International Journal of Research in Engineering and Science (IJRES)
ISSN (Online): 2320-9364, ISSN (Print): 2320-9356
www.ijres.org Volume 4 Issue 4 ǁ April. 2016 ǁ PP.16-24

A Novel Method to Build a Fuzzy Decision Tree Based On Hedge Algebras

Lan L V T[1], Han N M[1], Hao N C[2]
[1]Information Technology Faculty, Hue University of Science, Hue University, Vietnam
[2]Department of Testing, Hue University, Vietnam

Abstract: The training set plays an important role in the training process. When the value domains of the attributes are not homogenous, that is, their values may be in number or linguistic, we need a method to treat the inhomogeneous data of the training set. Hedge algebra is a useful tool to make the training set homogeneous by changing the values of mixed domain to a homogeneous data domain that only contains linguistics or number values. In the homogeneous process of the value domains, we have to know the values ψ_{min} and ψ_{max}. However, in reality, we do not know both of the values ψ_{min} and ψ_{max} exactly. In this paper, we present a method to determine the linguistic values when we only know the sub-interval $[\psi_1, \psi_2]$ without knowing the super-interval $[\psi_{min}, \psi_{max}]$ exactly.

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

I. INTRODUCTION

Let *M* is a training data set, all elements in *M* have a common set of structures, including the pair <*attribute*, *value*>, one of these attributes represents class and attributes are called predictive attributes or classification attributes. The classification is finding the rules for putting objects into one of the classes based on a training data set. There are many approaches classification problems, such as: Fisher linear discriminant function, Naïve Bayes, Logistic, neural networks, decision trees, ... In which, the decision tree method is popular due to its method of visualization, understanding and its performance [2,18]. To build a decision tree, at each node in the need to determine an appropriate attribute to check, divide the data into subsets. On the training sample *M*, basically, the classification algorithms must perform two steps:

Step 1: Selected *A* attribute with values $a_1, a_2, ..., a_n$;

Step 2: With *A* attribute is selected, we create a node of the tree and then divided the samples corresponding to this node into the corresponding set $M_1, M_2, ..., M_n$; Then again performed [17]. This is the division with the results obtained from step 1, this means that the quality of the resulting tree depends largely on choosing attributes and dividing the samples in each node. Because of this, the algorithm must calculate the amount of information received on the attributes and select the corresponding attributes have the best information to make the split node on the tree, in order to reach the tree with fewer nodes but has high predictability [2,12,18].

In the real world, business data is very diverse and complex because they are stored to serve many different jobs, many attributes were homogeneous domain before saving value, but also exist many domain attributes value not homogeneous [5,7,8,12]. When the attributes is not a homogeneous pattern appears in training sample, learning algorithms to build the tree can not proceed. Hence, the need to pre-process data to obtain homogenised training data set. The problem is that we have to deal with how to get the result is positive.

Example 1. For data table DIEUTRA stored on the laptop buying customer at a company as Table 1, we want to select the training sample to build decision trees to predict whether customers buy.

Table 1. The training set with some inhomogeneous attributes (*Salary*)

CardID	Full name	Habitat	Knowledge	Household economy	Salary	Computer
M01087	Le Van Binh	Rural	Law	Not good	Low	No
M02043	Nguyen Thi Hoa	City	IT	Not good	52	Yes
M02081	Tran Binh Tham	City	History	Not bad	20	Yes
M02046	Tran Thi Huong	City	History	Rather	High	Yes
M03087	Nguyen Thi Lai	Rural	History	Rather	High	No
M03025	Vu Tuan Hoa	Rural	IT	Rather	Very high	Yes
M03017	Le Ba Linh	City	Law	Not bad	35	No
M04036	Bach An Lai	City	Law	Rather	100	Yes
M04037	Ly Thi Hoa	City	History	Not bad	50	Yes

M04042	Vu Quang Bình	Rural	Law	Not bad	Very high	Yes
M04083	Nguyen Thi Hoa	Rural	IT	Not bad	Less low	Yes
M05041	Lê Xuan Hoan	City	IT	Not good	55	Yes
M05080	Tran Que Chung	Rural	History	Not bad	50	No

In recent years, hedge algebra is researched by many authors in the country and foreign authors and have significant results, especially in reasoning approximately and fuzzy control problem [1, 6, 11 -17, 21]. The use of hedge algebra to handing the domain of linguistic values are not homogeneous data showed very positive results [6, 8].

For example, domain of attribute *Salary* in Table 1 are homogeneous following: *{Less high, low, possibly high, less low, high, high, very high, less low, very high, possibly high, very high, less low, possibly high, possibly high* } or domain of value after quantitative linguistic values following: {45, 24, 52, 34, 64, 64, 79, 35,100, 50, 79, 40, 55, 50} with classical domain of value *Salary* in training data set is determined *Dom*(*Salary*) = [ψ_{min}, ψ_{max}] = [20,100]. The decision tree is obtained after training shown in Figure 1 [8].

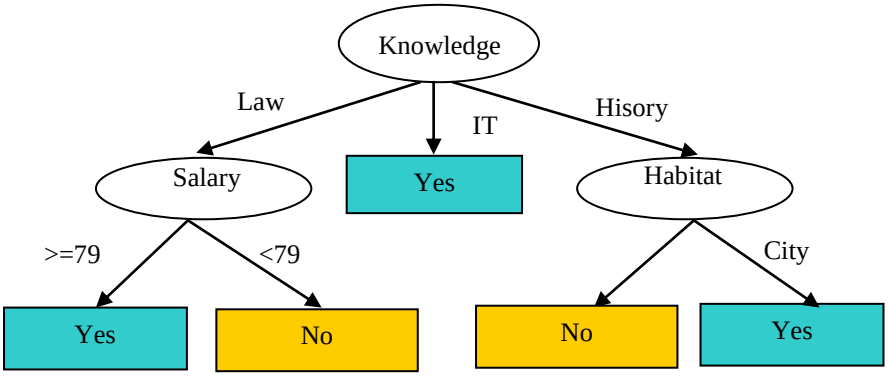

Fig 1. Decision tree is created after homogeneity of values

However, the quantitative linguistic values, do not always do well to find the value ψ_{min}, ψ_{max} in the data set. With the domain can not find the classical domain of value [ψ_{min}, ψ_{max}] in the attribute is considering of training date set, we had to ask the opinion of experts to identify them and then continue to work, as collective training data set in table 2, we ask experts to determine [ψ_{min}, ψ_{max}] = [20,100] and then continue.

Table 2. Training data set with attribute *Salary* can not find domain [ψ_{min}, ψ_{max}]

Habitat	Knowledge	Household Economy	Salary	Computer
City	Law	Not good	Less high	No
Rural	Law	Not good	Low	No
City	IT	Not good	Possibly high	Yes
City	History	Not bad	Very low	Yes
City	History	Rather	High	Yes
Rural	History	Rather	65	No
Rural	IT	Rather	Very high	Yes
City	Law	Not bad	30	No
City	Law	Rather	Very high	Yes
City	History	Not bad	Possibly high	Yes
Rural	Law	Not bad	Very high	Yes
Rural	IT	Not bad	Less low	Yes
City	IT	Not good	Possibly high	Yes
Rural	History	Not bad	Possibly high	No

The opinions of experts not always done and more can not take full advantage of the information stored in the training data set. In this paper, we will present a way to be able to quantify the value of linguistic not find classical value domain [ψ_{min}, ψ_{max}] in the attributes are considered of training data set based on hedge algebra.

II. HEDGE ALGEBRA

Hedge algebra is one approach to detecting the algebraic structure of the value domain of the linguistic variable. In view of algebra, each value domain of the linguistic variable X can be interpreted as an algebra *AX = (X, G, H, ≤)*, in which *Dom (X) = X* is the term domain of linguistic variable X is generated from a set of primary generators $G = \{c^-, c^+\}$ by the impact of the hedges $H = H^- \cup H^+$; W is a neutral element; ≤ is a

semantically ordering relation on *X*, it is induced from the natural qualitative meaning of terms. Order structure induced directly so is the difference compared to other approaches.

When we add some special elements, then hedge algebra become an abstract algebra $\underline{\underline{X}} = (X, G, H, \Sigma, \Phi, \preceq)$, which Σ, Φ are two operators taking the limit of the set terms is generated when affected by the hedges in *H*. Alternatively, if the symbol $H(x) = \{h_1 \ldots h_p x / h_1, \ldots h_p \in H\}$, then $\Phi x = infimumH(x)$ and $\Sigma x = supremumH(x)$. Thus, hedge algebra $\underline{\underline{X}}$ is built on a foundation of hedge algebra $AX = (X, G, H, \preceq)$, where $X = H(G)$, Σ and Φ are two additional operators. Then $X = X \cup Lim(G)$ with $Lim(G)$ is the set of elements limited: $\forall x \in Lim(G), \exists u \in X: x = \Phi u$ or $x = \Sigma u$. The limitation elements are added to hedge algebra $\underline{\underline{X}}$ to make the new calculation meant and so $\underline{\underline{X}} = (X, G, H, \Sigma, \Phi, \preceq)$ called complete hedge algebra [1,6].

Definition 2.1. Let X be a hedge algebra. A function $f : X \to [0,1]$ is called quantitative semantics function of X if $\forall h, k \in H^+$ or $\forall h, k \in H^-$ and $\forall x, y \in X$, we have :

$$\left|\frac{f(hx) - f(x)}{f(kx) - f(x)}\right| = \left|\frac{f(hy) - f(y)}{f(ky) - f(y)}\right|$$

For a hedge algebra X and quantitative semantic function, we can define fuzziness of vague. Let *f* is a quantitative semantics function *of* X. Consider any $x \in X$, fuzziness of *x* is measured by the diameter of the set $f(H(x)) \subseteq [0,1]$.

Definition 2.2. Let X be a hedge algebra. A function $fm: X \to [0,1]$ is said to be a fuzziness measure of term in X if satisfied the following conditions:

1) $fm(c^-) + fm(c^+) = 1$ and $\sum_{-q \le i \le p, i \ne 0} fm(h_i u) = fm(u)$ for $\forall u \in X$. In this case, *fm* is called complete;

2) For the constants **0**, **W** and **1**, $fm(0) = fm(W) = fm(1) = 0$.

3) For $\forall x, y \in X$, $\forall h \in H$, $\dfrac{fm(hx)}{fm(x)} = \dfrac{fm(hy)}{fm(y)}$

That is, this proportion does not depend on specific elements and, hence, it is called the fuzziness measure of the hedge and denoted by $\mu(h)$.

Proposition 2.1. For each fuzziness measure *fm* on **X** the following statement hold:

1) $fm(hx) = \mu(h) fm(x)$, for every $x \in X$

2) $fm(c^-) + fm(c^+) = 1$

3) $\sum_{-q \le i \le p, i \ne 0} fm(h_i c) = fm(c)$, $c \in \{c^-, c^+\}$

4) $\sum_{-q \le i \le p, i \ne 0} fm(h_i x) = fm(x)$, for every $x \in X$

5) $\sum_{i=-q}^{-1} \mu(h_i) = \alpha$ and $\sum_{i=1}^{p} \mu(h_i) = \beta$, where α, β > 0 and α + β = 1.

Definition 2.3. Let *fm* be a fuzziness measure on X. A mapping: $X \to [0,1]$ which is induced by *fm* on X, is defined as follows:

1) $\upsilon(c^-) = W - \alpha . fm(c^-)$ and $\upsilon(c^+) = W + \alpha . fm(c^+)$

2) If $1 \le j \le p$ then:

$$\upsilon(h_j x) = \upsilon(x) + Sgn(h_j x)\left[\sum_{i=1}^{j} fm(h_i x) - \omega(h_j x)\, fm(h_j x)\right]$$

If $-q \le j \le -1$ then:

$$\upsilon(h_j x) = \upsilon(x) + Sgn(h_j x)\left[\sum_{i=j}^{-1} fm(h_i x) - \omega(h_j x)\, fm(h_j x)\right]$$

where $w(h_j x) = \frac{1}{2}\left[1 + Sgn(h_j x)Sgn(h_q h_j x)(\beta - \alpha)\right] \in \{\alpha, \beta\}$

Definition 2.4. A mapping *IC*: *Dom*(A_i) $\rightarrow$ [0,1] use change the value in *R* into [0,1] defined as follows:

- If $LD_{Ai} = \emptyset$ và $D_{Ai} \neq \emptyset$ then $\forall \omega \in Dom(A_i)$ we have: $IC(w) = 1 - \frac{\psi_{\max} - \omega}{\psi_{\max} - \psi_{\min}}$, với $Dom(A_i) = [\psi_{min}, \psi_{max}]$ is a classical domain of value of A_i.

- If $D_{Ai} \neq \emptyset$, $LD_{Ai} \neq \emptyset$ then $\forall \omega \in Dom(A_i)$ we have $IC(\omega) = \{\omega * \nu(\psi_{\max LV})\}/\psi_{\max}$, với $LD_{Ai} = [\psi_{minLV}, \psi_{maxLV}]$ is a domain of linguistic value of A_i. If we select parameters *W* and fuzziness measure of the hedges so that $\nu(\psi_{\max LV}) \approx 1.0$ then $(\{\omega * \nu(\psi_{\max LV})\}/\psi_{\max}) \approx 1 - \frac{\psi_{\max} - \omega}{\psi_{\max} - \psi_{\min}}$

III. QUANTITATIVE LINGUISTIC VALUES WHEN NOT FIND CLASSICAL DOMAIN OF VALUE

For any inhomogeneities attribute *A*, we will change it to linguistic values and then to change about homogeneous values. In training data set was given in table 1, we will build a hedge algebra *Salary* as follows:

$\underline{X}_{Salary} = (X_{Salary}, G_{Salary}, H_{Salary}, \leq)$, với $G_{Salary} = \{high, low\}$,

$H^+_{Salary} = \{more, very\}$, $H^-_{Salary} = \{possibly, less\}$, where *very* > *more* and *less* > *possibly* ; $W_{Salary} = 0.6$, $fm(low) = 0.4$, $fm(high) = 0.6$, $\mu(very) = 0.35$, $\mu(more) = 0.25$, $\mu(possibly) = 0.20$, $\mu(less) = 0.20$.

Thus, we have : $fm(very\ low) = 0.35 \times 0.4 = 0.14$, $fm(more\ low) = 025 \times 0.4 = 0.10$, $fm(less\ low) = 0.2 \times 0.4 = 0.08$, $fm(possibly\ low) = 0.2 \times 0.4 = 0.08$. Because *very low* < *more low* < *low* < *possibly low* < *less low* so :

$I(very\ low) = [0,0.14]$, $I(more\ low) = [0.14,0.24]$, $I(possibly\ low) = [0.24,0.32]$, $I(less\ low) = [0.32,0.4]$.

We have: $fm(very\ high) = 0.35 \times 0.6 = 0.21$, $fm(more\ high) = 025 \times 0.6 = 0.15$, $fm(less\ high) = 0.2 \times 0.6 = 0.12$, $fm(possibly\ high) = 0.2 \times 0.6 = 0.12$. Because *less high* < *possibly high* < *high* < *more high* < *very high* so:

$I(less\ high) = [0.4,0.52]$, $I(possibly\ high) = [0.52,0.64]$, $I(more\ high) = [0.64,0.79]$, $I(very\ high) = [0.79,1]$.

Thus, given U_{Salary}= {45, low, 52, 34, *high*, *high*, *very high*, 35, 100, 50, *very*, *less low*, 55, 50}, $[\psi_{min}, \psi_{max}] = [20,100]$.

We have computed and obtain: $IC(\omega)$ = {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}. The vague of attribute *Salary* are :

{less high, low, possibly high, less low, high, high, very high, less low, very high, possibly high, very high, less low, possibly high, possibly high}

so after quantitative values of attribute *Salary* will be obtain values:

{45, 24, 52, 34, 64, 64, 79, 35,100, 50, 79, 40, 55, 50}.

However, the process of quantitative the linguistics values above only feasible if we can find a classical domain of value $[\psi_{min}, \psi_{max}]$ of attribute is considering, here is [20, 100]. In case, this domain is not found, the algorithm is not inapplicable.

3.1. Quantitative the linguistic values when knowing a sub interval of $[\psi_{min}, \psi_{max}]$ and all values of IC(ω)

Let A_i is a inhomogeneities attribute, then we have $Dom(A_i) = D_{Ai} \cup LD_{Ai}$ but marginal value $[\psi_{min}, \psi_{max}]$ respectively classical domain of value D_{Ai} of A_i do not determine, we have known a sub interval $[\psi_1, \psi_2]$ respectively linguistic value $[\psi_{LV1}, \psi_{LV2}]$ of LD_{Ai} and all vague quantitative of $IC(\omega)$.

For example attribute *Salary* in Table 2, vague of attribute *Salary* are {*less high*, *low*, *possibly high*, *less low*, *high*, *high*, *very high*, *less low*, *very high*, *possibly high*, *very high*, *less low*, *possibly high*, *possibly high* }. 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}. Here, we do not know $[\psi_{min}, \psi_{max}]$ respectively linguistic value $[\psi_{minLV}, \psi_{maxLV}]$ = [*very low*, *very high*] but knowing sub interval is $[\psi_1, \psi_2]$ = [30, 65] respectively linguistic value is $[\psi_{LV1}, \psi_{LV2}]$ = [*less low*, *more high*].

Then, because $IC(\omega) = 1 - \frac{\psi_{\max} - \omega}{\psi_{\max} - \psi_{\min}}$ so all values of ω in between $[\psi_1, \psi_2]$ will be right with this rule. Because value ω will be proportioned to the radius $f(H(x)) \subseteq [0,1]$, mean, if $\omega_1 > \omega_2$ large then $IC(\omega_1) >$

$IC(\omega_2)$ and $\frac{\omega_1}{IC(w_1)} = \frac{\omega_2}{IC(w_2)}$ when all $IC(\omega_1)$, $IC(\omega_2)$ on the same side of W. Thus, linguistic value quantitative is computed as follows:

Step 1: For ω that linguistic value respectively in $[\psi_{LV1}, \psi_{LV2}]$, we have: $\omega = IC(w)(\psi_2 - \psi_1) + \psi_1$

Step 2: For ω that linguistic value respectively in $[\psi_{LV2}, \psi_{maxLV}]$, we compute the sequential increase, according to interval $\psi_{LV2}..\psi_{maxLV}$, where $\omega_i = \psi_2 \frac{IC(w_2)}{IC(w_i)}$ and shifted position ψ_{LV2} into the position I have just found.

Step 3: For ω that linguistic value respectively in $[\psi_{minLV}, \psi_{LV1}]$, we compute sequential decrease according interval $\psi_{LV1}..\psi_{minLV}$, where $\omega_i = \psi_1 \frac{IC(w_1)}{IC(w_i)}$ and shifted positions back into position ψ_{LV1} i have just found.

Example 2. Let $\underline{X}_{Salary}$ be a hedge algebra to describe inhomogeneous attributes *Salary* in table 2 as follows: $\underline{X}_{Salary} = (X_{Salary}, G_{Salary}, H_{Salary}, \leq)$, where $G_{Salary} = \{high, low\}$, $H^+_{Salary} = \{more, very\}$, $H^-_{Salary} = \{possibly, less\}$. $W_{Salary} = 0.6$, $fm(low) = 0.4$, $fm(high) = 0.6$, $\mu(very) = 0.35$, $\mu(more) = 0.25$, $\mu(possibly) = 0.20$, $\mu(less) = 0.20$. Domain of linguistic values is *{less high, low, possibly high, less low, high, high, very high, less low, very high, possibly high, very high, less low, possibly low, possibly high}*. $IC(\omega)$ = {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}. Knowing sub interval have domain of value is $[\psi_1, \psi_2]$ = [30, 65] corresponding with domain of linguistic value is $[\psi_{LV1}, \psi_{LV2}]$ = [*less low, more high*]. We have:

fm(*very low*) = 0.35 × 0.4 = 0.14, *fm*(*more low*) = 025 × 0.4 = 0.10, *fm*(*less low*) = 0.2 × 0.4 = 0.08, *fm*(*possibly low*) = 0.2 × 0.4 = 0.08. Because *very low* < *more low* < *low* < *possibly low* < *less low* so :

I(*very low*) = [0, 0.14], *I*(*more low*) = [0.14, 0.24], *I*(*possiby low*) = [0.24, 0.32], *I*(*less low*) = [0.32, 0.4].

fm(*very high*) = 0.35 × 0.6 = 0.21, *fm*(*more high*) = 0.25 × 0.6 = 0.15, *fm*(*less high*) = 0.2 × 0.6 = 0.12, *fm*(*possibly high*) = 0.2 × 0.6 = 0.12. Because *less high* < *possibly high* < *high* < *more high* < *very high* so: *I*(*less high*) = [0.4, 0.52],*I*(*possibly high*) = [0.52, 0.64], *I*(*more high*) = [0.64, 0.79], *I*(*very high*) = [0.79, 1].

Setp 1: Compute ω that have linguistic value in [*less low, more high*]

$\omega_{less\ low} = IC(\omega_{less\ low})(\psi_2 - \psi_1) + \psi_1 = 0.4(65-30)+30=44$

$\omega_{less\ high} = IC(\omega_{ess\ high})(\psi_2 - \psi_1) + \psi_1 = 0.52(65-30)+30=48$

$\omega_{possiby\ high} = IC(\omega_{possiby\ high})(\psi_2 - \psi_1) + \psi_1 = 0.64(65-30)+30=52$

Step 2: Compute ω that have linguistic value in [*more high, very high*]

$\omega_{more\ high} = \psi_2 \times IC(\omega_{possibly\ high})/IC(\omega_{more\ high}) = 65 \times 0.64 / 0.52 = 80$

$\omega_{very\ high} = \psi_2 \times IC(\omega_{more\ high})/IC(\omega_{very\ high}) = 80 \times 0.79 / 0.64 = 99$

Setp 3: Compute ω that have linguistic value in [*very low, less low*]

$\omega_{possibly\ low} = \psi_1 \times IC(\omega_{less\ low})/IC(\omega_{possibly\ low}) = 30 \times 0.32 / 0.4 = 24$

$\omega_{more\ low} = \psi_1 \times IC(\omega_{possibly\ low})/IC(\omega_{more\ low}) = 24 \times 0.24 / 0.32 = 18$

$\omega_{very\ low} = \psi_1 \times IC(\omega_{more\ low})/IC(\omega_{very\ low}) = 18 \times 0.14 / 0.24 = 10$

Thus, the domain of value after values quantitative are : {48, 18, 52, 30, 80, 80, 99, 30, 99, 52, 99, 30, 52, 52}. Decision tree after training shown in figure 2.

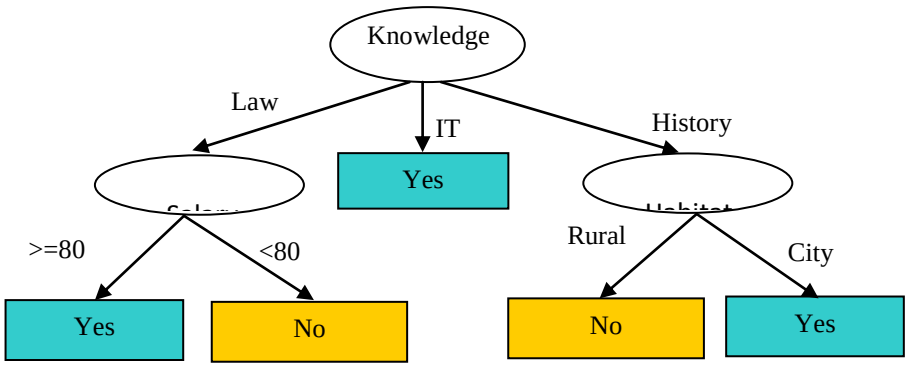

Fig 2. Decision tree is created after value of attribute quantitative

3.2. Quantitative the linguistics values when knowing a sub interval of [ψ_{min}, ψ_{max}] and do not determine all values of IC(ω)

Let A_i is a inhomogeneities attribute, then we have $Dom(A_i) = D_{Ai} \cup LD_{Ai}$ but marginal value [ψ_{min}, ψ_{max}] respectively classical domain of value D_{Ai} of A_i do not determine, we have known a sub interval [ψ_1, ψ_2] respectively linguistic value [ψ_{LV1}, ψ_{LV2}] of LD_{Ai}, mean $\upsilon(\psi_{LV1}) = IC(\psi_1)$ and $\upsilon(\psi_{LV2}) = IC(\psi_2)$. And then, we have to find the remaining values of $IC(\omega_i)$, mean $IC(\omega_i)$ satisfy $IC(\psi_i) < IC(\psi_1)$ or $IC(\psi_i) > IC(\psi_2)$.

Because $IC(\omega) = 1 - \dfrac{\psi_{\max} - \omega}{\psi_{\max} - \psi_{\min}}$ so all values of ω in between [ψ_1, ψ_2] will be right with this rule,

mean $IC(\omega) = 1 - \dfrac{\psi_2 - \omega}{\psi_2 - \psi_1}$ where $\omega \in [\psi_2 - \psi_1]$. Therefore, we can build a hedge algebra respectively.

According method to build a hedge algebra in section 2, we have fuzziness of values in hedge algebra are a sub interval of [0,1]. So, a set of this sub interval of the value of the same level long will create partitions of [0.1]. Partition with values greater lengths from going smoother and infinitely greater length when the length of the interval in the partition decreases to 0.

Therefore, the linguistic value is a linear ordering so we will divide the corresponding sub interval into smaller partitions to determine the length of the interval [0, $\upsilon(\psi_i)$] or [$\upsilon(\psi_i)$, 1], so that we can determine the values for the linguistic values. This is where to calculate the IC(ω) not in interval [ψ_1, ψ_2] by dividing the interval in succession to determine the IC (iω) respectively. Thus, we have stepped to compute as follows:

Step 1: Building a hedge algebra in [ψ_1, ψ_2] to compute values of $IC(\omega)$ respectively of values in [ψ_1, ψ_2] này.

Step 2: Compute partitions of $IC(\omega)$ is repeated as follow:

1) If $\psi_i < \psi_1$ then:

- Partitioned [0, $\upsilon(\psi_1)$] into [0, $\upsilon(\psi_i)$] and [$\upsilon(\psi_i)$, $\upsilon(\psi_1)$]
- Compute $fm(h_i) \sim fm(h_1) \times I(\psi_1)$ and $fm(h_1) = fm(h_1) - fm(h_i)$

2) If $\psi_i > \psi_2$ then:

- Partitioned [$\upsilon(\psi_2)$, 1] into [$\upsilon(\psi_2)$, $\upsilon(\psi_i)$] và [$\upsilon(\psi_i)$, 1]
- Compute $fm(h_i) \sim fm(h_2) \times I(\psi_2)$ và $fm(h_2) = fm(h_2) - fm(h_i)$

3) Compute values of $IC(\omega_i)$ and ψ_i at position *i*. Assign position i are the position 1 and continue to compute backward with the remaining value with $\psi_i < \psi_1$ or assigned position i are the position 2 and continue charging forward with the remaining value $\psi_i > \psi_2$.

Step 3: Quantitative the linguistic values with the calculation section 1, where known, all values of $IC(\omega)$.

Example 3. Let a training data set as Table 3. We will quantify the linguistic values of attribute Salary.

Table 3. A training data set with inhomogeneous attribute *Salary* and can not find domain [ψ_{min}, ψ_{max}]

Habit	Knowledge	Household Economy	Salary	Computer
City	Law	Not good	48	No
Rural	Law	Not good	Low	No
City	IT	Not good	53	Yes
City	History	Not bad	Very low	Yes
City	History	Rather	High	Yes
Rural	History	Rather	80	No
Rural	IT	Rather	Very high	Yes
City	Law	Not bad	30	No
City	Law	Rather	80	Yes
City	History	Not bad	50	Yes
Rural	Law	Not bad	Very high	Yes
Rural	IT	Not bad	Less low	Yes
City	IT	Not good	55	Yes
Rural	History	Not bad	50	No

Because a training data set with inhomogeneous attribute *Salary* so we have to homogeneous values of attribute *Salary*. We have*: Dom*(*Salary*) = $D_{Salarry} \cup LD_{Salarry}$. $D_{Salarry}$ = {30, 48, 50, 53, 55, 80}*;* $\psi_1 = 30$; $\psi_2 = 80$. $LD_{Salarry}$ = {*very low, low, less low, high, very high*}. The linguistic values with classical value outside [ψ_1, ψ_2]: {*less low, very high*}.

Step 1: Compute values of $IC(\omega)$ of *Salary* respectively in [ψ_1, ψ_2] = [30, 80]. We have: $D_{Salarry}$={30, 48, 50, 53, 55, 80}*;*$LD_{Salarry}$*={low, less low, high}.* Let $\underline{X}_{Salarry}$ a hedge algebra and is denoted as follows:

$\underline{X}_{Salarry}$ = ($X_{Salarry}$, $G_{Salarry}$, $H_{Salarry}$, ≤), where $G_{Salarry}$ = {*high, low*}, $H^{+}_{Salarry}$ = {*More, Very*}, $H^{-}_{Salarry}$ = {*possibly, less*}. $W_{Salarry}$ = 0.4, *fm*(*low*) = 0.4, *fm*(*high*) = 0.6, μ(*very*) = 0.35, μ(*more*) = 0.25, μ(*possibly*) = 0.20, μ(*less*) = 0.20. We have: *fm*(*very low*) = 0.35 × 0.4 = 0.14, *fm*(*more low*) = 0.25 × 0.4 = 0.1, *fm*(*possibly low*) = 0.2 × 0.4 = 0.08, *fm*(*less low*) = 0.2 × 0.4 = 0.08,

Becase *very low* < *more low* < *low* < *possibly low* < *less low* so: *I*(*very low*) = [0, 0.14], *I*(*more low*) = [0.14, 0.24], *I*(*possibly low*) = [0.24, 0.32], *I*(*less low*) = [0.32, 0.4].

fm(*very high*) = 0.35 × 0.6 = 0.21, *fm*(*more high*) = 025 × 0.6 = 0.15, *fm*(*possibly high*) = 0.2 × 0.6 = 0.12, *fm*(*less high*) = 0.2 × 0.6 = 0.12,

Because *less high* < *possibly high* < *high* < *more high* < *very high*, so: *I*(*less high*) = [0.4, 0.52], *I*(*possibly high*) = [0.52, 0.64], *I*(*more high*) = [0.64, 0.79], *I*(*very high*) = [0.79, 1].

Dom(Salary) = {48, *low*, 53, *very low*, *high*, 80, *very high*, 30, 80, 50, *very high*, *less low*, 55, 50}.

Let ψ_1 = 80 ∈ X_{Salary}, thus for every ω ∈ *Num*(*Salary*), *IC*(ω) = {0.36, 0.24, 0.46, _, 0.64, 1, _, 0, 1, 0.40, _, 0.32, 0.50, 0.40}.

Step 2: The values outside the interval is computed by finding the appropriate partitions of the fuzzy interval to insert outliers in this interval..

Because *very high* >*more high*, so we will partition [0.79,1] respectively with |*I(large)*|. Thus, we have:

fm(*very high*) ~ *fm(more high)* × *I*(*more high*) = 0.21 × 0.79 = 0.17. So *I*(*more high*) = [0.79, 0.96], *I*(*very high*) = [0.96, 1]. Thus, $\psi_{\text{very high}}$ = 97.

Because *very low* < *more low* so we will partition [0, 0.14] respestively with |*I(low)*|. *fm*(*very low*) ~ *fm(more low)* × *I*(*more low*) = 0.14 × 0.14 = 0.02. So *I*(*more low*) = [0.02, 0.14], *I*(*very low*) = [0, 0.02]. Thus, $\psi_{very\ low}$ = 4.

Step 3: Compute *IC*(ω) with [ψ_1, ψ_2] = [4, 97] is repeated. Thus, we have: *IC*(ω) = {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}.

Thus, attribute Salary is quantitative have values are: {48, 26, 52, 4, 64, 79, 97, 29, 79, 50, 97, 41, 54, 50}. Decision tree after training shown by Figure 3.

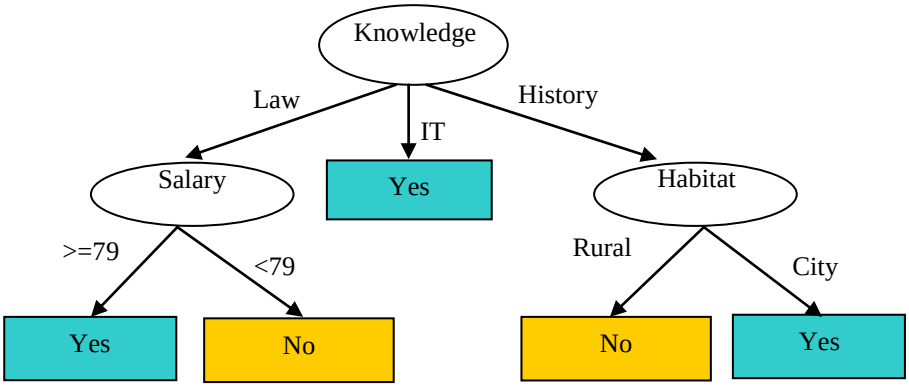

Fig 3. Decision tree is created after attribute quantitative

3.3. Evaluation

The decision tree which is created after dealing with the values outside the interval [ψ_{min}, ψ_{max}] as shown in Figure 3 is feasible because it shows the similarity with the tree in the Figure 1 created by the dataset completely determined in the interval [ψ_{min}, ψ_{max}] and in accordance with the expert opinion.

In case we cannot deal with the values outside the interval [ψ_{min}, ψ_{max}] as the proposal, it means that we bypass the values outside the interval [ψ_{min}, ψ_{max}] in the training set and it is considered as the "error" cases. Now, we have *IC*(ω) = {0.36, 0.24, 0.46, 0.64, 1, 0, 1, 0.40, 0.32, 0.50, 0.40}. The decision tree which is created after the learning process is shown in the Figure 4.

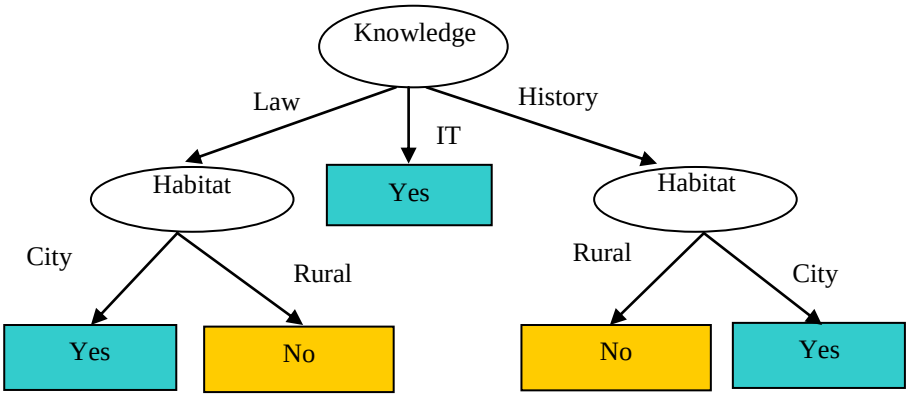

Fig 4. The decision tree created without dealing with the values outside the interval [ψ_{min}, ψ_{max}]

Compare to the tree created by the training set completely determined in the interval [ψ_{min}, ψ_{max}] or with the best expert opinion as shown in the Figure 4, we recognized that the tree shown in the Figure 4 does not reflect the reality. This fact is completely suitable because dealing with by this way causes the training set to be lost seriously. Thus, the quantifying method for the linguistic value when we do not know the interval [ψ_{min}, ψ_{max}] and only know the sub-interval con [ψ_1, ψ_2] as proposed is feasible.

IV. CONCLUSION

This paper evaluated the complexity of the training data set are selected from business data, analyze the variety of domains of attribute values and complexity indicates the quantitative the linguistic values. On the basis of hedge algebra, by considering efficiency as homogeneous as values of attributes not homogeneous in the sample according to the linguistic value or classical value, the paper points out a way knowledge to be able to identify clear value for the linguistic value in limited conditions, so that we can train the decision tree consistent with reality.

REFERENCES

[1]. D. T. Long, "A method to build rule fuzzy systems semantic-based hedge algebra and application to classification", Mathematic doctor thesis, IOIT, 2010.
[2]. D. V. Ban, L. M. Thanh, L. V. T. Lan, "A method for choosing a training set and a learning algorithm to build a decision tree in data mining", Journal of Computer Science and Cybernetics, vol. 23, no. 4, pp. 317-326, 2007.
[3]. N. C. Ho, "Fuzzy set theory and soft computing", Collection of lectures in Autumn school about fuzzy systems and applications, Mathematical Institute of Vietnam, pp. 51-92, 2006.
[4]. N. C. Ho, "A topological completion of refined hedge algebras and a Model of fuzziness of linguistic terms", Fuzzy Set and Systems, vol. 158, no. 4, pp. 436-451, 2007.
[5]. N. C. Hao, N. C. Ho, "An approach for approximate data in fuzzy databases", Journal of Computer Science and Cybernetics, vol. 23, no. 2, pp. 110-121, 2007.
[6]. N. C. Hao, "Fuzzy databases with data manipulating based on hedge algebra", Mathematic doctor thesis, IOIT, 2008.
[7]. L. V. T. Lan, "Dependency data of training set and its effect on the classification in data mining", Hue University Journal of Research, vol. 19, no. 53, pp. 55-64, 2009.
[8]. L. V. T. Lan, N. M. Han, N. C. Hao, "An approach for choosing a training set to build a decision tree based on hedge algebra", Proceeding of the 6th National Conference on Fundamental and Applied Information Technology Research (FAIR), Publisher of Natural Sciences and Technology, 2013, pp. 251-258.
[9]. A. K. Bikas, E. M. Voumvoulakis and N. D. Hatziargyriou, "Neuro-Fuzzy Decision Trees for Dynamic Security Control of Power Systems", Intelligent System Applications to Power Systems (ISAP), 15th International Conference, Curitiba, 2009, pp. 1-6.
[10]. A. Chida, "Enhanced Encoding with Improved Fuzzy Decision Tree Testing Using CASP Templates", Computational Intelligence Magazine, IEEE, vol. 7, issue 4, pp 55-60, 2012.
[11]. Chang, Robin L. P. Pavlidis, "Fuzzy Decision Tree Algorithms", Man and Cybernetics, IEEE, vol. 7, issue 1, pp 28-35, 2007.
[12]. P. Dorian, "Data Preparation for Data Mining", Morgan Kaufmann, 1999.

[13]. R. A. Daveedu, S. Jaya, D. Lavanya, "Construction of Fuzzy Decision Tree using Expectation Maximization Algorithm", International Journal of Computer Science and Management Research, vol. 1, issue 3, pp 416-424, 2012.
[14]. A. Fernandez, M. Calderon, E. Barrenechea, "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, vol. 9, pp. 39-46.
[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, vol. 3, pp. 1457 – 1462.
[16]. K. Sachdeva, M. Hanmandlu, A. Kumar, "Real Life Applications of Fuzzy Decision Tree", International Journal of Computer Applications, vol. 42, no. 10, pp 24-28, 2012.
[17]. A. H. Hefny, S. A. Ghiduk, A. A. 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., vol. 1, pp 55-63, 2010.
[18]. H. T. Bao, "Introduction to knowledge discovery and data mining", Institute of Information Technology, National Center for Natural Science for Technology, 2000. [Online] Available: http://www.jaist.ac.jp/~bao
[19]. N. C. Ho, H. V. Nam, "An algebraic approach to linguistic hedges in Zadeh's fuzzy logic", Fuzzy Sets and Systems, vol.129, issue 2, pp. 229-254, 2002.
[20]. S. Moustakidis, G. Mallinis, N. Koutsias, J. B. Theocharis, V. Petridis, "SVM-Based Fuzzy Decision Trees for Classification of High Spatial Resolution Remote Sensing Images", Geoscience and Remote Sensing, IEEE, vol.129, issue 2, pp 149-169, 2012.
[21]. O. Dorokhov, V. Chernov, "Application of the fuzzy decision trees for the tasks of alternative choices", Transport and Telecommunication Institute, Lomonosova 1, Riga, LV-1019, Latvia, vol. 12, no. 2, pp 4-11, 2011.