 97]$. Lúc này ta có: $IC(\omega)$ = {0.47, 0.24, 0.52, 0, 0.64, 0.81, 1, 0.27, 0.81, 0.49, 1, 0.40, 0.54, 0.49}.

Vậy thuộc tính *LươngTháng* sau khi được định lượng có giá trị là: **{48, 26, 52, 4, 64, 79, 97, 29, 79, 50, 97, 41, 54, 50}**. Cây quyết định sau khi huấn luyện như hình 3.



Hình 3. Cây quyết định được tạo sau khi định lượng thuộc tính khi chỉ biết đoạn con của $[\psi_{min}, \psi_{max}]$

VI. Kết luận

Bài báo đã đánh giá tính phức tạp của dữ liệu huấn luyện được chọn từ dữ liệu nghiệp vụ, phân tích tính đa dạng của miền trị thuộc tính đồng thời chỉ ra tính phức tạp khi định lượng giá trị ngôn ngữ. 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 trong điều kiện hạn chế, 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 Tường 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. 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)
10. Chida, A: Enhanced Encoding with Improved Fuzzy Decision Tree Testing Using CASP Templates, Computational Intelligence Magazine, IEEE (2012).
11. Chang, Robin L. P. Pavlidis: Fuzzy Decision Tree Algorithms, Man and Cybernetics, IEEE (2007).
12. Dorian, P.: Data Preparation for Data Mining, Morgan Kaufmann (1999).
13. 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).
14. 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).
15. 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).
16. Kavita Sachdeva, Madasu Hanmandlu, Amioy Kumar: Real Life Applications of Fuzzy Decision Tree, International Journal of Computer Applications (2012).
17. 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).
18. Ho Tu Bao: Introduction to knowledge discovery and data mining, Institute of Information Technology National Center for Natural Science (2000).
19. 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).
20. 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).
21. 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).

A METHOD TO DETERMINE THE LINGUISTIC VALUES IN TRAINING DATA SET TO BUILD A FUZZY DECISION TREE

Le Van Tuong Lan[1], Nguyen Mau Han[1], Nguyen Cong Hao[2]

[1] Faculty of Information Technology, College of Sciences, Hue University

[2] Center for Information Technology, Hue University

lvtlan@yahoo.com, nmhan2005@yahoo.com, nchao@hueuni.edu.vn

Abstract: Sample training data set plays an important role in the training process. When the value of the attribute domain may be value or linguistic, we need a method to homogenise sample training data set. Hedge algebra is a useful tool to make the training set homogeneous by changing the values of mixed domain to homogeneous data domain that only contains linguistics or values. In the process of homogeneous data domain, we have to know the values ψ_{min}, ψ_{max}. However, in reality, we do not know the values ψ_{min}, ψ_{max} exactly. In this paper, we present a method to determine the linguistic values when we only know the sub values $[\psi_1, \psi_2]$ without knowing the values $[\psi_{min}, \psi_{max}]$ exactly.

Keywords: Training data set, Linguistic values, Hedge algebra, Fuzzy decision tree.