Annales Univ. Sci. Budapest., Sect. Comp. 45 (2016) 319–335

AN ALGORITHM TO BUILDING A FUZZY DECISION TREE FOR DATA CLASSIFICATION PROBLEM BASED ON THE FUZZINESS INTERVALS MATCHING

Le Van Tuong Lan, Nguyen Mau Han and Nguyen Cong Hao
(Hue University, VietNam)

Communicated by Le Manh Thanh

(Received October 12, 2016; accepted November 21, 2016)

Abstract. Nowaday, on demand to reflect the real world, so we have many imprecise stored business data warehouses. The precise data classification cannot solve all the requirements. Thus, fuzzy decision tree classification problem have role is important of fuzzy data mining problem. The fuzzy decision classification based on fuzzy set theory have some limitations derived from the inner selves of it. The hedge algebra with many advantages has become a really useful tool for solving the fuzzy decision tree classification. However, sample data homogenise process based on quantitative methods of the hedge algebra with some restrictions remain appear because of error in the process and not the result tree truly versatile. So, the fuzzy decision tree obtained not always have high predictable. In this paper, we using fuzziness intervals matching an approach hedge algebra, we proposed the inductive learning method HAC4.5 fuzzy decision tree to obtain the fuzzy decision tree with high predictable.

1. Introduction

The real world is infinite while our language are limited, inevitably appears the phrase are inexact or imprecise. Therefore, in practice, the business data

Key words and phrases: Hedge algebra, linguistic, homogenise, fuzzy decision tree, HAC4.5.
2010 Mathematics Subject Classification: 11A07, 11A25, 11N25, 11N64.

warehouse stored imprecise is inevitable, so the precise data classification can not solve all the requirements. The fuzzy classification problem has been studied by many scientists with different approaches [1], [5–7], [15–16], [19–25]. including fuzzy decision tree classification is very interesting, because of the intuitive and effective training model.

1. Building fuzzy decision tree based on fuzzy set theory, such as: Zadeh LA., Chang, Fullér R., Hesham A, Ishibuchi H., Lee C.S. George, Wang T., Lee H, Wei-Yuan Cheng, Chia-Feng Juang, etc [8–10], [15], [18], [26–32]. Scientists follow this approach has given many solutions approach based on fuzzy set theory combined with neural networks, genetic, support vector machines to solve limitations problems of precise classification. However, there are still encountering the limitations stem from fuzzy set theory:

- It is difficult to simulate complete structure of linguistic that people use to reason. Order structure induced on the concepts represented by linguistic values are not shown on the fuzzy.

- In reasoning process, sometimes we need to approximate the linguistic value that is to find a linguistic value that value its meaning with a fuzzy set approximation given, which caused a complex and error for approximation process and depends on the subjective.

2. Zengchang Q., Jonathan Lawry, Yongchuan Tang, have determined the linguistic values for fuzzy data set and building linguistic decision tree (LDT) using the thought of the ID3 algorithm of precise decision tree with the node corresponding the linguistic attribute (LID3) [12], [13], [31], [32]. However:

- This approach will give the multilevel tree, there is a large horizontal division at the linguistic node when set the large liguistic values of the fuzzy attribute (Figure 1), so easily become over-curious. In addition, at this node, we cannot use a binary division of the C4.5 algorithm, because not order between linguistic values.

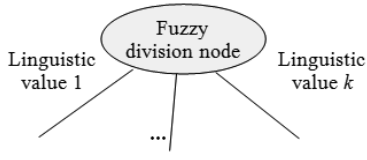

Figure 1. Multilevel position according to linguistic value at fuzzy attribute

- Moreover, with the precise values in the domain of the fuzzy attribute training data set, a sub interval of the precise values will be mapped become a linguistic value should be more errors.

For example: with Mushroom training data (Figure 2), the classification of Mushroom for the *Habitat* and *Population* attributes having more errors by the training data contains precise data and imprecise data.

	GillSize	GillColor	StalkShape	StalkRoot	StalkSurface	StalkSurface	StalkColor	StalkColor	VeilType	Habitat	VeilColor	RingNumber	RingType	SporePrint	Population	Classes
2	n	k	e	e	s	s	w	w	p	78	w	o	p	k	25	poisonous
3	b	k	e	c	s	s	w	w	p	55	w	o	p	n	25	edible
4	b	n	e	c	s	s	w	w	p	34	w	o	p	n	More Dense	edible
5	n	n	e	e	s	s	w	w	p	78	w	o	p	k	50	poisonous
6	b	k	t	e	s	s	w	w	p	78	w	o	e	n	25	edible
7	b	n	e	c	s	s	w	w	p	55	w	o	p	k	25	edible
8	b	g	e	c	s	s	w	w	p	Possibly Dry	w	o	p	k	15	edible
9	b	n	e	c	s	s	w	w	p	78	w	o	p	n	10	edible
10	n	p	e	e	s	s	w	w	p	Possibly Dry	w	o	p	k	15	poisonous
11	b	g	e	c	s	s	w	w	p	78	w	o	p	k	Rare	edible
12	b	g	e	c	s	s	w	w	p	55	w	o	p	n	15	edible
13	b	n	e	c	s	s	w	w	p	78	w	o	p	k	15	edible
14	b	w	e	c	s	s	w	w	p	34	w	o	p	n	Less Rare	edible
15	n	k	e	e	s	s	w	w	p	78	w	o	p	n	50	poisonous
16	b	n	t	e	s	f	w	w	p	Less Dry	w	o	e	k	1	edible
17	n	k	e	e	s	s	w	w	p	78	w	o	p	n	50	edible
18	b	k	t	e	s	s	w	w	p	Very Wet	w	o	e	n	15	edible
19	n	n	e	e	s	s	w	w	p	More Dry	w	o	p	k	15	poisonous
20	n	n	e	e	s	s	w	w	p	34	w	o	p	n	15	poisonous
21	n	k	e	e	s	s	w	w	p	55	w	o	p	n	15	poisonous
22	b	k	e	c	s	s	w	w	p	78	w	o	p	n	10	edible
23	n	n	e	e	s	s	w	w	p	Very Dry	w	o	p	n	15	poisonous
24	b	k	e	c	s	s	w	w	p	55	w	o	p	n	25	edible
25	b	w	e	c	s	s	w	w	p	Wet	w	o	p	n	15	edible
26	b	g	e	c	s	s	w	w	p	78	w	o	p	k	10	edible
27	n	n	e	e	s	s	w	w	p	55	w	o	p	n	25	poisonous

Figure 2. Picture of Mushroom data

3. An approach based on hedge algebra is proposed since 1990 by N.C. Ho and W. Wechler has several advantages. Because, according to this approach, each linguistic value of linguistic variable is an element of hedge algebras structure so we can matching it.

According to the approach algebras, we can homogeneous fields that data includes precise data and imprecise data. N.C. Ho, N.C. Hao, L.X. Viet, T.T. Son, Long N.V, Nam H.V, [2–4], [11–14] [17] showed semantic quantitative methods could data homogeneous into number value or linguistic value and how to query the data on this field. Therefore, we can to learn classification on homogeneous sample set.

Build decision tree problem can use the algorithms to build decision tree such as C4.5, SLIQ, ... to learn [20–21], [23], [25] with the binary division node is calculated by division point based on linguistic values order and determined the value corresponding in hedge algebra is built.

However, homogeneous method based on hedge algebra semantic quantitative method have some errors, because fuzziness measure of imprecise value is sub interval of [0,1], thus values that approximate can be partitioned in two different sub intervals so different data classification results. Beside, result tree

is also difficult prediction in these cases to predict where there is an overlap fuzzy division point. For example, we need to predict this case $[x_1, x_2]$, where $x_1 < x$ and $x_2 > x$ at the fuzzy division node in Figure 3.

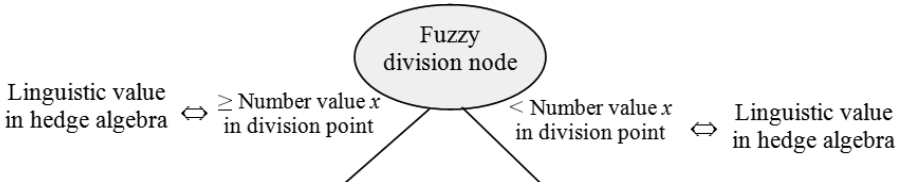

Figure 3. Binary division point by linguistic value or number value of fuzzy attribute based on hedge algebra semantic quantitative method

In this paper, we will propose a fuzzy decision tree learning method with the heterogeneous training sample set based on fuzziness interval matching method with purpose to minimize errors in the process to predict.

The paper is organized as follows. In the second section, fuzzy interval matching method will be recalled. In Section 3 improvement from the HAC4.5 algorithm for fuzzy data classification will be propose. Section 4 will be devoted to experimental and discuss. Some conclusions will be given in Section 5.

2. Building a fuzziness intervals matching method based on hedge algebra

Hedge algebra is one approach to detecting 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 $\underline{X} = (X, G, H, \leq)$, in which $Dom(X) = X$ is the terms 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; $\leq$ is an 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 $\underline{X} = (X, G, H, \Sigma, \Phi, \leq)$, 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 $H(x) = \{h_1 ... _p x | h_1, ..., h_p \in H\}$, then $\Phi_x = infimum\, H(x)$ and $\Sigma_x = supremum\, H(x)$. Thus, hedge algebra X is built on foundation of hedge algebra $\underline{X} = (X, G, H, \leq)$, 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 $\underline{X}$ to make the new calculation meant and so $\underline{X} = (X, G, H, \Sigma, \Phi, \leq)$ called complete hedge algebra. The quantitative semantics function (ν), fuzziness measure function (fm), sign function (SGN) and the properties of hedge algebra can reference in the relevant documents [2–3].

2.1. Definition of fuzziness intervals

Definition 1. ([2]) A fuzziness interval of $x \in X$ is denoted by $I(x)$ is a sub interval of $[0, 1]$ and have length is determined by fuzziness measure of x, i.e. $fm(x) = |I(x)|$.

For every term x, the fuzziness interval of $x \in X$ is a subinterval of $[0, 1]$ of length $fm(x)$, denoted by $I_{fm}(x)$, which will be constructed by induction on the length of x as follows:

i) For x of length 1, i.e. $x \in \{c^+, c^-\}$, $I_{fm}(c^+)$ and $I_{fm}(c^+)$ are intervals which constitute a partition of $[0, 1]$ and satisfy the conditions that $c^- \leq c^+$ implies $I_{fm}(c^-) \leq I_{fm}(c^+)$, $|I_{fm}(c^+)| = fm(c^+)$ and $|I_{fm}(c^+)| = fm(c^+)$, where $|I(x)|$ denotes the length of $I(x)$ and the notation $U \leq V$ means that, for $\forall x \in U, \forall y \in V$, we have $x \leq y$.

ii) Suppose that $I_{fm}(x)$ has been defined and $|I_{fm}(x)| = fm(x)$, for all x of length k. Then, the fuzziness intervals $\{I_{fm}(h_i x) : i \in q\hat{}p\}$ are constructed so that they constitute a partition of $I_{fm}(x)$ and satisfy the conditions that $|I_{fm}(h_i x)| = fm(h_i x)$ and $\{I_{fm}(h_i x) : i \in [-q\hat{}p]\}$ is a linearly ordered set, whose order is induced by that of the set $\{h_{-q}x, h_{-q+1}x, ..., h_p x\}$.

When $l(x) = k$, we denoted $I(x)$ instead of $I_{fm}(x)$, $X_k = \{\forall x \in X : l(x) = k\}$ is the set of elements in X that have length equal k, $I_k = \{I_k(x) : x \in X_k\}$ is the set of fuzziness interval level k.

Definition 2. Two the fuzziness intervals are called equal, denoted $I(x) = I(y)$, if they are determined by the same value ($x = y$), i.e. we have $I_L(x) = I_L(y)$ and $I_R(x) = I_R(y)$. Where $I_L(x)$ and $I_R(x)$ are point the tip of the left and right of fuzziness interval $I(x)$. Otherwise, we denoted by $I(x) \neq I(y)$.

Theorem 1. ([2]) *Let $\underline{X} = (X, G, H, \leq)$ be a hedge algebra, we have:*

i) *If* $\text{sign}(h_p x) = +1$, *then*

$$I(h_{-q}x) \leq I(h_{-q+1}x) \leq \cdots \leq I(h_{-1}x) \leq I(h_1 x) \leq I(h_2 x) \leq \cdots \leq I(h_p x)$$

and if $\text{sign}(h_p x) = -1$, *then*

$$I(h_{-q}x) \geq I(h_{-q+1}x) \geq \cdots \geq I(h_{-1}x) \geq I(h_1 x) \geq I(h_2 x) \geq \cdots \geq I(h_p x).$$

ii) *The set* $I_k = \{I_k(x) : x \in X_k\}$ *is a partition of* $[0,1]$.

Proposition 1. $\forall x, y \in X$, *we determine two the fuzziness intervals* $I_k(x)$ *and* $I_l(y)$. *And then they or without inheriting relation, or relation with each other if* $\exists I_v(z) \in I_v, v \leq \min(l,k), I_L(z) \leq I_L(y), I_R(z) \geq I_R(y)$, $I_L(z) \leq I_L(x), I_R(z) \geq I_R(x)$, *i.e.* $Iv(z) \supseteq I_k(x)$ *and* $I_v(z) \supseteq I_l(y)$, *i.e.* x, y *is generated by* z, $x = h_{in}...h_{i1}z, y = k_{jm}...k_{j1}z, \forall h_i, k_j \in H$.

2.2. The fuzziness intervals matching

Let $\underline{X} = (X, G, H, \leq)$ be a hedge algebra and a interval value $[a,b]$. For comparison a value $x \in X$ with $[a,b]$, the first, we can to change $[a,b]$ into sub interval of $[0,1]$. Because, the fuzziness of x if a sub interval of $[0,1]$, thus, for comparison a value $x \in X$ and sub interval of $[0,1]$, we only consider intersection of two sub intervals of $[0,1]$ corresponding.

From [2], fore each $x \in X, I(x) \subseteq [0,1]$ and $|I(x)| = fm(x)$, $[I_a, I_b] = [f(a), f(b)] \subseteq [0,1]$ the same to change $[a,b]$ into sub interval of $[0,1]$.

i) For each $[I_a, I_b]$ if exist $x \in X$ so that $[I_a, I_b] \subseteq I(x)$ then $[a,b] = {}_{|x|}x$, (see Figure 4).

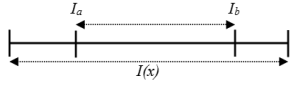

Figure 4. The relationship in case $[I_a, I_b] \subseteq I(x)$

ii) For each $[I_a, I_b]$ so that $[I_a, I_b] \not\subset I(x) \forall x, x_1 \in X$ then: for each x and x_1, supposed that $x < x_1$ if $|[I_a, I_b] \cap I(x)| \geq |[I_a, I_b]|/\pounds$ then $[a,b] = {}_{|x|}x$, where $\pounds$ is number of interval $I(x_i) \subseteq [0,1]$ so that $[I_a, I_b] \cap I(x_i) \neq \emptyset$, (see Figure 5).

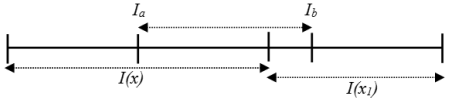

Figure 5. The relationship in case $[I_a, I_b] \subseteq I(x)$

Otherwise, if $|[I_a, I_b] \cap I(x_1)| \geq |[I_a, I_b]|/\pounds$ then $[a,b] = {}_{|x_1|}x_1$, (see Figure 6).

iii) For each $[I_a, I_b]$ and $x \in X$ so that $[I_a, I_b] \cap I(x) = \emptyset$ then exist $z \in X$ so that $[I_a, I_b] \subseteq I(z)$ and $I(x) \subseteq I(z)$ then $[a,b] = {}_{|z|}x$, (see Figure 7).

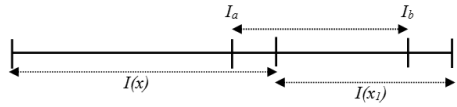

Figure 6. The relationship in case $[I_a, I_b] \not\subset I(x)$

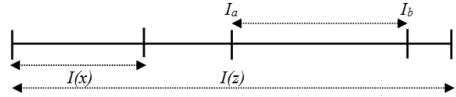

Figure 7. The relationship in case $[I_a, I_b] \cap I(x) = \emptyset$

Definition 3. Let $[a_1, b_1]$ and $[a_2, b_2]$ are two difference intervals corresponding two the fuzziness intervals $[I_{a1}, I_{b1}]$, $[I_{a2}, I_{b2}] \subseteq [0, 1]$. We said that interval $[a_1, b_1]$ precede order $[a_2, b_2]$ or $[a_2, b_2]$ behind order $[a_1, b_1]$, is written $[a_1, b_1] < < [a_2, b_2]$ *or* $[I_{a1}, I_{b1}] < [I_{a2}, I_{b2}]$ if:

i) $b_2 > b_1$ (i.e. $I_{b2} > I_{b1}$);

ii) if $I_{b2} = I_{b1}$ (i.e. $b_2 = b_1$) then $I_{a2} > I_{a1}$ (i.e. $a_2 > a_1$).

Then, we said that the sequence of interval $[a_1, b_1]$, $[a_2, b_2]$ is the sequence have two post-preorder relations.

Theorem 2. *Let* $[a_1, b_1]$, $[a_2, b_2], \ldots, [a_k, b_k]$ *are* k *difference intervals each other. Then, we always obtained a sequence with* k *interval from above intervals by post-preorder relations.*

Proof. Clearly, for each k difference intervals each other, such as: $[a_1, b_1]$, $[a_2, b_2], \ldots,$ $[a_k, b_k]$, we always find interval $[a_i, b_i]$ of sequence, where $a_i = = \min(a_1, a_2, ..., a_n)$.

If there are many intervals $[a_j, b_j]$, $i = 1..k$ and $a_j = a_i$ then we will select interval $[a_i, b_i]$ is a interval that b_i the smallest of value b_j. The selection b_i always only to finding, because intervals is given difference each other. Thus, if $a_i = a_j$ then $b_i \neq b_j$ (by the Definition 2).

After to finding the first interval $[a_i, b_i]$ of the sequence, we continue to finding the second interval, etc. After k step to finding and sorting, we obtained the sequence with k interval that elements of the sequence are sorting according to post-preorder relation. ■

3. The HAC4.5 algorithm for fuzzy decision tree data classification problem

3.1. Introduction

The C4.5 algorithm is improved by Quinlan [32]. The C4.5 algorithm, at each node of the tree, C4.5 chooses the attribute of the data that most effectively splits its set of samples into subsets enriched in one class or the other. The splitting criterion is the normalized. The attribute with the highest normalized information gain is chosen to make the decision.

Because fuzzy attribute of the training sample set have partitived according to the fuzzy interval is a sub interval of $[0, 1]$, and domain of value are sorted linear order according to post-preorder relation. We can compare to divide threshold of the this set of value at any interval $I(x) = [I_a, I_b] \subseteq [0, 1]$ similarity as continuous number values in the C4.5 algorithm.

Finding threshold to allow split based on information gain ratio of thresholds in D at that node. Information gain ratio of thresholds for attribute A is number attribute in D at that node.

Suppose that attribute A is a fuzzy attribute have partitived according to the fuzzy interval and there are k difference intervals already sort order according to post-preorder relation: $[I_{a1}, I_{b1}] < [I_{a2}, I_{b2}] < \cdots < [I_{ak}, I_{bk}]$.

We have k thresholds are computed: $Th_i^{HA} = [I_{ai}, I_{bi}], (1 \leq i < k)$. At each threshold Th_i^{HA}, the set of data D of this node are divided into two sets: $D_1 = \{\forall [I_{aj}, I_{bj}] \,|[I_{aj}, I_{bj}] \leq Th_i^{HA})\}$ *and* $D_2 = \{\forall [I_{aj}, I_{bj}] \,|[I_{aj}, I_{bj}] > Th_i^{HA})\}$.

Then, we have:

$$Gain^{HA}(D, Th_i^{HA}) = Entropy(D) - \frac{|D_1|}{|D|} \times Entropy(D_1) - \frac{|D_2|}{|D|} \times Entropy(D_2)$$

$$SplitInfo^{HA}(D, Th_i^{HA}) = -\frac{|D_1|}{|D|} \times log_2 \frac{|D_1|}{|D|} - \frac{|D_2|}{|D|} \times log_2 \frac{|D_2|}{|D|}$$

$$GainRatio^{HA}(D, Th_i^{HA}) = \frac{\mathrm{Gain^{HA}(D, Th_i^{HA})}}{\mathrm{SplitInfo^{HA}(D, Th_i^{HA})}}$$

Based on compute information gain ratio of thresholds, we select threshold that information gain ratio is the biggest to split D into two subsets.

3.2. The HAC4.5 algorithm

Input: Training data set D.
Output: Fuzzy decision tree S.
Method:
For each (fuzzy attribute X in D)

Begin
Built a hedge algebra $\underline{X_k}$ corresponding with fuzzy attribute X;
Transform number values and linguistic values of X into intervals $\subseteq [0, 1]$;
End;
Set of leaf node S; S = D;
For each (leaf node L in S)
If (L homogenise) or (L set of attribute is empty) then
L.Label = *Class name*;
Else
Begin
X is attibute have GainRatio or GainRatioHA is the biggest;
L.Label = *Attribute name* X;
If (L *is fuzzy attribute*) Then
Begin
T = Threshold have $GainRatio^{HA}$ is the biggest;
$S1 = \{Ix_i | Ix_i \subseteq L, Ix_i \leq T\}$;
S1.Father node = L;
S1.Attribute = L.Attribute − X;
$S_2 = \{Ix_i | Ix_i \subseteq L, Ix_i > T\}$;
S_2.Father node = L;
S_2.Attribute = L.Attribute - X;
$S = S + S_1 + S_2 - L$;
End
Else
If (L is *continuous attribute*) then
Begin
T = Threshold have GainRatio is the biggest;
$S_1 = \{x_i | x_i \in L, x_i \leq T\}$;
S1.Father node = L;
S1.Attribute = L.Attribute - X;
$S_2 = \{x_i | x_i \in L, x_i > T\}$;
S2.Father node = L;
S2.Attribute = L.Attribute - X;
$S = S + S_1 + S_2 - L$;
End
Else { L is discrete attribute }
Begin
$P = \{x_i | x_i \in K, x_i single\}$;
For (each $x_i \in P$) do
Begin
$S_i = \{x_j | x_j \in L, x_j = x_i\}$;

S_i. Father node = L;
S_i.Attribute = L.Attribute −X ;
$S = S + S_i$;
End;
$S = S - L$;
End;
End;

3.3. Evaluting algorithm

For m is number of the attribute, n is number of training sample set, the complex of the C4.5 algorithm is $O(m \times n \times log\, n)$. In the HAC4.5 algorithm, the first, we loss $O(n^2)$ for each fuzzy attribute, we partitive the fuzzy intervals. After that, the complex of the algorithm at loop step by attribute m_i is $O(n \times log\, n)$ if m_i is not fuzzy attribute, otherwise, then complex of the algorithm is $O(n \times n \times log\, n)$ because we additional losses $O(n)$ to finding the thresholds of the fuzzy intervals for this attribute. Thus, the complex of the HAC4.5 algorithm is $O(m \times n^2 \times log\, n)$.

The soundness of the algorithm is inferred from soundness of the C4.5 algorithm and matching method in section 2.

Because used idea of the C4.5 algorithm so at this division node cannot division by partial k, avoid situation spread by horizontally, thus, the result tree not overfitting. Additional the cost $O(n)$ is not too big should be can accept in the training process, moreover, the training process only done once and used to predict for several times. Due to the partition of the training process based on the concept of interval partition correlation, so the fuzzy decision tree will be obtained can be used to predict in the case by point or interval become advantages for the process of prediction.

4. Experimental evaluation

Experimental programs is installed in the Java language (Eclipse Mars Release (4.5.0) by computer with configuration: Processor Intel $Core^{TM}$i5-2450 CPU @2.50GHz (4CPUs), 2.50 GHz, RAM 4GB, System type 64 bit for three algorithms: the C4.5, based on point homogenise matching method and interval matching by HAC4.5 with two the training sample sets are Mushroom and Adult.

- The training sample set Mushroom there are more 8000 records include 22 attributes, there are two attributes Habitat and Population contain precise data and imprecise data. We have divided 5000 records for the training sample set, in 3000 records remain, we random select 2000 records for testing.

- The training sample set Adult 40000 records include 14 attributes contain discrete data, continuous data, logic and imprecise data, there aretwo attributes Age and HoursPerWeek containprecise data and imprecise data. We have divided 20000 records for the training sample set, in 20000 records remain, we random select 5000 records for testing.

4.1. The results of Mushroom data classification

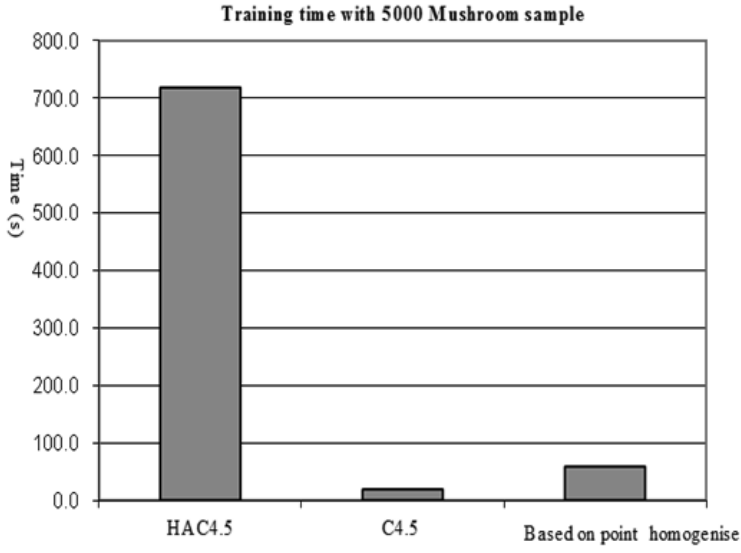

Figure 8. Matching training time in Mushroom sample

Training with 5000 Mushroom sample	
Algorithm	Time (s)
HAC4.5	717.3
C4.5	18.9
Based on point homogenise	58.2

Table 1. Matching training in Mushroom data.

Testing ratio from 100 to 2000 Mushroom data sample

Algorithm	100	500	1000	1500	2000
HAC4.5	82.0%	81.0%	86.1%	88.9%	91.5%
C4.5	57.0%	54.8%	51.2%	66.2%	70.0%
Based on point homogenise	71.0%	72.2%	72.6%	77.9%	77.2%

Table 2. Matching testing ratio in Mushroom data

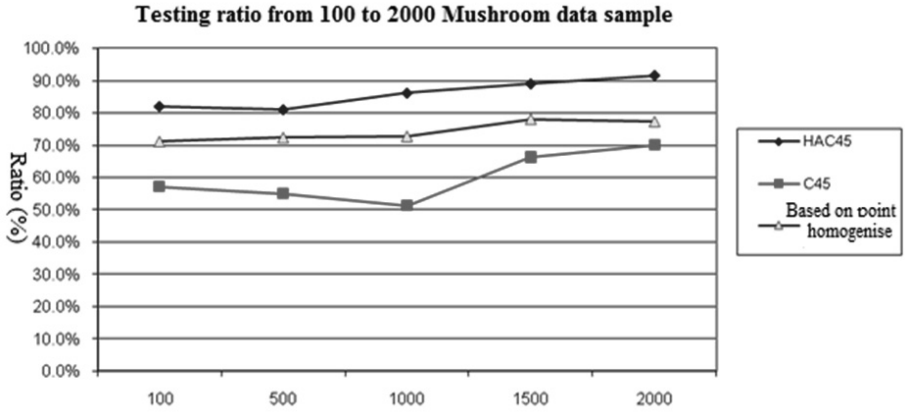

Figure 9. Matching testing ratio from 100 to 2000 in Mushroom data sample

4.2. The results of Adult predict data

Training time in 20000 sample.

Algorithm	Time (s)
HAC4.5	1863.7
C4.5	479.8
Based on point homogenise	589.1

Table 3. Matching training in Adult data

Algorithm	1000	2000	3000	4000	5000
HAC4.5	92.3%	91.5%	93.0%	95.0%	96.1%
C4.5	84.5%	85.7%	85.9%	86.2%	85.7%
Based on point homogenise	87.0%	86.2%	87.4%	87.5%	86.6%

Table 4. Matching testing ratio in Adult data

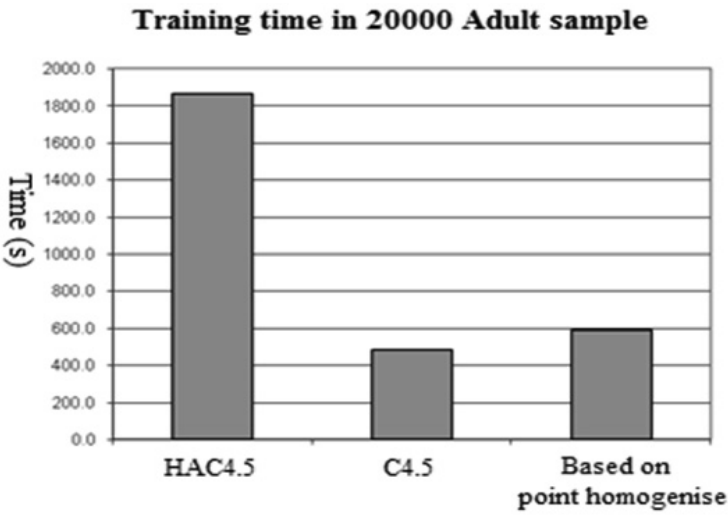

Figure 10. Matching traning time in Adult.

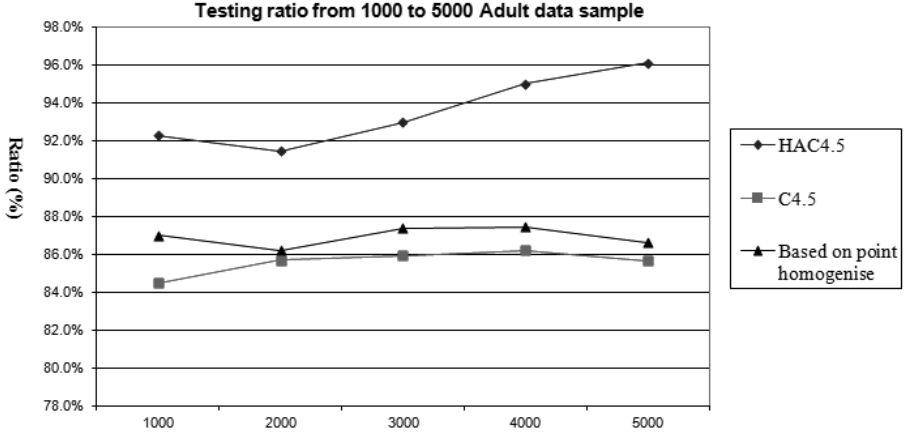

Figure 11. Matching testing ratio from 1000 to 5000 in Adult data sample

Algorithm	1000	2000	3000	4000	5000
HAC4.5	2.4	4.7	7.2	9.7	12.1
C4.5	1.4	2.8	4.1	5.5	6.0
Based on point homogenise	2.2	4.6	7.1	9.2	11.8

Table 5. Matching testing time in Adult data Testing time from 1000 to 5000 sample in Adult data (s)

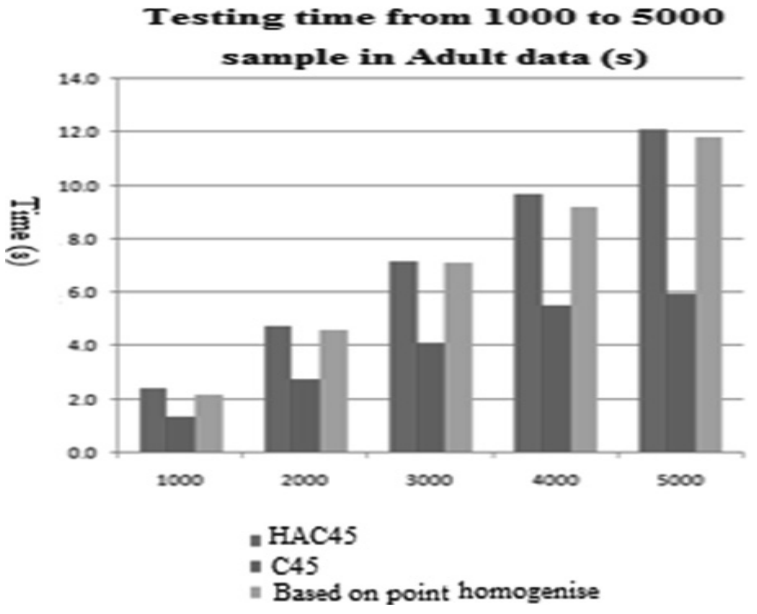

Figure 12. Matching testing time from 1000 to 5000 in Adult data sample

4.3. Experimental evaluation

Installation three the C4.5 algorithm, based on point homogenise matching method and HAC4.5, evaluation results on the same data set as the Mushroom and the Adult, we have obtained:

- The cost of time: The C4.5 algorithm always is the fastest time in all the samples even during training or testing, because it ignores the vague values in the sample set should not loss process time.

We already homogenise the sample set based on the point matching method and then we used this sample set for tree training so have to defined hedge algebras corresspoding the fuzzy attribute and the cost to homogenise original value should take time longer than the C4.5 algorithm.

Because, initially, we through the process of building the hedge algebras for the fuzzy attribute and the cost to convert the value into sub interval of $[0, 1]$, moreover, at each loop step to need additional time to divided select, so the HAC4.5 algorithm is possibly slow than other algorithms.

- **The predictable result.** Because the C4.5 algorithm ignore vague values in the sample set, only interest the precise values so loss data in the fuzzy attribute, thus predictable results not good.

At the fuzzy attribute, we will be building a hedge algebra and use it to homogenise the training sample set by point matching, then we are obtained the homogenise training sample set include precise data and imprecise data, so the result tree is trained will be better. However, in this case, predictable result not good, because for the partitive by fuzzy point will be errors corresponding the precise values at division point.

The predictable result of the HAC4.5 is the best of all, because in the tree training process, we already process vague values but the precise values not change, so there is no errors in partition process.

Although the HAC4.5 to spend more time for training process but it is an effectively method because the result tree with predictable better than other algorithms. Furthermore, we only doing the training process one that predictable based on result tree can doing several times. So, the cost of time of the HAC4.5 algorithm is acceptable.

5. Conclusions

The fuzzy decision tree classification problem is an important role in the process of data mining. However, the fuzzy decision tree classification based on fuzzy set theory have many disadvantages. The hedge algebra have many advantages has become a really useful tool for solving the decision tree classification problems. Recognizing the limitations of quantitative semantics methods the training process, the paper was used hedge algebra to proposed a fuzzy interval matching method, it was based on a new method to inductive learning fuzzy decision tree used the algorithm HAC4.5 effectively was proposed. The time optimization of the HAC4.5 algorithm will be consider in the future paper.

References

[1] **Duong Thang Long,** Method to built fuzzy rule system based on hedge algebra semantic and applied for classification problem, IOIT, 2010.

[2] **Nguyen Cong Hao,** Fuzzy databases with data manipulation based on hedge algebra, Thesis of Doctor mathematic, IOIT, 2008.

[3] **Nguyen Cat Ho and Tran Thai Son,** On distance between linguistic values in hedge algebra, *Journal of Computer Science and Cybernetics*, **11(1)** (1995), 10–20.

[4] **Le Xuan Viet,** Semantic quantitative linguistic values of linguistic variable inhedge algebra and applied, Thesis of Doctor mathematic, IOIT, 2009.

[5] **Bikas, A.K., E. M. Voumvoulakis and N.D. Hatziargyriou,** Neuro-Fuzzy Decision Trees for Dynamic Security Control of Power Systems, Department of Electrical and Computer Engineering, NTUA, Athens, Greece, 2008.

[6] **Abonyi, J., J.A. Roubos and M. Setnes,** Learning fuzzy classification rules from labeled data, *Information Sciences*, **150** (2003).

[7] **Chandra, B.,** *Fuzzy SLIQ Decision Tree Algorithm*, IEEE, 2008.

[8] **Chang, Robin L.P. Pavlidis, Theodosios,** *Fuzzy Decision Tree Algorithms, Man and Cybernetics*, IEEE , 2007.

[9] **Fuller, R.,** *Neural Fuzzy Systems*, Physica-Verlag, Germany, 1995.

[10] **Hesham A.H., S.G. Ahmed and 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., pp. 55-63, Oct. 2010.

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

[12] **Ho, N.C. and H.V. Nam,** An algebraic approach to linguistic hedges in Zadeh's fuzzy logic, *Fuzzy Sets and Systems*, **129** (2002), 229–254.

[13] **Ho, N.C. and W. Wechler,** Hedge algebras: an algebraic approach to structures of sets of linguistic domains of linguistic truth variables, *Fuzzy Sets and Systems*, **35(3)** (1990), 281–293.

[14] **Ho, N.C. and W. Wechler,** Extended algebra and their application to fuzzy logic, *Fuzzy Sets and Systems*, **52** (1992), 259–281.

[15] **Ishibuchi, H. and T. Nakashima,** Effect of rule weights in fuzzy rule-based classification systems, *IEEE Trans. on Fuzzy Systems*, **9(4)** (2001).

[16] **James, F., I. Smith and T.H. Nguyen,** Genetic program based data mining of fuzzy decision trees and methods of improving convergence and reducing bloat, Data Mining, Intrusion Detection, Information Assurance, 2007

[17] **Lan L.V.T., N.M. Han and N.C. Hao,** A novel method to build a fuzzy decision tree based on hedge algebras, *International Journal of Research in Engineering and Science*, **4(4)** (2016), 16–24.

[18] **Lee, C., S. George and T.C. Lin,** Neural fuzzy systems, *A Neuro-Fuzzy Synergism to Intelligent Systems*, Prentice-Hall International, Inc, 1995.

[19] **Moustakidis, S., G. Mallinis, N. Koutsias, J.B. Theocharis and Petridis,** SVM-Based Fuzzy Decision Trees for Classification of High Spatial Resolution Remote Sensing Images, Geoscience and Remote Sensing, IEEE, 2012.

[20] **Manish, M., J. Rissanen and R. Agrawal,** SLIQ: A Fast Scalable Classifier for Data Mining, IBM Almaden Reseach Center, 1996.

[21] **Manish, M., J. Rissanen and R. Agrawal,** SPRINT: A Fast Scalable Classifier for Data Mining, IBM Almaden Reseach Center, 1996.

[22] **Peer, F., D. Parveen and M. Sathik,** Fuzzy decision tree based effective IMine indexing, *International Journal of Computer Technology and Electronics Engineering* (*IJCTEE*), **1(2)** (2011).

[23] **Quinlan, J.R.,** Simplifying decision trees, *International Journal of Man-Machine Studies*, **27** (1987), 221–234.
http://www.mlrg.cecs.ucf.edu/MLRG_documents/c4.5.pdf

[24] **Tajiri, R.H., Z.M. Eduardo, B.Z. Bruno and S.M. Leonardo,** A new approach for fuzzy classification in relational databases, *Database and Expert Systems Applications*, Springer, pp. 511–518, 2011.

[25] **Ruggieri, S.,** *Efficient C4.5*, University Di Pisa, 2000.

[26] **Wang, T. and H. Lee,** Constructing a fuzzy decision tree by integrating fuzzy sets and entropy, *ACOS'06 Proceedings of the 5th WSEAS international conference on Applied computer science*, World Scientific and Engineering Academy and Society, USA, 2006, pp. 306–311.

[27] **Wei-Yuan Cheng and Chia-Feng Juang,** A fuzzy model with online incremental SVM and margin-selective gradient descent learning for classification problems, *IEEE Transactions on Fuzzy systems*, **22(2)** (2014), 324–337.

[28] **Zadeh, L.A.,** Fuzzy sets, *Information and Control*, **8** (1985), 338–358.

[29] **Zadeh, L.A.,** Fuzzy sets and fuzzy information granulation theory, Beijing Normal University Press, China, 2000.

[30] **Zengchang, Q. and J. Lawry,** *Linguistic Decision Tree Induction*, Department of Engineering Mathematics, University of Bristol, United Kingdom, 2007.

[31] **Zengchang, Qin and Yongchuan Tang,** Linguistic decision trees for classification, Uncertainty Modeling for Data Mining, Springer, 77–119, 2014.

[32] **Zhang, J. and Honavar,** Learning decision tree classifiers from attribute-value taxonomies and partially specified data, *Proceedings of the International Conference on Machine Learning*, Washington DC, 2003.

L. V. T. Lan, N. M. Han and N. C. Hao
Hue University
3 Le Loi st.
Hue City
VietNam
lvtlan@yahoo.com
nmhan2005@yahoo.com
nchao@hueuni.edu.vn