A Comparative Study on Heuristic Algorithms for Generating Fuzzy Decision Trees

X.-Z. Wang, D. S. Yeung, *Senior Member, IEEE*, and E. C. C. Tsang, *Member, IEEE*

***Abstract*—Fuzzy decision tree induction is an important way of learning from examples with fuzzy representation. Since the construction of optimal fuzzy decision tree is NP-hard, the research on heuristic algorithms is necessary. In this paper, three heuristic algorithms for generating fuzzy decision trees are analyzed and compared. One of them is proposed by the authors. The comparisons are two fold. One is the analytic comparison based on expanded attribute selection and reasoning mechanism; the other is the experimental comparison based on the size of generated trees and learning accuracy. The purpose of this study is to explore comparative strengths and weaknesses of the three heuristics and to show some useful guidelines on how to choose an appropriate heuristic for a particular problem.**

***Index Terms*—Approximate reasoning, fuzzy decision trees, fuzzy rules, heuristic algorithms, learning, learning from fuzzy examples.**

I. Introduction

THERE have been many methods for constructing decision trees from collections of crisp examples [16]. The decision trees generated by these methods are useful in building knowledge-based expert systems. Due to the rapid growth of uncertainty in the knowledge-based systems, it is found that using crisp decision trees alone to acquire imprecise knowledge is not enough. Uncertainty such as fuzziness and ambiguity should be incorporated into the process of learning from examples such as decision tree induction. Approaches to fuzzy decision tree generation have been suggested by many authors (e.g., [7], [8], [18], [19], [21], [26], [27]).

The fuzzy decision tree with minimal number of leaf-nodes is usually thought to be optimal. It can be regarded as a commonsense application of Occam's Razor [1]. However, the optimal (fuzzy) decision tree generation has been proved to be NP-hard [6], [20]. Therefore, the research on heuristic algorithms is necessary. The heuristic information used in constructing fuzzy decision trees can be various and each heuristic may be better than the other in some aspects. Mainly three heuristics for generating fuzzy decision trees could be found from the existing references. The first is called Fuzzy ID3 which was initially proposed by Quinlan in the crisp case [15] and was developed by many researchers. (e.g., [7], [8], [18], [21]). The heuristic information used in fuzzy ID3 is the minimal fuzzy entropy. The second was suggested by Yuan and Shaw [27]. This heuristic is distinct from fuzzy ID3 heuristic and uses the minimal ambiguity (nonspecificity) of a possibility distribution to select expanded attributes. Recently, the authors have proposed the third heuristic [26], which uses the maximum classification importance of attribute contributing to its consequent to select the expanded attributes and is specified in the comparative process of following sections. Despite these heuristics for generating fuzzy decision trees, little is known regarding their comparative strengths and weaknesses. In this paper, three heuristics are analyzed and compared. The comparisons are two fold. One is the analytic comparison based on expanded attribute selection and reasoning mechanism. The other is the experimental comparison based on the size of trees and learning accuracy. The purpose of this comparative study is to explore the strengths and the weaknesses of each of the three heuristics and to show some useful guidelines on how to choose an appropriate heuristic for a particular problem.

Throughout this paper, X represents a discrete universe of discourse, $F(X)$ denotes the set of all fuzzy subsets defined on X. For $X = \{e_1, e_2, \cdots, e_N\}$ and $A \in F(X)$, A sometimes is represented as $A = A(e_1)/e_1 + \cdots + A(e_N)/e_N$, and the concept of cardinality of a fuzzy set is used. It is well known that the cardinality of a fuzzy set is a generalization of the number of elements in a crisp set. Although there are arguments and questions on the cardinality of finite fuzzy sets (e.g., [22]), this paper selects a common form of the cardinality of a fuzzy set, i.e., $M(A) = \Sigma_{i=1}^{N} A(e_i)$. When the crisp case is considered, $M(A)$ is the number of elements of a crisp set.

II. Heuristic Algorithms for Fuzzy Decision Tree Generation

In this section, a formal definition for a fuzzy decision tree is given. The heuristic algorithm for generating a fuzzy decision tree, which has two components (a criterion of expanded attribute selection and an approximate reasoning mechanism), is discussed. The generic procedure of fuzzy decision tree generation is outlined.

A. Learning From Examples and Fuzzy Decision Tree

We first formulate a problem of learning from examples with fuzzy representations. Consider a set of examples $\{e_1, e_2, \cdots, e_N\}$ which is defined as the universe of discourse X. (In short, X is denoted by $\{1, 2, \cdots, N\}$). Let $A^{(1)}, \cdots, A^{(n)}$ and $A^{(n+1)}$ be a set of fuzzy attributes where $A^{(n+1)}$ denotes a classification attribute. Each

Manuscript received September 29, 1999; revised April 16, 2000 and October 21, 2000. This work was supported by a Research Fellowship Grant G-YY12 from the Hong Kong Polytechnic University. The work of X.-Z. Wang was supported in part by Natural Science Foundation of Hebei Province Grant 698 139. This paper was recommended by Associate Editor L.O. Hall.

X.-Z. Wang is with the Department of Computing, Hong Kong Polytechnic University, Kowloon, Hong Kong and also with the Department of Mathematics, Hebei University, Baoding, China (csxzwang@comp.polu.edu.hk).

D. S. Yeung and E. C. C. Tsang are with the Department of Computing, Hong Kong Polytechnic University, Kowloon, Hong Kong.

Publisher Item Identifier S 1083-4419(01)02504-3.

TABLE I
SMALL TRAINING SET WITH FUZZY REPRESENTATION

No.	Outlook			Temperature			Humidity		Wind		Sports Plan		
	Sunny	Cloudy	Rain	Hot	Mild	Cool	Humid	Normal	Windy	Not_	V	S	W
1	0.9	0.1	0.0	1.0	0.0	0.0	0.8	0.2	0.4	0.6	0.0	0.8	0.2
2	0.8	0.2	0.0	0.6	0.4	0.0	0.0	1.0	0.0	1.0	1.0	0.7	0.0
3	0.0	0.7	0.3	0.8	0.2	0.0	0.1	0.9	0.2	0.8	0.3	0.6	0.1
4	0.2	0.7	0.1	0.3	0.7	0.0	0.2	0.8	0.3	0.7	0.9	0.1	0.0
5	0.0	0.1	0.9	0.7	0.3	0.0	0.5	0.5	0.5	0.5	0.0	0.0	1.0
6	0.0	0.7	0.3	0.0	0.3	0.7	0.7	0.3	0.4	0.6	0.2	0.0	0.8
7	0.0	0.3	0.7	0.0	0.0	1.0	0.0	1.0	0.1	0.9	0.0	0.0	1.0
8	0.0	1.0	0.0	0.0	0.2	0.8	0.2	0.8	0.0	1.0	0.7	0.0	0.3
9	1.0	0.0	0.0	1.0	0.0	0.0	0.6	0.4	0.7	0.3	0.2	0.8	0.0
10	0.9	0.1	0.0	0.0	0.3	0.7	0.0	1.0	0.9	0.1	0.0	0.3	0.7
11	0.7	0.3	0.0	1.0	0.0	0.0	1.0	0.0	0.2	0.8	0.4	0.7	0.0
12	0.2	0.6	0.2	0.0	1.0	0.0	0.3	0.7	0.3	0.7	0.7	0.2	0.1
13	0.9	0.1	0.0	0.2	0.8	0.0	0.1	0.9	1.0	0.0	0.0	0.0	1.0
14	0.0	0.9	0.1	0.0	0.9	0.1	0.1	0.9	0.7	0.3	0.0	0.0	1.0
15	0.0	0.0	1.0	0.0	0.0	1.0	1.0	0.0	0.8	0.2	0.0	0.0	1.0
16	1.0	0.0	0.0	0.5	0.5	0.0	0.0	1.0	0.0	1.0	0.8	0.6	0.0

fuzzy attribute $A^{(j)}$ consists of a set of linguistic terms $T(A^{(j)}) = \{T_1^{(j)}, \cdots, T_{m_j}^{(j)}\}$ $(j = 1, 2, \cdots n+1)$. All linguistic terms are defined on the same universe of discourse X. The value of the ith example e_i with respect to the jth attribute, denoted by μ_{ij}, is a fuzzy set defined on $T(A^{(j)})$ $(i = 1, \cdots, N, j = 1, 2, \cdots n+1)$. In other words, fuzzy set μ_{ij} has a form of $\mu_{ij}^{(1)}/T_1^{(j)} + \mu_{ij}^{(2)}/T_2^{(j)} + \cdots + \mu_{ij}^{(m_j)}/T_{m_j}^{(j)}$ where $\mu_{ij}^{(k)}$ denotes the corresponding membership degree $(k = 1, \cdots, m_j)$.

To illustrate these notations, we consider an example shown in Table I which describes a small training set of learning from fuzzy examples [27]. The universe of discourse is $X = \{1, 2, \cdots, 16\}$. Five fuzzy attributes and their linguistic terms are

$$
\begin{aligned}
A^{(1)} &= \text{Outlook},\ T(A^{(1)}) = \{T_1^{(1)}, T_2^{(1)}, T_3^{(1)}\} \\
&= \{\text{Sunny, Cloudy, Rain}\} \subset F(X) \\
A^{(2)} &= \text{Temperature},\ T(A^{(2)}) = \{T_1^{(2)}, T_2^{(2)}, T_3^{(2)}\} \\
&= \{\text{Hot, Mild, Cool}\} \subset F(X) \\
A^{(3)} &= \text{Humidity},\ T(A^{(3)}) = \{T_1^{(3)}, T_2^{(3)}\} \\
&= \{\text{Humid, Normal}\} \subset F(X) \\
A^{(4)} &= \text{Wind},\ T(A^{(4)}) = \{T_1^{(4)}, T_2^{(4)}\} \\
&= \{\text{Windy, Not\_Windy}\} \subset F(X) \\
A^{(5)} &= \text{Sports-plan},\ T(A^{(5)}) = \{T_1^{(5)}, T_2^{(5)}, T_3^{(5)}\} \\
&= \{\text{V, S, W}\} \subset F(X)
\end{aligned}
$$

where $A^{(5)}$ is the classification attribute and the three symbols, V, S and W, denote three sports to play on weekends; volleyball, swimming and weightlifting, respectively. Each linguistic term (corresponding to a column of Table I) is a fuzzy subset defined on X. For instance, $T_1^{(1)}$ = Sunny = 0.9/1 + 0.8/2 + 0.0/3 + $\cdots$ + 1.0/16. As to the value of the ith example e_i with respect to the jth attribute, one can easily observe such as μ_{11} = 0.9/Sunny + 0.1/Cloudy + 0.0/Rain.

According to [16], a graph $G = (V, E)$ consists of a finite, nonempty set of nodes V and a set of edges E. If the edges are ordered pairs (v, w) of nodes, then the graph is said to be directed. A path in a graph is a sequence of edges of the form $(v_1, v_2), (v_2, v_3), \cdots, (v_{n-1}, v_n)$ which is from v_1 to v_n with length $n-1$. A directed tree is a graph without cycles, satisfying the following properties.

1) There is exactly one node, called root, which no edges enter.
2) Every node except the root has exactly one entering edge.
3) There is a unique path from the root to each node.

If (v, w) is an edge in a tree, then v is called the father of w, and w is a son of v. If there is a path from v to w $(v \neq w)$ then v is a proper ancestor of w and w is a proper descendent of v. A node with no proper descendent is called a leaf.

Definition 1: Consider the above formulated problem of learning from examples. A fuzzy decision tree is defined as a directed tree with the following properties.

1) Every node except the root is a fuzzy subset which can be represented as a conjunction of several linguistic terms $\cap_{j \in J} T_{k_j}^{(j)}$ where $J \subset \{1, 2, \cdots, n\}$, $k_j \in \{1, 2, \cdots, m_j\}$ and $\cap$ denotes the intersection among fuzzy sets.
2) Every leaf corresponds to one or more linguistic terms of the classification attribute.

For a fuzzy decision tree, the root is considered a special fuzzy subset $\{1, 1, \cdots, 1\}$ defined on X. While all linguistic terms become nominal symbols, i.e., all fuzzy sets become crisp sets, Definition 1 describes a crisp decision tree. Fuzzy decision trees can be built by using different heuristic algorithms to train Table I (e.g., Figs. 1–3).

B. Heuristic Algorithm for Generating Fuzzy Decision Tree

A fuzzy decision tree with the smallest number of leaf-nodes is usually called optimal. The construction of a optimal fuzzy decision tree, however, has been proven to be NP-hard [6], [20]. (Intuitively speaking, the complexity of a NP-hard problem

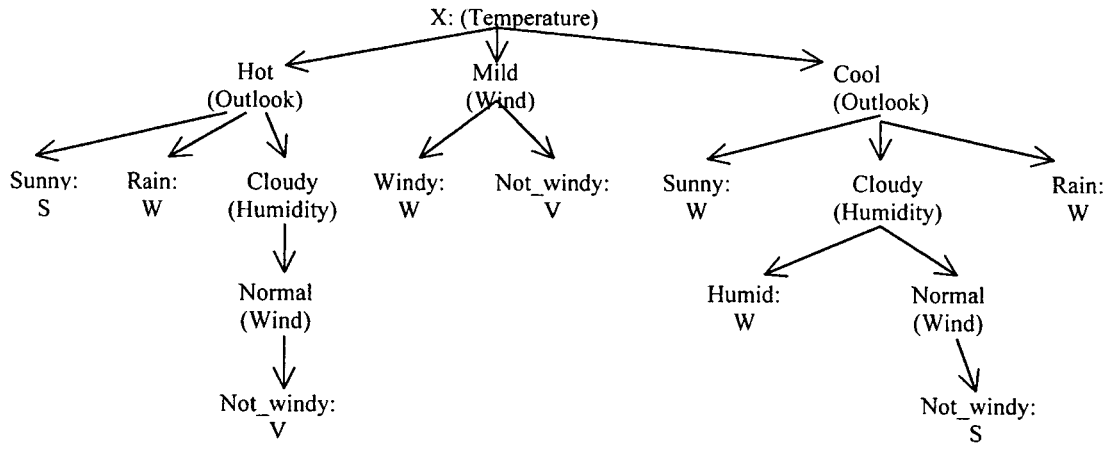

Fig. 1. Fuzzy decision tree by using fuzzy ID3 heuristic to train Table I ($\alpha = 0.4$, $\beta = 0.95$).

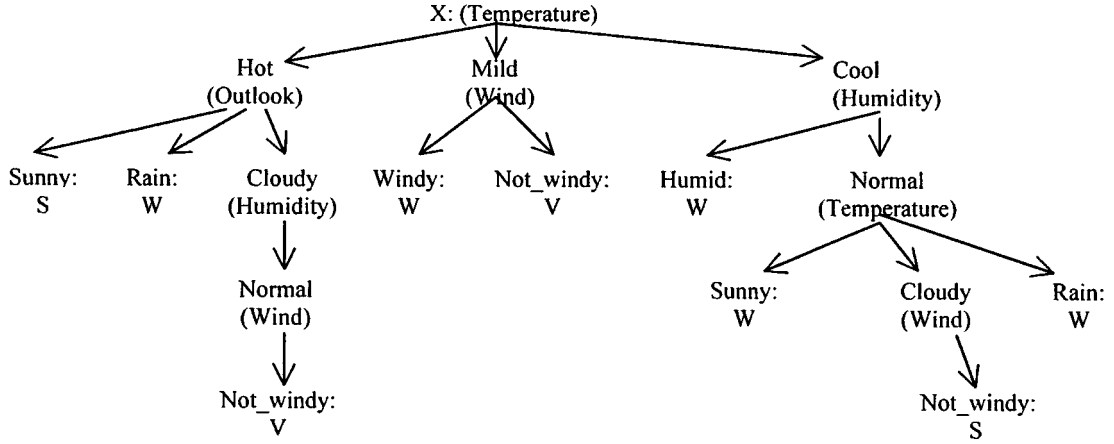

Fig. 2. Fuzzy decision tree by using Yuan and Shaw's [27] heuristic to train Table I ($\alpha = 0.4$, $\beta = 0.95$).

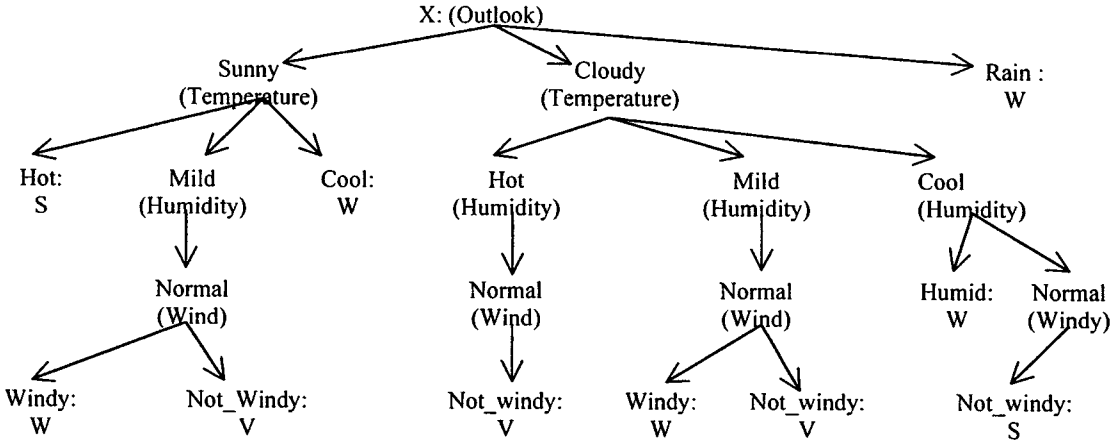

Fig. 3. Fuzzy decision tree by using our proposed heuristic to train Table I ($\alpha = 0.4$, $\beta = 0.95$).

exponentially increases with the problem-size. It is considered to be unrealistic to design an exact algorithm for a NP-hard problem at present. For more understanding on NP-hard, one can refer to [2].) Therefore, researchers had to investigate heuristic algorithms for generating relatively better trees. The heuristic algorithm in this paper contains two components, namely 1) a criterion of expanded attribute selection and 2) an approximate reasoning mechanism. For given heuristic

information and a leaf-standard, the general procedure for generating fuzzy decision trees can be described as follows.

Consider $(1, 1, \cdots, 1)_N$ as the first candidate node where N is the number of training examples.

WHILE there exist candidate nodes
DO select one using the search strategy; if the selected one is not a leaf, then generate its son-nodes according to an expanded attribute obtained by the given heuristic. These son-nodes are regarded as new candidate nodes.

Before training the initial data, the α-cut is usually used for the initial data [27]. The purpose of using α-cut is to reduce the fuzziness. The α-cut of a fuzzy set A is defined as $A_\alpha(x) = \begin{cases} A(x) & A(x) \geq \alpha \\ 0 & A(x) < \alpha \end{cases}$. Raising α can reduce the fuzziness of initial data, but the too big α may result in empty examples. Usually, α is in the interval (0, 0.5]. Generally speaking, if the training result is satisfactory then the used algorithm is considered to be able to handle the existing fuzziness and using α-cut is unnecessary. Otherwise, α-cut should be used to improve the learning performance. The selection of the value of α depends on problem–domain and users' requirements. (A more detailed explanation for α-cut could be found in [27]).

Focusing on the expanded attribute selection and reasoning mechanism, we will compare three powerful heuristics algorithms for fuzzy decision tree generation. They are heuristic **I** (the well known fuzzy ID3, e.g., [18]), heuristic **II** given in [27], and our recently proposed heuristic **III** in [26].

III. Comparison of Expanded Attribute Selection

A. Outline of Three Expanded Attribute Selection Methods

1) Heuristic I: Fuzzy ID3: One powerful heuristic for generating crisp decision trees is called ID3. The earlier version of ID3, which is based on minimum classification information-entropy to select expanded attributes, was proposed by Quinlan in 1986 [15]. As the increasing uncertainty is incorporated into the knowledge-based system, the fuzzy version of ID3 has been suggested by several authors. (e.g., [7], [8], [18], [21]). Classification information-entropy based on probabilistic models, i.e., Shannon entropy, is a well-known concept to describing probabilistic distribution's uncertainty. Subsequently, this concept was extended to describe the possibilistic distribution's uncertainty, called fuzzy entropy. A typical extension was given in [3]. Here a possibilistic distribution refers to a vector whose components are in [0,1] while a probabilistic distribution is a possibility distribution with the property that the sum of all components is equal to 1.

For a probabilistic distribution, each component is considered as the probability with which the corresponding event occurs. A possibilistic distribution is usually considered as a fuzzy set (vector), and each component of the vector, i.e., the membership degree, is regarded as the possibility with which the corresponding event occurs. For the difference and consistency between probability and possibility, on can refer to [28]. The difference between the uncertainty described by the entropy of a probabilistic distribution and described by the fuzzy entropy of a possibilistic distribution is that the former attains its maximum at all components being 0.5 but the latter does not. Fuzzy ID3 uses the fuzzy entropy of a possibilistic distribution. We briefly describe the expanded attribute selection of fuzzy ID3 as follows.

Consider a nonleaf node S having n attributes $A^{(1)}, \cdots, A^{(n)}$ to be selected. For each $k(1 \leq k \leq n)$, the attribute $A^{(k)}$ takes m_k fuzzy subsets (linguistic terms), $T_1^{(k)}, \cdots, T_{m_k}^{(k)}$. $A^{(n+1)}$ denotes the classification attribute, taking values $T_1^{(n+1)}, \cdots, T_m^{(n+1)}$.

Definition 2: For each attribute value (fuzzy subset), $T_i^{(k)}(1 \leq k \leq n, 1 \leq i \leq m_k)$, its relative frequencies concerning the jth fuzzy class $T_j^{(n+1)}(1 \leq j \leq m)$ at the considered nonleaf node S is defined as $p_{ij}^{(k)} = M(T_i^{(k)} \cap T_j^{(n+1)} \cap S)/M(T_i^{(k)} \cap S)$.

Definition 3: At the considered nonleaf node S, the fuzzy classification entropy of $T_i^{(k)}(1 \leq k \leq n, 1 \leq i \leq m_k)$ is defined as $Entr_i^{(k)} = -\Sigma_{j=1}^m p_{ij}^{(k)} \log_2 p_{ij}^{(k)}$.

Definition 4: The averaged fuzzy classification entropy of the kth attribute is defined as $E_k = \Sigma_{i=1}^{m_k} w_i Entr_i^{(k)}$ in which w_i denotes the weight of the ith value $T_i^{(k)}$ and is defined as $w_i = M(S \cap T_i^{(k)})/\Sigma_{j=1}^{m_k} M(S \cap T_j^{(k)})$.

Fuzzy ID3 heuristic aims to search for an attribute such that its averaged fuzzy classification entropy attains minimum, i.e., selecting such an integer k_0 (the k_0th attribute) that $E_{k_0} = Min_{1 \leq k \leq n} E_k$.

2) Heuristic II: Yuan and Shaw's Method: Another existing powerful heuristic is chosen from the article [27]. Instead of using minimum fuzzy entropy, this heuristic uses the minimum classification ambiguity to select expanded attributes. The classification ambiguity is called nonspecificity (also called U-uncertainty). It is the only function that satisfies all nine requirements for a possibilistic measure of uncertainty [5], [9]. Continuing to use the notations in part 1) of this section, Yuan and Shaw's heuristic is briefly formulated below.

Definition 5: At the considered nonleaf node S, the classification ambiguity of $T_i^{(k)}(1 \leq k \leq n, 1 \leq i \leq m_k)$ is defined as $Ambig_i^{(k)} = \Sigma_{j=1}^m (\pi_{ij}^{(k)} - \pi_{i,j+1}^{(k)}) \ln j$ in which $(\pi_{i1}^{(k)}, \pi_{i2}^{(k)}, \cdots, \pi_{im}^{(k)})$ with descending order $(\pi_{i,m+1}^{(k)} = 0)$ is a permutation of $(\tau_{i1}^{(k)}, \tau_{i2}^{(k)}, \cdots, \tau_{im}^{(k)})$ which is a normalization of $(p_{i1}^{(k)}, p_{i2}^{(k)}, \cdots, p_{im}^{(k)})$, i.e., $\tau_{ij}^{(k)} = p_{ij}^{(k)} / \max_j p_{ij}^{(k)}$.

Definition 6: The averaged classification ambiguity of the kth attribute is defined as $G_k = \Sigma_{i=1}^{m_k} w_i Ambig_i^{(k)}$ in which w_i is defined identically as Definition 4.

Yuan and Shaw's heuristic aims to search for an attribute such that its averaged classification ambiguity attains minimum, i.e., selecting such an integer k_0 (the k_0th attribute) that $G_{k_0} = Min_{1 \leq k \leq n} G_k$.

3) Heuristic III: Our Proposed Method: Our proposed new heuristic [26] was based on the concept of degree of importance of attribute contributing to the classification. This concept was firstly proposed by Pawlak [14] while investigating the reduction of knowledge. It was used to extract the minimum indispensable part of equivalent relations. In [26] we extended this concept to a fuzzy case and then used it to select the expanded attribute at a considered node while generating fuzzy decision trees. This paper gives a revised version of the

TABLE II
TWO TYPES OF CARS

No.	Colors	Max-speed	Size	Made in
1	White	Low	Full	Factory A
2	White	Mid	Full	Factory B
3	Black	Low	Compact	Factory A
4	Black	High	Compact	Factory B

proposed heuristic. Before describing the heuristic, we look at the intuitive background of this heuristic.

A fundamental problem we are interested in is whether or not each attribute-value has the same degree of importance contributing to the fuzzy classification. To answer this question, we consider a simple crisp case (Table II) which is obtained by observing cars manufactured by two different car factories.

Intuitively, a rule can be extracted from Table II. That is, the max-speed of cars made in Factory A is low. The second attribute max-speed seems to be more important than the other attributes. This intuitive importance of an attribute can be defined formally by the number of inconsistent examples added due to removing an attribute from the table. An example x is called inconsistent with respect to a fixed example y if all attribute-values of x are the same as the corresponding values of y but their classifications are different. It is easy to see that the number of inconsistent examples with respect to the first example in Table II is one when removing the max-speed attribute, and is zero when removing either colors or size. Therefore, according to this viewpoint, the attribute max-speed is considered to be indispensable with respect to classification and to be more important than the other attributes.

The above degree of importance of attributes contributing to the classification is generalized to the fuzzy case in the next definition.

Definition 7: Using the notations introduced at the beginning of part A in Section II. Let $\mu_{ik} = (\mu_{ik}^{(1)}, \cdots, \mu_{ik}^{(j)}, \cdots, \mu_{ik}^{(m_k)})$ be the value of the ith example with respect to the kth attribute, $\omega_{ik} = (\mu_{ik}^{(1)}, \cdots, \mu_{ik}^{(j-1)}, \mu_{ik}^{(j+1)}, \cdots, \mu_{ik}^{(m_k)})$, ν_i be the value of the ith example with respect to the classification $(1 \le i \le N, 1 \le k \le n)$, i.e., ν_i is a fuzzy set defined on $\{T_1^{(n+1)}, T_2^{(n+1)}, \cdots, T_m^{(n+1)}\}$ (the set of linguistic terms of the classification attribute $A^{(n+1)}$), SM is a selected similarity measure, $\lambda_{ip}^{(j)} = \wedge_{q \neq k}(SM(\mu_{iq}, \mu_{pq})) \wedge SM(\omega_{ik}, \omega_{pk})$ (where $\wedge$ denotes minimum, $i \neq p$), and $\sigma_{ip} = SM(\nu_i, \nu_p)$ $(i \neq p)$. Then, for the k-th attribute $A^{(k)}$ $(1 \le k \le n)$, the importance degree of its jth linguistic term $T_j^{(k)}$ $(1 \le j \le m_k)$ contributing to the classification is defined as

$$\theta_j^{(k)} = \frac{1}{N(N-1)} \sum_i \sum_{p \neq i} g^+(\lambda_{ip}^{(k)} - \sigma_{ip}) \quad \text{where}$$

$$g^+(x) = \begin{cases} x & x > 0 \\ 0 & x \le 0. \end{cases}$$

Definition 8: The averaged degree of importance of the kth attribute $A^{(k)}$ is defined as $P_k = \Sigma_{j=1}^{m_k} w_j \theta_j^{(k)}$ in which w_j is defined identically as Definitions 4 or 6.

The key points included in Definition 8 are that 1) the similarity measure between attribute-values or between classification-values replaces checking whether two attribute-values or two clusters are identical; 2) the number of inconsistent examples becomes vague; and 3) the function g^+ is used to express the inconsistent degree.

Our proposed heuristic aims to search for an attribute such that its averaged degree of importance contributing to the classification attains maximum, i.e., selecting such an integer k_0 (the k_0th attribute) that $P_{k_0} = \max_{1 \le k \le n} P_k$.

Respectively, using the three expanded attribute selection methods, we can train Table I where the cut-standard is chosen to be α = 0.4. To facilitate comparison, we purposely select β = 0.95 as the leaf-standard without any restriction although such selection results in three impractical and complicated trees Figs. 1–3.

B. Comparison of Three Expanded Attribute Selection Criteria

Proposition 1: For fixed k and i, consider two functions given in Definitions 3 and 5

$$Entr_i^{(k)} = -\sum_{j=1}^{m}; \{p_{ij}^{(k)} \log_2 p_{ij}^{(k)} \quad \text{and}$$

$$Ambig_i^{(k)} = \sum_{j=1}^{m} (\pi_{ij}^{(k)} - \pi_{i,j+1}^{(k)}) \ln j.$$

Within the area $\{0 \le p_{ij}^{(k)} \le 1 | j = 1, 2, \cdots m\}$, the first function attains its minimum at a vector of which each component is either 0 or 1, and the second attains its minimum at a vector in which one component is 1 but the other components are 0. Here we make the appointment $0 \log_2 0 = \lim_{x \to 0} (x \log_2 x) = 0$.

The proof is in the Appendix located at the end of this paper. Based on Definitions 4 and 6, this proposition indicates that fuzzy ID3 heuristic aims averagely to search the expanded attribute with relative frequencies as close to 0 or 1 as possible while Yuan and Shaw's method aims averagely to search one with relative frequencies as close to 0 (except for the maximum frequency) as possible.

It is easy to see from Proposition 1 that the minimum of the function $Ambig_i^{(k)}$ implies the minimum of the function $Entr_i^{(k)}$ and the inverse is invalid. Particularly, if $\{0 \le p_{ij}^{(k)} \le 1 | j = 1, 2, \cdots m\}$ is a probabilistic distribution then the two minima are equivalent.

Proposition 2: From Proposition 1, functions $Entr_i^{(k)}$ and $Ambig_i^{(k)}$ attain their maxima at $p_{i1}^{(k)} = \cdots = p_{im}^{(k)} = e^{-1}$ and $p_{i1}^{(k)} = \cdots = p_{im}^{(k)} = 1$, respectively.

Proposition 1 implies that when all frequencies are 1 the fuzzy entropy is 0 but the nonspecificity attains maximum. That indicates such a situation in which different expanded attributes will be selected by using fuzzy ID3 heuristic and Yuan and Shaw's heuristic. However, Proposition 1 indicates that if Yuan and Shaw's heuristic selects an expanded attribute with very small value of $Ambig_i^{(k)}$ then Fuzzy ID3 will select the same expanded attribute at the same nonleaf node. Moreover, for the frequency distribution, the smaller the nonspecificity, the closer it is to a probabilistic distribution. Thus Proposition 1 intuitively indicates that the two heuristics are likely to select the same expanded attribute while the nonspecificity is small.

Particularly, if the two heuristics select the same expanded attribute at the root with small fuzzy entropy and nonspecificity, then the two heuristics for selecting expanded attributes are gradually consistent . That is, the expanded attribute selection of fuzzy ID3 heuristic, to some extent, is identical to the one of Yuan and Shaw's heuristic. From Figs. 1 and 2, one can clearly see the process of expanded attributes of the two heuristics. The result shows that the expanded attributes of the two heuristics are the same at most nonleaf nodes.

Our proposed heuristic is based on the maximum degree of importance of attribute contributing to the fuzzy classification. It aims, on the considered node with several attributes to be chosen, to select an attribute whose contribution to classification is maximal. We have the following Proposition 3.

Proposition 3: Under an assumption of uniform distribution, either maximum or minimum degree of importance implies maximum fuzzy entropy when the classification is crisp and implies maximum nonspecificity when the classification is fuzzy. The uniform distribution assumption is formulated in the proof (see Appendix).

Here we would like to illustrate this proposition in the crisp situation. Consider a group of cases with symbolic attributes. The classification takes only two symbols, namely, "+" and "−." Let α be a value of an attribute A. Then this group of cases can be categorized into two subsets $\{A = \alpha\}$ and $\{A \neq \alpha\}$. Further, for each case $e \in \{A = \alpha\}$, a subset of cases $I_e = \{u \mid$ all attribute-values of cases e and u are the same except attribute $A\}$ can be generated. If the number of positive cases is greater than or equal to the number of negative cases then I_e is assigned to a symbol "+" (denoted by $Sign(I_e)$ = "+") else to a symbol "−" (denoted by $Sign(I_e)$ = "−"). Proposition 3 assumes that $\{Sign(I_e)|e \in \{A = \alpha\}\}$ is distributed uniformly. This assumption indicates that the number of $Sign(I_e)$ with "+" is approximately equal to the number of $Sign(I_e)$ with "−." The degree of importance of α, i.e., the number of inconsistent cases caused by deleting α, is determined by the class of $e \in \{A = \alpha\}$. Particularly, if the class is identical to $Sign(I_e)$ for each $e \in \{A = \alpha\}$ then the degree of importance of α attains minimum, and if the class is opposite to $Sign(I_e)$ for each $e \in \{A = \alpha\}$ then the degree of importance of α attains maximum. Due to the assumption of uniform distribution, either maximum or minimum degree of importance of α implies that the node $\{A = \alpha\}$ has the maximum classification entropy.

Proposition 3 indicates that the relation between our proposed heuristic and the others is very complicated. It implicitly proposes that there exists such an attribute at which the maximum (minimum resp.) degree of importance and maximum (minimum resp.) entropy can be achieved simultaneously at a node. For example, we consider the crisp case shown in Table III where there are eight examples, three attributes, and two classes.

Since the crisp classification entropy can be regarded as a special case of fuzzy case, one can directly compute the classification entropies of the three attributes, which are 0.33, 0.32, and 0.29, respectively. On the other hand, we can also compute the degree of importance of each attribute contributing to the classification. The degree of importance of the first attribute, for instance, is equal to 1 + 0 + 0 + 0 + 1 + 1 + 2 + 1 = 6 (that is the total number of examples with the same attribute-values and the different class by removing the first attribute). Similarly, one can obtain the degrees of importance of the second attribute and the third attribute, that are 4 and 0, respectively.

TABLE III
CRISP TRAINING SET

A1	A2	A3	Class
0	0	0	1
1	0	1	1
2	2	2	1
0	2	1	1
1	0	0	1
1	1	0	0
2	0	0	0
2	1	0	1

C. Complexity Caused by Expanded Attribute Selection

With regard to the complexity of the fuzzy decision tree, the relation among the three heuristics is nondeterministic, dependent mainly on the expanded attribute selection. We first consider the computation effort while expanding a nonleaf node and then consider the size of trees. From Definitions 3–8, we have the following proposition.

Proposition 4: While expanding the same nonleaf node in terms of the three heuristics, the following assertion is valid:

$$\text{CE(fuzzy ID3)} \approx \text{CE(Yuan and Shaw's method)} \leq \text{CE(our method)}$$

where CE represents the term Computation-Efforts which refers mainly to the number of times of operations such as addition, multiplication, max, min, etc.

The number of leaves is an important index to measure the size of a tree. Obviously, the bigger the number of leaves, the more the complexity. The generic standard of leaf-node is a frequency-threshold, that is, a node (fuzzy set) is regarded as a leaf if the relative frequency of some class at the node exceeds a given threshold. In the process of training Table I to generate the fuzzy decision trees Figs. 1–3, we have made use of the generic standard (without restrictions) where the threshold is set to 0.95. From Proposition 1, we have the following proposition.

Proposition 5: If a node is judged to be a leaf in Yuan and Shaw's heuristic with very small ambiguity then it is a leaf in ID3 as well.

Proposition 5 indicates intuitively that the size of the tree generated by ID3 is generally smaller than that by Yuan and Shaw's heuristic. Moreover, Proposition 3 implicitly shows that the number of leaves given by our heuristic is generally bigger than the number given by Fuzzy ID3 or Yuan and Shaw's heuristic. We do not have an exact relation among the node-numbers of these generated trees. However, Propositions 4 and 5 together with the experiments in Section V show a nonrigorous relation, that is, for the complexity, Fuzzy ID3 $\leq$ Yuan and Shaw's method $\leq$ our method.

TABLE IV
SUMMARY OF THE EMPLOYED DATABASES

Database	Domain	Source	Classes	Attributes [b]	Examples
Rice Taste	Food	[14]	2[a]	5	105
Iris	Biological	[4]	3	4	150
Mango Leaves	Biological	[15]	3	18	166
Thyroid Gland	Medical	[13]	3	5	215
Pima India Diabetes	Medical	[13]	2	8	768

[a] The continuous output is separated into two categories by positive values and negative values.
[b] All attributes are numerical.

TABLE V
SUMMARY OF ANALYTIC COMPARISON

	Fuzzy ID3 heuristic	Yuan and Shaw's heuristic	Our proposed heuristic
Heuristic information	Fuzzy entropy	Non-specificity of possibility distribution	Importance of attributes contributing to classification
Criterion of expanded attribute	Minimum fuzzy entropy	Minimum non-specificity	Maximum importance degree
Expanded attributes used in the tree	Partially same as Yuan and Shaw's heuristic	Partially same as fuzzy ID3 heuristic	Not same as the others
Reasoning mechanism	Minimum-maximum operation of memberships	Multiplication-addition operation of memberships	Weighted average of similarity
Reasoning accuracy	A(as a standard)	Less than A	Greater than A
Comprehensibility of tree	Lower than B	Higher than B	B(as a standard)
Complexity	C(as a standard)	Almost same as C	Greater than C

D. Discussion

On a nonleaf node, ID3 aims on average to select an attribute such that the components of the frequency vector are as close to zero or one as possible whereas Yuan and Shaw's heuristic aims on average to select an attribute such that only one component is as close to one as possible and other components are as close to zero as possible. Propositions 2 and 3 indicate that there is a significant difference between fuzzy ID3 and Yuan and Shaw's method when training examples have much classification uncertainty (ambiguity). Moreover, Propositions 1 and 2 show that these two heuristics are gradually consistent when classification ambiguity is small, which gives us such guidelines that ID3 heuristic is better if more than one fuzzy set in the consequent part of the generated fuzzy rule is acceptable (e.g., IF A THEN B or C) and Yuan and Shaw's heuristic is better if it is unacceptable. To a great extent, the experimental results in Section V demonstrate this comparison assertion (Table VII) where more than one fuzzy set in the consequent part is unacceptable. Synthesizing the above propositions, we can reasonably consider that the main strengths of Fuzzy ID3 are the generation of a tree with small size (leaf-nodes) whereas the main strengths of Yuan and Shaw's method are the nonspecificity existing in the classification can be effectively handled.

Our method selects an attribute which has the most influence to the classification as the expanded attribute, and on average, it can arrange the degree of importance of attributes in descending order. The main strength of our method is that the algorithm can generate weighted fuzzy rules. A weighted fuzzy rule means that, for each proposition of antecedent of the rule, a weight is assigned to indicate the degree of importance of the proposition. It can be universally accepted that, in a fuzzy production rule, different propositions should have a different degree of importance contributing to the consequent. For example, in medical diagnosis systems it is common to observe that a particular symptom combined with other symptoms may lead to a possible disease. Doctors often assign a weight to each symptom in order to show the relative degree (weight) of importance of each symptom leading to the consequent (a disease), which means that for a disease one symptom is possibly more important than another symptom.

The weight assignment is an important but difficult problem. Usually the weight values are given by domain specialists in terms of their experience. Due to the domain experience, it is difficult to measure the weight values. This paper makes an initial attempt to measure the weight by a type of degree of importance. Noting that domain experts usually consider a proposition to be important only if the loss of this proposition will result in many errors, the importance of a proposition specified by domain experts is coherent with the importance of an attribute-value given in Definition 7. Thus our proposed heuristic gives a new way to roughly acquire the weights while generating the fuzzy decision tree (fuzzy rules). That is, for a linguistic term of an expanded attribute, its degree of importance (which is defined in Definition 7) is regarded as the weight of the corresponding proposition. In this way, while converting the tree into a set of rules, a number of weighted fuzzy rules are derived.

Compared with traditional fuzzy rules (FRs), weighted fuzzy rules (WFRs) have at least two advantages. One is that WFRs can enhance the representation power of FRs [23] and the other is that WFRs can improve the learning accuracy due to the use of weights as well as the operator pair (addition, multiplication), which is reflected partially in the following sections.

Moreover, an arrangement of attributes according to their degrees of importance can be given. This arrangement provides some guidelines for the robustness, that is, the loss of some attribute-values with small importance of classification does not heavily affect the prediction accuracy for novel cases. This comparison assertion is also partially reflected in the experiments of Section V (Table VI).

IV. Comparison of Reasoning Mechanism

A. Rule Forms Generated by the Three Algorithms

Definition 9: A simple fuzzy rule takes a form "IF antecedent THEN consequent (CF)" where the antecedent is the intersection of some fuzzy sets $A_1 \cap A_2 \cap \cdots \cap A_n$ and the consequent is one fuzzy set B (A_i and B are defined on the same universe of discourse), CF is the certainty factor defined as M(antecedent $\cap$ consequent)/M(antecedent). A weighted fuzzy rule [23], [24] takes a form "IF antecedent THEN consequent (CF., Lw, Th)" where the antecedent, the consequent and the certainty factor are defined as that in the simple form, Lw = (Lw_1, Lw_2, $\cdots$, Lw_n) is the local weight vector and Th = (Th_1, Th_2, $\cdots$, Th_n) is the threshold for firing the rule.

The following are the forms of fuzzy production rules with parameters extracted from the fuzzy decision trees generated by the three heuristics discussed above.

Fuzzy ID3 method: IF A_j $(j = 1,2,\cdots,n)$ THEN $(B_1,\cdots,B_m); CF., (p_1,\cdots,p_m)$.

Yuan and Shaw's method: IF A_j $(j = 1,2,\cdots,n)$ THEN $B_k; CF$.

Our method: IF A_j $(j = 1,2,\cdots,n)$ THEN $B_k; CF., (Lw_1,\cdots,Lw_n)$.

We consider the comprehensibility together with the reasoning mechanism for the three forms of fuzzy production rules. The parameter CF denotes the certainty factor. The rules generated by fuzzy ID3 have the parameter vector $(p_1,\cdots,p_m)$ for representing the classification distribution. That is, the crisp decision is not given since all components of the vector $(p_1,\cdots,p_m)$ are used in the reasoning mechanism [8], [18]. It usually makes the meaning of the consequent unclear, and therefore, lowers the comprehensibility to some extent. From our method, weighted fuzzy production rules can be extracted. The weight parameter vector $(Lw_1,\cdots,Lw_n)$ has the clear meaning, i.e., it indicates the degree of each proposition contributing the consequent B_k. Its many advantages have been demonstrated in our previous work (e.g., [23]–[25]). Noting that the fuzzy rule extracted from Yuan and Shaw's tree includes only one parameter CF, one can see that the comprehensibility of Yuan and Shaw's tree is better than that of our tree which is, in turn, better than that of fuzzy ID3 tree, that is

$$\begin{aligned}&\text{Compreh(fuzzy ID3 tree)}\\ &\quad\le \text{Compreh(Our tree)}\\ &\quad\le \text{Compreh(Yuan and Shaw's tree)}.\end{aligned}$$

Fuzzy decision tree induction is usually reported to have poor learning-accuracy [17], [20], [27]. To some extent, the inclusion of several parameters can improve the learning-accuracy [8] but simultaneously lowers the comprehensibility. Practically, one needs to make a trade-off between the accuracy and the comprehensibility.

B. Reasoning Mechanism

Inference in an ordinary decision tree is executed by matching one branch (a path) starting from the root and ending at a leaf-node to which a class is attached as the inference result. On the other hand, in a fuzzy decision tree, more than one branch must be matched approximately. The reasoning mechanism used in fuzzy ID3 (e.g., [18]) consists of the following three key points.

a) For the operation to aggregate membership values for the path of edges, the multiplication is adopted from many alternatives.
b) For the operation of total membership value of the path of edge and the certainty of the class attached to the leaf-node, the multiplication is also adopted.
c) For the operation to aggregate certainties of the same class from the different paths of edges, the addition is adopted from several alternatives.

In Yuan and Shaw's method, simple fuzzy rules can be extracted from the generated tree. Each path of branches from root to leaf can be converted to a simple fuzzy rule with the antecedent representing attributes on the passing branches from root to leaf and the consequent representing the class(es) labeled at the leaf. Because of the fuzziness, more than one fuzzy rule can be applied at the same time for one object classification. As a result, the object is classified into different classes with different degrees. The reasoning mechanism used in Yuan and Shaw's method is briefly described as follows [27].

a) Take the minimum of memberships of A_i $(i = 1,2,\cdots,n)$ (in Definition 8) of the antecedent, which is regarded as the degree of the inferred conclusion.
b) Take the maximum of the different memberships as the degree of class when two or more rules are applied to classify the object into the same class with different memberships.
c) When only one class is required, the class with highest degree is selected.

In our proposed method, weighted fuzzy rules can be extracted from the generated trees. The weight assigned to each proposition is regarded as the degree of importance of the attribute-value contributing to the classification. (The degrees of importance are defined in Definition 7 and are computed in selecting the expanded attribute). Like the case of a simple fuzzy rule, each path of branches from root to leaf can be converted to a weighted fuzzy rule. Our matching algorithm uses the weighted average of membership to classify an object into different classes with different degrees. We describe the reasoning mechanism used in our method as follows [23].

Let $C = (\lambda_1, \lambda_2, \cdots, \lambda_n)$ be an object to be classified, F be a group of weighted fuzzy production rules extracted from the tree, and there are k clusters. The initial state of C with respect to classification is set to be $(x_1, x_2, \cdots, x_k) = (0, 0, \cdots, 0)$.

Step 1) From the group F, select a rule R: IF A THEN B $[CF., Th, Lw]$ where the antecedent A is supposed to be $(A_1, A_2, \cdots, A_n)$; the consequent B to be $(b_1, b_2, \cdots, b_m)$; and the local weight Lw to be $(Lw_1, Lw_2, \cdots, Lw_n)$.

Step 2) Compute the membership degree of λ_j belonging to A_j: $SM_j = A_j(\lambda_j)$ where the membership function of the fuzzy set A_j is denoted by itself.

Step 3) Let Th = $(Th_1, Th_2, \cdots, Th_n)$ be the threshold. If for each proposition A_j the inequality $SM_j \ge$

Th_j holds, then the rules are executed. Compute the overall weighted average SM_W of similarity measures as $SM_W = \Sigma_{j\in T}\ (Lw_j/\Sigma_{i\in T}\ Lw_i)SM_j$ where $T = \{j : SM_j \geq Th_j\}$.

Step 4) Modify the matching-consequent according to one of the beforehand given modification strategies:

- 4a) more or less form: $B^* = \min\{1, B/SM_W\}$;
- 4b) membership-value reduction form: $B^* = B * SM_W$;
- 4c) keeping the consequent of the rule unchanged (no modification): $B^* = B$.

Step 5) Compute the certainty degree of B^* as $CF_{B^*} = CF * SM_W$;

Step 6) Put $x_j = \max(CF_{B^*}, x_j)$ where j is a number corresponding to the cluster of the consequent B^*.

Repeat the above six steps until each rule within the group F has been applied to the object C.

Let us now analyze and compare the mechanism used in the above three heuristic methods. Theoretically the main difference between the reasoning mechanism used in fuzzy ID3 and Yuan and Shaw's method is that the operators are (addition, multiplication) denoted by (+, •) in the former but are (max, min) in the latter. The operator pair (max, min) emphasizes the effect of maximum membership degree whereas (+, •) puts stress on the effect of averaged memberships. Usually, the reasoning accuracy for novel examples of the operator pair (+, •) is better than that of the pair (max, min), which is partially demonstrated by the experiments in Section V. It is very likely that the poor reasoning accuracy of (max, min) is due to the fact that the lost information available for classification caused by (max, min) is more than that caused by (+, •). Similar to fuzzy ID3, the operators used in our proposed reasoning mechanism are also (+, •). The reasoning mechanism used in fuzzy ID3 emphasizes the effect of averaged memberships whereas ours emphasizes the effect of weighted average of memberships. It is clear that the weighted average is a generalization of the traditional average model. Due to the inclusion of weight, the learning accuracy of our reasoning method for some learning problems is superior to that of the method used in fuzzy ID3. The experimental results of the three matching algorithms are shown in Section V.

C. Remark

The heuristic algorithm for generating a fuzzy decision tree and the reasoning mechanism are discussed separately. Two kinds of operators, namely, (+, •) and (max, min), can be employed in the reasoning algorithms and, in the sense of prediction accuracy, (+, •) is shown to be better than (max, min) experimentally. However, the comprehensibility of (max, min) is obviously better than that of (+, •). Noting that both (+, •) and (max, min) can be regarded as special cases of T-norm and S-norm (i.e., T-conorm), one can expect that the reasoning performance is further improved by the inclusion of T-norm and S-norm in the reasoning algorithm.

The results of analytic comparison from Sections III and IV are summarized in Table V, which are partially demonstrated by experiments of the following section.

V. Experiments

In this section, five selected databases are used. Experimental comparisons are made with respect to the three heuristics discussed above, based on the following issues: the number of total nodes, the number of leaf-nodes, the training accuracy, the testing accuracy and the capability of tolerating noisy data. Partial results of analytic comparisons listed in previous sections are experimentally demonstrated in this section.

A. Brief Introduction to Databases Used in Our Experiments

The five databases employed for experiments are obtained from various sources. Their features are briefly described below and summarized in Table IV.

1) Rice taste data: This database was used by Nozaki [12] to verify a simple and powerful algorithm for fuzzy rule generation. It contains 105 cases with five numerical attributes. The classification attribute is continuous. According to positive values and negative values of the classification attribute, cases are categorized into two classes in our experiments.
2) Iris data: This was the original data Fisher used to illustrate the discriminant analysis [4]. It contains 150 cases of three different kinds of flowers. Each case consists of four numerical attributes.
3) Mango leaf data: This set was used by Pal [13] to investigate the automatic feature extraction based on fuzzy techniques. It provides the information on different kinds of mango-leaf with 18 numerical attributes for 166 patterns (cases). It has three classes representing three kinds of mango.
4) Thyroid gland data [11]: This set contains 215 cases of three different kinds of thyroid grand. Each case consists of five numerical attributes.
5) Pima India diabetes data [11]: This database contains 768 cases related to the diagnosis of diabetes (268 positive and 500 negative). It has eight numerical attributes.

B. Experimental Procedures

Noting that all attributes of the selected five databases are numerical, we need to fuzzify these numerical attributes into linguistic terms. We make use of the following simple algorithm for generating triangular type of membership functions [10], where the number of linguistic terms is given in advance.

Let X be the considered data set. We intend to cluster X into k linguistic terms $T_j, j = 1, 2, \cdots, k$. For simplicity, we assume the type of membership to be triangular. An iteration algorithm for obtaining these linguistic terms is selected from [27].

In our experiments, the number of linguistic terms for each attribute of the five databases is taken to be three, the parameter α specified in Section II-B for reducing the fuzziness in training process is set to 0.35 and the leaf criterion is taken to be 0.75. The learning accuracy, computational complexity, and the robustness are used to compare the performance of the three methods. The learning accuracy includes training accuracy

TABLE VI
SUMMARY OF EXPERIMENTS FOR THREE HEURISTICS: I—FUZZY ID3 HEURISTIC, II—YUAN AND SHAW'S HEURISTIC, III—OUR PROPOSED HEURISTIC

Database	Number of nodes			Number of leaves			Training accuracy			Testing accuracy			Accuracy after removing an attribute		
	I	II	III	I	II	III	I	II	III	I	II	III	I	II	III
Rice taste	12.2	14.2	16.0	7.2	8.3	9.7	0.86	0.82	0.88	0.84	0.82	0.84	0.66	0.68	0.76
Iris	4.0	12.8	14.5	3.0	7.5	8.8	0.96	0.97	0.97	0.96	0.97	0.97	***	0.52	0.88
Mango-leaf	28.8	29.0	33.7	15.5	17.3	20.8	0.86	0.85	0.90	0.80	0.78	0.86	0.68	0.66	0.84
Thyroid Gland	11.8	12.5	15.2	5.8	7.7	7.8	0.82	0.80	0.86	0.78	0.75	0.78	0.72	0.75	0.75
Pima Indian diabetes	42.8	45.5	53.2	26.8	31.3	34.7	0.75	0.73	0.80	0.72	0.70	0.77	0.65	0.68	0.73

*** No result (since only one attribute is used).

TABLE VII
EXPERIMENTAL RESULTS FOR THREE HEURISTICS: I—FUZZY ID3 HEURISTIC, II—YUAN AND SHAW'S HEURISTIC, III—OUR PROPOSED HEURISTIC) AND TWO DATABASES (WHERE 30% OF EXAMPLES ARE CHANGED TO BELONGING TO MORE THAN ONE CLASS WITH DIFFERENT POSSIBILITY)

Database	Number of nodes			Number of leaves			Training accuracy			Testing accuracy		
	I	II	III	I	II	III	I	II	III	I	II	III
Mango-leaf	32.8	28.6	35.0	18.3	16.5	22.2	0.80	0.91	0.82	0.75	0.86	0.78
Thyroid Gland	16.4	18.2	18.8	8.5	9.7	10.0	0.78	0.88	0.79	0.76	0.85	0.75

and testing accuracy, the complexity is regarded as the numbers of nodes and leaves, and the robustness refers to the prediction accuracy by dropping an attribute (with smallest importance to classification) from the extracted rules. For each considered database, 50% of cases are uniformly and randomly chosen as the training set and the remaining 50% of cases are held for testing. This procedure is repeated six times for the given cut standard $= 0.35$ and the leaf standard $= 0.75$. The number of nodes, the number of leaves, the training accuracy, and the testing accuracy are regarded as the average of the six. These targets are considered due to the following reasons.

a) The number of nodes represents the complexity degree of the generated tree (the complexity of extracted fuzzy rules), which is closely related to the time-complexity and space-complexity. The number of leaves corresponds to the number of extracted fuzzy rules. It is reasonable to argue that a simple tree is considered to be superior to a complex one [1], [20].
b) The training accuracy and the testing accuracy are the two most important factors for fuzzy decision trees. The training accuracy refers to the correctness rate of testing the training set by the extracted rules. Usually the training accuracy of the crisp learning without noise can attain 100% but the fuzzy learning cannot. The testing accuracy represents the capability of predicting classes of novel examples.

The experimental results are summarized in Table VI.

An important aspect of inductive learning is the sensitivity to imperfection and imprecision in the data. One type of imperfection in our experiments is the lack of attributes. We consider the testing accuracy by deleting the attribute with minimal degree of importance contributing to classification. The result is listed in the last column of Table VI.

To check the capability of handling ambiguity of classification, we select two databases in which 30% of cases are changed to belonging to more than one class with different possibility. That means the classification distribution is no longer a probability distribution but a possibility distribution. For example, the classification for an example can become (0.60, 0.95, 0.00, 0.00) in which each component denotes the possibility of belonging to the corresponding class and the sum of all components is no longer equal to 1. (In [27], the authors have illustrated this situation clearly). The experimental results are shown in Table VII where the number of linguistic terms is taken to be four.

C. Discussions

We analyze the experimental results shown in Tables VI and VII. The classification result of an example is a distribution $(p_1, p_2, \cdots, p_m)$ where m is the number of classes. When the classification of examples is a probability distribution ($p_i \in [0,1], \Sigma p_i = 1$), the number of total nodes and the number of leaves for fuzzy ID3 heuristic are less than that of the other two heuristics. The reason is that the tree induced by fuzzy ID3 heuristic tends to generate leaves on average as early as possible. From Table VII, one can see that, when the classification of examples is a possibility distribution ($p_i \in [0,1]$), the learning performance of Yuan and Shaw's heuristic is better than the other two. It is because Yuan and Shaw's heuristic is based on the reduction of ambiguity of the possibility distribution.

Since there are a lot of real-world learning problems in which the classification is a probability distribution rather than a possibility distribution, fuzzy ID3 is applied to real-world problems more widely than Yuan and Shaw's heuristic.

Regarding the training accuracy and the testing accuracy, our method is better than fuzzy ID3 which is better than Yuan and Shaw's method. The reason may be that the weight is included in our method and the operator pair (+, $\bullet$) has the performance better than the pair (min, max) for numerical attributes. It is worth noting that the learning accuracy depends heavily on the number of linguistic terms used in the tree generation. One may notice that except for Iris data the performance on other four databases is poor. In fact, it results from the small number of linguistic terms (three) and the lower leaf-standard (0.75). Generally raising the number of linguistic terms can improve the learning accuracy, but simultaneously increases the complexity of the tree (the number of nodes). When the number of linguistic terms of each attribute for the five databases increases to four or five, the learning accuracy which is better than the results shown in Table VI can be obtained.

From the last three columns of Table VI, one can see that, for the accuracy by dropping an attribute with small importance, our method is better than the other two. That shows, to some extent, the robustness of the set of fuzzy rules generated by our

heuristic. It seems that, in the aspect of robustness, the heuristic information of degree of importance of attribute contributing to classification introduced in our method is more important than the heuristic information of fuzzy entropy or ambiguity used in the other two methods.

VI. Concluding Remarks

In this paper, three heuristic algorithms for generating fuzzy decision trees have been compared and analyzed. It is hard to say which heuristic is the best but some remarks in particular aspects can be given. For Fuzzy ID3, the main strengths are that the algorithm can generate a relatively small tree without much computation effort. For Yuan and Shaw's method, the main strengths are that the algorithm can effectively handle the nonspecificity existing in classification. For our proposed heuristic, the main strengths are that the algorithm can generate weighted fuzzy rules with high learning accuracy. Synthesizing the comparative results in previous sections, we have the following nonrigorous relationships where the notations I, II, and III represent fuzzy ID3, Yuan and Shaw's heuristic, and our proposed method, respectively.

In applicability, I $\geq$ II and III
In complexity, I $\leq$ II $\leq$ III
In comprehensibility, II $\geq$ III $\geq$ I
In learning accuracy, III $\geq$ I $\geq$ II
In handling of classification ambiguity, II $\geq$ I and III
In robustness, III $\geq$ I and II

One may choose an appropriate heuristic for a particular problem according to the above comparative strengths and weaknesses of the three.

Appendix

Proof of Proposition 1: For the first function, one can directly check that, for each $j \in \{1, 2, \cdots m\}$ and $x_j \in (0, 1]$,

$$\frac{\partial^2}{\partial x_j^2}\left(-\sum_{j=1}^{m} x_j \log_2 x_j\right) = -\frac{1}{x_j \ \ln\ 2} < 0$$

which implies the first function is convex concerning each variable on (0,1]. A convex function attains its minimum at the extremes hence the first part of Proposition 1 is valid. To prove the second part, we can consider the function $g(x_1, \cdots, x_m) = \Sigma_{j=1}^{m} \ (x_j - x_{j+1}) \ \ln \ j$ in the area $\{(x_1, \cdots, x_m) | 1 = x_1 \geq x_2 \geq \cdots \geq x_m \geq x_{m+1} = 0\}$ without losing generality. Noting that the function g can be written as $g(x_1, \cdots, x_m) = \Sigma_{j=2}^{m} \ x_j \ \ln\ (j/j+1)$, it is easy to check that in the considered area g gets its minimum only at (1,0, $\cdots$, 0), which completes the proof.

Proof of Proposition 2: The validity of this proposition can be given by solving $(\partial/\partial p_{ij}^{(k)}) Entr_i^{(k)} = 0$ and noting $Ambig_i^{(k)} = \Sigma_{j=2}^{m} \ \pi_{ij}^{(k)} \ \ln\ (j/j+1)$.

Proof of Proposition 3: Consider the learning problem formulated in the Section II-A where the classification is assumed to be crisp. Given a linguistic term $T_j^{(k)}$ of attribute $A^{(k)}$, we examine the degree of importance of $T_j^{(k)}$ and the fuzzy entropy on $T_j^{(k)}$ (as a node). The degree of importance of $T_j^{(k)}$ is determined in terms of the classification change caused by removing $T_j^{(k)}$. For each case e, a fuzzy set defined on $\{e_1, e_2, \cdots, e_N\}$ can be given by $I_e = \{\lambda_{ip}^{(j)} | p = 1, \cdots, N,$ fixed j and fixed $i\}$ in which $\lambda_{ip}^{(j)}$ is defined in Definition 7 when $i \neq p$ and is 1 when $i = p$. (In the crisp case, I_e denotes the set of cases having the same attribute values as e except $A^{(k)}$). Based on I_e, a frequency vector $(f_1, f_2, \cdots, f_m)$ can be determined by $f_i = M(C_i \cap I_e)/M(I_e)$ where Ci represents the ith class $(i = 1, \cdots, m)$. Let $f_{n_1} = \max\{f_1, \cdots, f_m\}$ and $f_{n_2} = \min\{f_1, \cdots, f_m\}$, then for each case e, a pair $(n_1(e), n_2(e))$ is given. This proposition assumes that $\{(n_1(e), n_2(e)) | \ \ all\ cases\ e\}$ is distributed uniformly. Noting that the degree of importance of $T_j^{(k)}$ is $\theta_j^{(k)} = (1/N(N-1)) \ \Sigma_i \ \Sigma_{p \neq i} \ g^+(\lambda_{ip}^{(k)} - \sigma_{ip})$, thus for each case e, its degree of importance is biggest when its class is $n_1(e)$ and is smallest when its class is $n_2(e)$. Noting the uniform distribution of $(n_1(e), n_2(e))$, the crisp classification and Proposition 2, we have the consequent that either the biggest or the smallest degree of importance of $T_j^{(k)}$ corresponds to the node $T_j^{(k)}$ which has the maximum classification entropy.

If the classification is fuzzy then the frequency vector $(f_1, f_2, \cdots, f_m)$ fails to be a probabilistic distribution but a possibilistic distribution. Similar to the above derivation, we can complete the proof by paying attention to Proposition 2.

Acknowledgment

The authors would like to thank anonymous referees for their valuable comments and suggestions on revising this paper.

References

[1] A. Blumer, A. Ehrenfeucht, D. Haussler, and M. K. Warmuth, "Occam's razor," *Inform. Process. Lett.*, vol. 24, pp. 377–380, 1987.
[2] H. P. Christos, *Computational Complexity*. Reading, MA: Addison-Wesley, 1994.
[3] A. De Luca and S. Termini, "A definition of a nonprobabilistic entropy in the setting of fuzzy set theory," *Inform. Contr.*, vol. 20, pp. 301–312, 1972.
[4] R. Fisher, "The use of multiple measurements in taxonomic problems," *Ann. Eugenics*, vol. 7, pp. 179–188, 1936.
[5] M. Higashi and G. J. Klir, "Measures on uncertainty and information based on possibility distribution," *Int. J. Gen. Syst.*, vol. 9, pp. 43–58, 1983.
[6] L. Hyafil and R. L. Rivest, "Constructing optimal binary decision trees is NP-complete," *Inform. Process. Lett.*, vol. 5, no. 1, pp. 15–17, 1976.
[7] H. Ichihashi, T. Shirai, K. Nagasaka, and T. Miyoshi, "Neuro-fuzzy ID3," *Fuzzy Sets Syst.*, vol. 81, pp. 157–167, 1996.
[8] B. Jeng, Y.-M. Jeng, and T.-P. Liang, "FILM: A fuzzy inductive learning method for automated knowledge acquisition," *Dec. Support Syst.*, vol. 21, pp. 61–73, 1997.
[9] G. J. Klir and M. Mariano, "On the uniqueness of possibilistic measure of uncertainty and information," *Fuzzy Sets Syst.*, vol. 24, pp. 197–219, 1987.
[10] T. Kohonen, *Self-Organization and Associate Memory*. New York: Springer-Verlag, 1988.
[11] C. Merz and P. Murphy. (1996) "UCI repository of machine learning databases". [Online]. Available: ftp://ftp.ics.uci.edu/pub/machine-learning-databases.
[12] K. Nozaki, H. Ishibuchi, and H. Tanaka, "A simple but powerful heuristic method for generating fuzzy rules from numerical data," *Fuzzy Sets Syst.*, vol. 86, pp. 251–270, 1997.
[13] S. K. Pal, "Fuzzy set theoretic measures for automatic feature evaluation," *Inform. Sci.*, vol. 64, pp. 165–179, 1992.
[14] Z. Pawlak, *Rough Sets: Theoretical Aspects of Reasoning about Data*. Norwell, MA: Kluwer, 1991.

[15] J. R. Quinlan, "Induction of decision trees," *Mach. Learn.*, vol. 1, pp. 81–106, 1986.
[16] S. R. Safavian and D. Landgrebe, "A survey of decision trees classifier methodology," *IEEE Trans. Syst., Man, Cybern.*, vol. 21, pp. 660–674, 1991.
[17] J. W. Shavlik, R. J. Mooney, and G. G. Towell, "Symbolic and neural learning algorithms: An experimental comparison," *Mach. Learn.*, vol. 6, no. 2, pp. 111–143, 1991.
[18] M. Umanol, H. Okamoto, I. Hatono, H. Tamura, F. Kawachi, S. Umedzu, and J. Kinoshita, "Fuzzy decision trees by fuzzy ID3 algorithm and its application to diagnosis systems," in *IEEE Int. Conf. on Fuzzy Systems*, June 26–29, 1994, pp. 2113–2118.
[19] X. Z. Wang and J. R. Hong, "On the handling of fuzziness for continuous-valued attributes in decision tree generation," *Fuzzy Sets Syst.*, vol. 99, no. 3, pp. 283–290, 1998.
[20] X. Z. Wang, B. Chen, G. Qian, and F. Ye, "On the optimization of fuzzy decision trees," *Fuzzy Sets Syst.*, vol. 112, no. 2, pp. 117–125, 2000.
[21] R. Weber, "Fuzzy-ID3: A class of methods for automatic knowledge acquisition," in *2nd International Conf. on Fuzzy Logic and Neural Networks*, Iizuka, Japan, July 17–22, 1992, pp. 265–268.
[22] M. Wygralak, "Questions of cardinality of finite fuzzy sets," *Fuzzy Sets Syst.*, vol. 102, no. 2, pp. 185–210, 1999.
[23] D. S. Yeung and E. C. C. Tsang, "Weighted fuzzy production rules," *Fuzzy Sets Syst.*, vol. 88, no. 3, pp. 299–313, 1997.
[24] ——, "A comparative study on similarity-based fuzzy reasoning methods," *IEEE Trans. Syst., Man, Cybern. B*, vol. 27, pp. 216–226, Apr. 1997.
[25] ——, "A multilevel weighted fuzzy reasoning algorithm for expert systems," *IEEE Trans. Syst., Man, Cybern. B*, vol. 28, pp. 149–158, Apr. 1998.
[26] D. S. Yeung, X. Z. Wang, and E. C. C. Tsang, "Learning weighted fuzzy rules from examples with mixed attributes by fuzzy decision trees," in *Proc. IEEE Int. Conf. on Systems, Man, and Cybernetics*, Tokyo, Japan, Oct. 12–15, 1999, pp. 349–354.
[27] Y. Yuan and M. J. Shaw, "Induction of fuzzy decision trees," *Fuzzy Sets Syst.*, vol. 69, no. 2, pp. 125–139, 1995.
[28] L. A. Zadeh, "Fuzzy sets as a basis for a theory of possibility," *Fuzzy Sets Syst.*, vol. 100, pp. 9–34, 1999.

X.-Z. Wang received the B.Sc. and M.Sc. degrees in mathematics from Heibei University, Baoding, China, in 1983 and 1992, respectively, and the Ph.D. degree in computer science from Harbin Institute of Technology, China, in 1998.

From 1983 to 1998, he worked as a Lecturer, an Associate Professor, and a Full Professor in the Department of Mathematics, Hebei University. Since 1998, he has worked as a Research Fellow at the Department of Computing, Hong Kong Polytechnic University, Kowloon. His main research interests include inductive learning and fuzzy representation, fuzzy measures and integrals, neuro-fuzzy systems and genetic algorithms, and feature extraction.

D. S. Yeung (M'89–SM'99) received the Ph.D. degree in applied mathematics from Case Western Reserve University, Cleveland, OH, in 1974.

He is the Chair Professor of the Department of Computing of the Hong Kong Polytechnic University, Kowloon. He was an Assistant Professor of mathematics and computer science at Rochester Institute of Technology, Rochester, NY, and has worked as a Research Scientist in the General Electric Corporate Research Centre, and a System Integration Engineer at TRW, Inc. His current research interests include expert-neural network hybrid systems, off-line handwritten Chinese character recognition, and fuzzy expert systems.

Dr. Yeung was the President of the IEEE Computer Society Chapter of Hong Kong from 1991 to 1992.

E. C. C. Tsang (S'93–M'93–S'93–M'98) received the B.Sc. degree in computer studies from the City University of Hong Kong (formerly the City Polytechnic of Hong Kong) in 1990 and the Ph.D. degree in computing at the Hong Kong Polytechnic University (HKPU), Kowloon, in 1996.

He is an Assistant Professor with the Department of Computing at HKPU. His main research interests are in the area of fuzzy expert systems, fuzzy Petri nets, fuzzy neural networks, machine learning, and genetic algorithms.