ELSEVIER

Fuzzy Sets and Systems 99 (1998) 283–290

FUZZY sets and systems

On the handling of fuzziness for continuous-valued attributes in decision tree generation

Wang Xizhao[a,*], Jiarong Hong[b]

[a] *Department of Mathematics, Hebei University, Baoding 071002, Hebei, People's Republic of China*
[b] *Department of Computer science, Harbin Institute of Technology, Harbin 150001, People's Republic of China*

Received August 1996; revised December 1996

Abstract

In this paper, fuzziness existing in the process of generating decision trees by discretizing continuous-valued attributes is considered. In a sense a better way to express this fuzziness via fuzzy numbers is presented using possibility theory. The fact that selection of membership functions in a class of symmetric distributions does not influence the decision tree generation is proved. The validity of using the tree to classify future examples is explained. On the basis of likelihood possibility maximization, the existing algorithm is revised. The revised algorithm leads to more reasonable and more natural decision trees. © 1998 Published by Elsevier Science B.V. All rights reserved.

Keywords: Possibility theory; Measures of information; Knowledge acquisition; Learning; Decision trees

1. Introduction

Learning algorithms based on decision tree generation are of the most powerful heuristics in inductive learning. One optimally inductive learning method is to generate all possible decision trees that correctly classify the training set and to select the simplest of them. The number of such trees is finite but very large, so the method is computationally burdensome and is feasible only for very small training set. ID3 is an algorithm designed for generating a reasonably good decision tree without much computation [8]. The ID3 algorithm can generate a simpler decision tree by using minimization of class information entropy but cannot guarantee generation of the simplest tree. So far, the ID3 algorithm, which can conveniently denote the information structure between attributes of concepts and attribute values, has been one of the most powerful algorithms. Some scholars, depending on their different needs, present many extensions of the ID3 algorithm, such as GID3 [1] and GID3* [4]. Other algorithms based on decision tree generation appeared in quick succession, e.g. N2 [2] and C4 [9].

Generally, attributes in a learning problem can be divided into two classes, namely, discrete-valued attributes and continuous-valued attributes. The former are regarded as nominal (categorical) notions while the latter, as real numbers. The above algorithms assume that all attribute values are nominal. Continuous-valued attributes must, therefore, be discretized prior to attribute selection. There are various ways for discretization but a practical one is binary partition which means that a continuous-valued attribute is discretized during decision tree generation by partitioning its range into two intervals. A threshold value, T, for the continuous-valued attribute A is

* Corresponding author.

PII S 0165-0114(97)00030-4

Table 1
A leaf node in a decision tree

Example	1	2	3	4	5	6	7	8	9	10
Value of attribute A	20	21	22	50	56	60	68	72	75	81
Class	−	−	−	+	+	+	+	+	+	+

determined and the set $[A \leqslant T]$ is assigned to the left branch, whereas $[A > T]$ is assigned to the right branch. The threshold value T is called a cut point. This method for selecting a cut point, which is used in the ID3 algorithm and its variants, involves choosing a particular discretization among several possible ones. In [5], a result about the information entropy minimization heuristic used in discretizing continuous-valued attributes is derived.

However, when a discretization is chosen for a given continuous-valued attribute A, values of the attribute A will possess fuzziness with respect to branching. To illustrate this kind of fuzziness, we consider a leaf node having 10 examples in a decision tree and a continuous-valued attribute A (see Table 1). By computing the information entropy, the best cut point, T, for discretization should be in the interval $(22, 50)$. When the left branch $[A \leqslant T]$ and the right branch $[A > T]$ are used to classify future examples, the value of T is usually taken to be the midpoint of the interval, namely 36, without considering concrete structure of attribute values. Obviously, this method is not very reasonable because each value in the interval $(22, 50)$ has the possibility for it to appear as positive example or as negative example. For solving this problem, a new technique, soft thresholds, has been presented by Quinlan in 1993 (see [10]). In that case, some kind of weighting is used to soften absolute thresholds. In this paper, we further discuss this problem and consider whether the selection of the cut point T (i.e. the threshold) can be by fuzziness.

Using possibility theory, this paper deals with this kind of fuzziness and gives us

- a better understanding of fuzziness in the process of discretization,
- a better selection of membership functions in order to describe this fuzziness,
- the influence upon decision tree generation and
- the validity of using the tree to classify future examples.

In Section 2 we discuss the procedure of the algorithm for decision tree generation with handling fuzziness and in Section 3 we give the theoretical foundation for supporting this algorithm and prove our main results of this paper.

2. Algorithm for decision tree generation with handling fuzziness

2.1. Selecting the best cut point

In the process of decision tree generation, a binary partition is usually regarded as a discretization for continuous-valued attributes. This partition should be chosen so as to provide useful classification information with respect to the classes to which the examples in the attribute's range belong. Often a cut point, T, is selected for a continuous-valued attribute A such that the set "$A \leqslant T$" is assigned to the left branch while "$A > T$" is assigned to the right branch. That is, the space of all examples will be divided into two parts by using a cut point. To explain the "best" cut point, we need the following definition (see [5]).

Let E be the set of all examples considered, and let there be k classes $C_1, C_2, \ldots, C_k$, with the property $E = \bigcup_{i=1}^{k} C_i$, $C_i \cap C_j = \emptyset$ $(i \neq j)$, $S \subset E$ $(S \neq \emptyset)$. Then the class entropy of the subset S is defined as

$$Entr(S) = -\sum_{i=1}^{k} P(C_i, S) \log P(C_i, S),$$

where $P(C_i, S) = |S \cap C_i|/|S|$ $(i = 1, 2, \ldots, k)$. $|\bullet|$ denotes the number of elements of a set and the logarithm may be to any convenient base (we define $x \log x = 0$ if $x = 0$).

Definition 1. Let S be a set of examples, A a continuous-valued attribute, and T a cut point. According to the value of the attribute A, the set S can be

Table 2

No.	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
A	15	17	19	21	22	24	26	28	30	31	35	36	37	39	42	43	45	46
Class	+	+	+	−	−	−	−	+	+	+	−	−	−	−	+	+	+	+

divided into two subsets, $S_1 = \{e \mid e \in S,\ A(e) \leqslant T\}$ and $S_2 = S - S_1$. Then, the class information entropy of the partition induced by T, denoted by $E(A, T, S)$, is defined as

$$E(A,T,S) = \frac{|S_1|}{|S|} Entr(S_1) + \frac{|S_2|}{|S|} Entr(S_2),$$

where $|\bullet|$ denotes the number of elements of a set.

Assume we are to select a continuous-valued attribute A for branching at a node having a set S of N examples. Let there be k classes $C_1, C_2, \ldots, C_k$, and suppose N examples do not have identical attribute values (if there are identical attribute values, the similar way for handling can be used). The N examples which are first sorted by increasing value of attribute A can be divided into M families ($M \leqslant N$), denoted by $F_1, F_2, \ldots, F_M$. Then there exists $j (j \leqslant k)$ for each $i (i \leqslant M)$ such that $F_i \subset C_j$, i.e. all examples of each family belong to the same class. For instance, Table 2 gives us a special case of $k = 2$, $N = 18$ and $M = 5$ for the attribute A, where $C_1 = \{1, 2, 3, 8, 9, 10, 15, 16, 17, 18\}, C_2 = \{4, 5, 6, 11, 12, 13, 14\}, F_1 = \{1, 2, 3\}, F_2 = \{4, 5, 6, 7\}, F_3 = \{8, 9, 10\}$, $F_4 = \{11, 12, 13, 14\}$, $F_5 = \{15, 16, 17, 18\}$.

Between each successive pair of families, a value is chosen. These values are usually regarded as the midpoints (see [5]), called candidate cut points. Hence, $M-1$ candidate cut points are obtained. Among these $M-1$ candidate cut points, $T_1, T_2, \ldots, T_{M-1}$, a particular point, e.g. T_A, can be selected such that the class information entropy of partition induced by the point T_A attains minimum. That is,

$$E(A, T_A, S) = \min_{1 \leqslant j \leqslant M-1} E(A, T_j, S).$$

Then T_A is regarded as the best cut point in the sense of class information entropy minimization. This determines a binary discretization for attribute A.

After all continuous-valued attributes have been discretized, a particular attribute should be selected for branching out of the node. In algorithms that use information entropy minimization for attribute selection, the attribute A^*, for which $E(A^*, T_{A^*}, S)$ is minimal, is the selected attribute among all continuous-valued attributes.

2.2. Revising the best cut point

μ is said to be a fuzzy number if it is a convex, closed fuzzy set on R (the real line). We call the set $\{x \mid \mu(x) > 0\}$ the support of the fuzzy number. A fuzzy number μ is called 0-symmetric if $\mu(0) = 1$ and $\mu(x) = \mu(-x)$ for each $x \in R$ (see [12]). The following is the well-known characteristic theorem of a fuzzy number.

Proposition 1. *Let μ be a continuous fuzzy number. Then μ has the following properties:*

(1) *There exists an interval $[m_\mu, n_\mu]$ such that $\mu(x) = 1$ for each $x \in [m_\mu, n_\mu]$.*
(2) *μ monotonically increases for $x < m_\mu$ and monotonically decreases for $x > n_\mu$.*
(3) $\mathrm{Lim}_{x \to \infty} \mu(x) = 0$.

For a given 0-symmetric fuzzy number μ with the property that the set $\{x \mid \mu(x) = 1\}$ consists of only one point, we denote

$$\left\{ \mu\left(\frac{x-a}{b}\right) \,\middle|\, a \in R, b > 0 \right\} \text{ by } \Omega_\mu$$

Ω_μ is called a family of fuzzy numbers, generated by μ, which is similar to the family of distribution functions in probabilistic statistics where a and b are location parameter and scale parameter, respectively.

Now we consider the fuzziness for branching and revise the selected cut point T_{A^*}. Let two families of examples be F_1 and F_2 (the family F_1 lies on the left of T_{A^*} while the family F_2 lies on the right of T_{A^*}, without loss of generality). F_1 and F_2 are regarded as two elements of the family Ω_μ. According to values of the selected attribute A^*, two fuzzy numbers can be estimated for describing F_1 and F_2 (see the following

Table 3
A problem of learning from example

No.	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
$A1$	15	17	19	21	22	24	26	28	30	31	35	36	37	39	42	43	45	46
$A2$	101	105	115	107	109	110	112	117	123	124	114	119	120	122	126	127	129	130
$A3$	310	312	313	316	318	330	331	315	320	322	332	333	335	337	325	327	340	345
Class	+	+	+	−	−	−	−	+	+	+	−	−	−	−	+	+	+	+

section). Two fuzzy numbers, whose supports are the real line, should have better membership functions in a sense, e.g. membership functions determined in the following way.

Let the ith family F_i have t examples which belong to the same class. The sorted values of the attribute A^* are $a_1, a_2, \ldots, a_t$ $(t \geqslant 2)$. Then a better membership function describing F_i, in a sense, is

$$\mu^{F_i}(x) = \exp\left(-\left[\frac{2x - a_1 - a_t}{a_t - a_1}\frac{\sqrt{2}}{2}\right]^2\right).$$

When $t = 1$, we take a sufficiently small, positive number δ and regard the membership function describing F_i as

$$\mu^{F_i}(x) = \exp\left(-\left[\frac{x - a_1}{\delta}\right]^2\right).$$

There are many forms of $\mu^{F_i}(x)$, but our main objective is to evaluate the cross point of two membership functions, and this does not depend on the concrete forms of $\mu^{F_i}(x)$ (see the following section).

The cross point of $\mu^{F_1}(x)$ and $\mu^{F_2}(x)$ is assumed to be $\overline{T_{A^*}}$ and the formula for computing this cross point is given in the following section. Suppose F_1 contains n sorted values $x_1, \ldots, x_n$ and F_2 contains m sorted values $y_1, \ldots, y_m$ $(x_n < y_1)$. It is clear that $\mu^{F_1}(x) < \mu^{F_2}(x)$ if $x > \overline{T_{A^*}}$ and $\mu^{F_1}(x) > \mu^{F_2}(x)$ if $x < \overline{T_{A^*}}$ for each $x \in [x_n, y_1]$. $\mu^{F_1}(x)$ and $\mu^{F_2}(x)$ represent the possibility with which x appears in F_1 and F_2, respectively. We prefer a bigger possibility with which x appears. That is to say, x should belong to the left branch if $x < \overline{T_{A^*}}$ and to the right branch if $x > \overline{T_{A^*}}$. The point $\overline{T_{A^*}}$, therefore, is regarded as the revised value of the best cut point for the selected attribute A^*.

When, in turn, attributes are to be selected for partitioning the child nodes, the discretization process must be performed again to rederive a new quantization based on each child node's own examples.

2.3. Procedure for generating decision tree

The procedure for decision tree generation with handling fuzziness is given as follows.

SELECT a node having N examples for branching

(a) Take an attribute and sort N examples by increasing the value of the attribute. Then, obtain several "families" (e.g. Table 2).

(b) Take a point (e.g. the midpoint) between each successive pair of families and regard it as a candidate cut point. Compute the class entropy induced by each candidate cut point and select the point with minimal entropy (the best cut point).

(c) Determine the best cut point for each attribute. Select the attribute whose best cut point, denoted by T, has minimal entropy among all attributes. Thus, two families which are located on the left of T and on the right of T are determined.

(d) Compute the cross point of two membership functions which describe these two families (The formula is given in Section 3.3). By the explanation given in Section 2.2, the cross point is regarded as the best revised cut point.

(e) According to the best revised cut point of the selected attribute, the tree branches at the selected node.

REPEAT the above process for a selected child node until classification ends.

2.4. Example

Consider a problem of learning from example given in Table 3 where there are 10 positive examples, 8 negative examples and 3 continuous-valued attributes ($A1, A2,$ and $A3$).

Using the algorithm listed in Section 2.3 and the algorithm in [5], we can get decision trees 1 and decision tree 2.

Decision tree 1
Root

{node1.($A2 \leqslant 121.38$)
  {node11.($A3 \leqslant 315.21$) : Positive
    node12.($A3 > 315.21$) : Negative
  }
node2.($A2 > 121.38$)
  {node21.($A2 \leqslant 122.13$) : Negative
    node22.($A2 > 122.13$) : Positive
  }
}

Decision tree 2
Root

{node1.($A2 \leqslant 122.50$)
  {node11.($A3 \leqslant 315.50$) : Positive
    node12.($A3 > 315.50$) : Negative
  }
node2.($A2 > 122.50$) : Positive
}

Consider $e = (21, 120, 315.4)$, the classification result is negative by using decision tree 1, but the classification result is positive by using decision tree 2. Assigning to example e to be negative is more reasonable because the possibility with which e appears in node 12 is larger than the possibility in node 11 where node 11 = [positive examples: 1, 2, 3, 8] and node 12 = [negative examples: 4, 5, 6, 7, 11, 12, 13].

3. Theoretical foundation of the algorithm

3.1. Class information entropy minimization

Information entropy minimization used in classification learning algorithms is the most powerful heuristic. It has many advantages in optimal learning although it cannot result in the simplest decision tree. The following proposition guarantees that the entropy induced by the best revised cut point obtained in our algorithms really attains minimum among entropies induced by all possible candidate cut points, if the best revised cut point does not lie inside of some family.

Proposition 2. *Let T be a cut point, for which $E(A,T,S)$ is minimal among entropies induced by all possible candidate cut points. Then, T is a boundary point, i.e. a cut point with the following property: In the sequence of examples sorted by the values of attribute A, there exist two examples $e1 \in S$ and $e2 \in S$ having different classes such that $A(e1) < T < A(e2)$, and there exist no other examples $e' \in S$ such that $A(e1) < A(e') < A(e2)$.*

Proof. It is a counterpart of Theorem 1 in [5].

3.2. Maximal likelihood possibility for appearance of sample

According to the possibility theory, a fuzzy number, μ, can be regarded as a possibility distribution. The membership degree $\mu(x)$ for each $x \in R$ is considered the possibility with which the point x appears. The details on possibility theory can be found in [3].

The integral of a membership function μ on the real line is called fuzzy entropy of the possibility distribution μ, denoted by $E[\mu]$ (see [12]). The fuzzy entropy denotes a kind of uncertainty (fuzziness) of the distribution. Obviously, the greater $E[\mu]$ is, the higher the uncertainty. Particularly, when $E[\mu]$ is almost zero, μ almost becomes a real number.

In the following, we take a fixed, 0-symmetric, continuous fuzzy number μ whose support is assumed to be R. Suppose the set $\{x \mid \mu(x) = 1\}$ consists of only one point, and consider Ω_μ, the possibility distribution family generated by μ defined as

$$\Omega_\mu = \left\{ Q(x;a,b) \mid Q_\mu(x;a,b) = \mu\left(\frac{x-a}{b}\right),\ a \in R,\ b > 0 \right\},$$

where a and b are the location parameter and the scale parameter, respectively, which is similar to the family of probabilistic distribution functions.

Consider a parameter estimation problem. A fuzzy number (possibility distribution) has membership function v which belongs to Ω_μ and the parameters of v, a and b remain to be determined. A crisp sample from the distribution v, $(x_1, x_2, \ldots, x_m)$, is known. The problem is how to reasonably estimate parameters a and b by using the sample.

To obtain estimators of a and b, we apply Maxmin μ/E estimation principle [3, 6, 11, 12]. Let the membership function be $Q_\mu(x; a, b)$, and $(x_1, x_2, \ldots, x_m)$ a sample. We denote by

$$L(a,b) = \min_{1 \leqslant i \leqslant m} Q_\mu(x_i; a, b) / E[Q_\mu]$$

the likelihood possibility with which the sample appears.

As a bigger possibility is preferred, we naturally require $L(a,b)$ is as big as possible. The Maxmin estimators of a and b are defined to be $\hat{a}$ *and* $\hat{b}$ satisfying $L(\hat{a}, \hat{b}) = \max_{a \in R, b>0} L(a,b)$. The Maxmin estimators of a and b have many advantages such as sufficiency and consistency (see [6, 11]). The following proposition gives us the formula of evaluating $(\hat{a}, \hat{b})$.

Proposition 3. *The Maxmin estimator of the parameter* $\theta = (a,b)$ *is*

$$(\hat{a}, \hat{b}) = ((x^{(1)} + x^{(m)})/2, (x^{(m)} - x^{(1)})/2c),$$

where $x^{(1)} \leqslant x^{(2)} \leqslant \cdots \leqslant x^{(m)}$ *are the ordered values of* $x_1, x_2, \ldots, x_m$ $(m \geqslant 2)$, *and* c *is a real number at which the function* $g(t) = t\mu(t)$ $(t \geqslant 0)$ *attains its maximum.*

Proof. Let $\int_{-\infty}^{\infty} \mu(t)\,dt = l$, and let $\wedge$ denote "min" and $\vee$ denote "max". It is easy to compute that $E[Q_\mu] = lb$. Therefore,

$$\begin{aligned} L(a,b) &= \bigwedge_{j=1}^{m} \mu\left(\frac{x_j - a}{b}\right) \Big/ (lb) \\ &= \bigwedge_{j=1}^{m} \mu\left(\frac{x^{(j)} - a}{b}\right) \Big/ (lb) \\ &= \left[\mu\left(\frac{x^{(1)} - a}{b}\right) \wedge \mu\left(\frac{x^{(m)} - a}{b}\right)\right] \Big/ (lb). \end{aligned}$$

For any given $b > 0$,

$$\begin{aligned} &\max_{a \in R} L(a,b) \\ &= \frac{1}{lb} \bigvee_{a \in R} \left[\mu\left(\frac{x^{(1)} - a}{b}\right) \wedge \mu\left(\frac{x^{(m)} - a}{b}\right)\right] \\ &= \frac{1}{lb} \mu\left(\frac{x^{(m)} - x^{(1)}}{2b}\right) \\ &= L\left(\frac{x^{(1)} + x^{(m)}}{2}, b\right) \end{aligned}$$

The validity of the second equality above results from the following fact: let μ be a 0-symmetric, convex function, $\alpha = a - (x^{(m)} + x^{(1)})/2b$, and $t = (x^{(m)} - x^{(1)})/2b$. Then

$$\begin{aligned} &\bigvee_{a \in R} \left[\mu\left(\frac{x^{(1)} - a}{b}\right) \wedge \mu\left(\frac{x^{(m)} - a}{b}\right)\right] \\ &= \bigvee_{a \in R} \left[\mu\left(\frac{a - x^{(1)}}{b}\right) \wedge \mu\left(\frac{x^{(m)} - a}{b}\right)\right] \\ &= \bigvee_{a \in R} [\mu(t - \alpha) \wedge \mu(t + \alpha)] \\ &\leqslant \bigvee_{a \in R} \left[\tfrac{1}{2}(\mu(t - \alpha) + \mu(t + \alpha))\right] \\ &\leqslant \bigvee_{a \in R} \mu(t) = \mu(t). \end{aligned}$$

Hence,

$$L\left(\frac{x^{(1)} + x^{(m)}}{2}, b\right) = \frac{2t\mu(t)}{l(x^{(m)} - x^{(1)})}.$$

By the assumption that $t\mu(t)$ attains the maximum at $t = c$, $L((x^{(1)} + x^{(m)})/2, b)$ attains its maximum at $b = (x^{(m)} - x^{(1)})/2c$. Therefore, the Maxmin μ/E estimation of parameter (a,b) is

$$(\hat{a}, \hat{b}) = ((x^{(1)} + x^{(m)})/2, (x^{(m)} - x^{(1)})/2c).$$

Hence, the proof is completed. □

Now, we discuss the revision of the best cut point. When N examples are sorted by increasing value of the attribute A, M families, $F_1, F_2, \ldots, F_M$, are obtained (see Sections 3.1 and 3.2). Suppose the best cut point is located between F_1 and F_2 (without loss of generality). Each family, F_i ($i = 1$ or 2), is regarded as a fuzzy number in the abstract and values of the attribute A in the family F_i are considered as a crisp sample of the fuzzy number. After the initial fuzzy number, μ, is chosen, the membership function describing F_i ($i = 1$ or 2) can be obtained by using Proposition 3. The revised value of the best cut point can be determined by computing the cross point of two membership functions which describe the families F_1 and F_2. The evaluation of the cross point and the selection of the initial fuzzy number are shown in the following subsections.

3.3. Steadiness of decision tree

From the above two subsections we know that the decision tree is generated by using a number of binary partitions induced by the best revised cut points. The best revised cut point is obtained by evaluating a cross point of two membership functions. So far, we have used a family Ω_μ which is generated by a 0- symmetric fuzzy number μ. We have no limitation of the shape of μ but continuity and R-support. A problem is which are the different shapes of μ, and how to influence the decision tree generation (i.e. how to influence the best revised cut point). The following proposition gives a satisfactory answer. To a large extent, the selection of membership functions does not influence decision tree generation.

Proposition 4. *Let μ be a given 0-symmetric fuzzy number with continuous membership function and R-support, and let $Q_\mu(x;a_1,b_1)$ and $Q_\mu(x;a_2,b_2)$, which are two fuzzy numbers describing a successive pair of families of attribute values, be generated by using Maxmin μ/E estimation. Then the cross point of these two membership functions does not depend on the selection of μ.*

Proof. Let the first sample be $(x^{(1)},x^{(2)},\ldots,x^{(m)})$ and the second sample $(y^{(1)},y^{(2)},\ldots,y^{(n)})$ $(n,m\geqslant 2)$. By Proposition 3, we know that

$$a_1=\frac{x^{(1)}+x^{(m)}}{2},\quad b_1=\frac{x^{(m)}-x^{(1)}}{2c},$$

$$a_2=\frac{y^{(1)}+y^{(n)}}{2},\quad b_2=\frac{y^{(n)}-y^{(1)}}{2c}.$$

To evaluate the cross point, we put $Q_\mu(x;a_1,b_1)=Q_\mu(x;a_2,b_2)$. According to Proposition 1, we obtain $(x-a_1)/b_1=-(x-a_2)/b_2$ which implies that the cross point is $T=(a_1b_2+a_2b_1)/(b_1+b_2)$, i.e.

$$T=\tfrac{1}{2}\{[(x^{(1)}+x^{(m)})(y^{(n)}-y^{(1)})+(y^{(1)}+y^{(n)})\times(x^{(m)}-x^{(1)})]/[(y^{(n)}-y^{(1)})+(x^{(m)}-x^{(1)})]\}$$

As shown in the above equality, the cross point, T, does not depend on the selection of μ. Hence, the proof is completed. □

Note. When a family contains only one sample point, e.g. x, we may regard $x-\varepsilon$ and $x+\varepsilon$ as a sample $(x^{(1)},x^{(2)})$ where ε is a sufficiently small, positive number. The above equality gives us a practical formula for computing the cross point (i.e. the best revised cut point).

3.4. Validity of using the tree to classify future examples

In the process of generating decision tree, we first select an attribute for branching at a node having a set of examples. Then, according to the best revised cut point obtained by evaluation, the left branch and the right branch are generated. These two branches can be described by two fuzzy numbers, denoted by μ_L and μ_R, respectively. Let the best revised cut point be T, and let two fuzzy numbers be $\mu_1=Q_\mu(x;a_1,b_1)$ and $\mu_2=Q_\mu(x;a_2,b_2)(a_1<a_2)$, which lead to the best revised cut point T. Then the left and right branches may be denoted by

$$\mu_L(x)=\begin{cases}\mu_1(x), & x>a_1,\\ 1, & x\leqslant a_1,\end{cases}\quad\text{and}$$

$$\mu_R(x)=\begin{cases}\mu_2(x), & x<a_2,\\ 1, & x\geqslant a_2,\end{cases}$$

respectively. It is clear that every value of the attribute in the interval $[a_1,a_2]$ possesses fuzziness and can be assigned to the left branch or the right branch. We have the following result.

Proposition 5. *When the decision tree is used to classify a future example, classification possesses the property that a value of the attribute considered is assigned to the left branch (the right branch) if and only if the possibility with which the value appears in the left branch (the right branch) is equal to or greater than the possibility with which the value appears in the right branch (the left branch).*

Proof. By Proposition 1, $\mu_L(x)$ monotonically decreases if $x>a_1$ and $\mu_R(x)$ monotonically increases if $x<a_2$. A new example, e, will belong to the left branch if $A(e)\leqslant T$ and will belong to the right branch if $A(e)>T$. This implies that $A(e)\leqslant T\Leftrightarrow\mu_L(A(e))\geqslant\mu_R(A(e))\Leftrightarrow$ the possibility

with which $A(e)$ appears in the left branch is greater than the possibility with which $A(e)$ appears in the right branch. Thus, the proof is completed. □

4. Conclusions

This paper presents the existence of fuzziness for continuous-valued attributes in decision tree generation; in a sense, gives a better way to express this kind of fuzziness by means of fuzzy numbers; proves that the decision tree generation does not depend on the selection of membership functions in a symmetric distributed family; and explains the validity of using the tree to classify future examples.

The algorithm in this paper is a modification of that in paper [5]. Whereas paper [5] regards the best cut point as the midpoint, this paper regards it as the cross point of two membership functions.

Comparing with paper [5], this algorithm has the following advantages:

(1) it presents a new method of softening the threshold value;
(2) it generates decision trees more reasonably and more naturally;
(3) the revised cut point selection is supported by the possibility theory;
(4) fuzziness in attributes is handled but rules (decision trees generated) are still crisp.

References

[1] J. Cheng, U.M. Fayyad, K.B. Irani and Z. Qian, Improved decision trees: a generalized version of ID3, *Proc. 5th Int. Conf. on Machine Learning* (Morgan Kaufmann, San Mateo, CA, 1988) 100–108.
[2] P. Clark and T. Niblett, The CN2 induction algorithm, *Machine Learning* **3** (1989) 261–284.
[3] D. Dubois and H. Prade, *Possibility Theory* (Plenum Press, New York, 1988).
[4] U.M. Fayyad and K.B. Irani, A machine learning algorithm (GID3*) for automated knowledge acquisition: improvements and extensions, GM Research labs, Warren MI (1991).
[5] U.M. Fayyad and K.B. Irani, On the handing of continuous-valued attributes in decision tree generation, *Machine Learning* **8** (1992) 87–102.
[6] W. Nather, On possibilistic inference, *Fuzzy Sets and Systems* **36** (1990) 327–337.
[7] W. Nather and M. Albrecht, Fuzzy model fitting based on the truth of the model, Freiberger, Forschungshefte D 187 (1987).
[8] J.R. Quinlan, Induction of decision tree, *Machine Learning* **1** (1986) 81–106.
[9] J.R. Quinlan, Probabilistic decision trees, in: Y. Kodratoff and R. Michalski, eds., *Machine Learning: An Artificial Intelligence Approach*, **3** (Morgan Kaufmann, San Mateo, CA, 1990).
[10] J.R. Quinlan, *C4.5: Programs for Machine Learning* (Morgan Kaufmann, San Mateo, CA, 1993).
[11] W. Xizhao and H. Minghu, Note on maxmin μ/E estimation, *Fuzzy Sets and Systems*, to appear.
[12] W. Zhenyuan and L. Shoumei, Fuzzy linear regression analysis of fuzzy-valued variables, *Fuzzy Sets and Systems* **36** (1990) 125–136.