Information Sciences 643 (2023) 119252

Contents lists available at ScienceDirect

Information Sciences

journal homepage: www.elsevier.com/locate/ins

Z-number-valued rule-based decision trees

Yangxue Li [a], Enrique Herrera-Viedma [a,*], Gang Kou [b], Juan Antonio Morente-Molinera [a,*]

[a] Andalusian Research Institute in Data Science and Computational Intelligence, Dept. of Computer Science and AI, University of Granada, 18071 Granada, Spain

[b] School of Business Administration, Southwestern University of Finance and Economics, Chengdu 611130, PR China

ARTICLE INFO

Keywords:
Z-numbers
Rule-based system
Decision trees
Z-number-valued rule
Classification
Information gain

ABSTRACT

As a novel architecture of a fuzzy decision tree constructed on fuzzy rules, the fuzzy rule-based decision tree (FRDT) achieved better performance in terms of both classification accuracy and the size of the resulted decision tree than other classical decision trees such as C4.5, LADtree, BFtree, SimpleCart and NBTree. The concept of Z-number extends the classical fuzzy number to model both uncertain and partial reliable information. Z-numbers have significant potential in rule-based systems due to their strong representation capability. This paper designs a Z-number-valued rule-based decision tree (ZRDT) and provides the learning algorithm. Firstly, the information gain is used to replace the fuzzy confidence in FRDT to select features in each rule. Additionally, we use the negative samples to generate the second fuzzy numbers that adjust the first fuzzy numbers and improve the model's fit to the training data. The proposed ZRDT is compared with the FRDT with three different parameter values and two classical decision trees, PUBLIC and C4.5, and a decision tree ensemble method, AdaBoost.NC, in terms of classification effect and size of decision trees. Based on statistical tests, the proposed ZRDT has the highest classification performance with the smallest size for the produced decision tree.

1. Introduction

Decision trees [8,41,16] are a common classification method, which is a machine learning algorithm with simple logic. The model of the decision tree is based on a series of if-then-else rules obtained from training data. It is easy to implement, highly interpretable, fully compatible with human intuitive thinking, and able to treat large-scale data. Many decision trees have been introduced to solve classification problems, such as Iterative Dichotomies 3 (ID3) [12], Successor of ID3 (C4.5) [9], classification and regression tree (CART) [6], CHi-squared automatic interaction detector (CHAID) [33], multivariate adaptive regression spline (MARS) [29], generalized, unbiased, interaction detection and estimation (GUIDE), conditional inference trees (CTREE) [27,17], pruning and building integrated in classification decision tree (PUBLIC) [26], and so on. And many ensemble training methods based on decision trees also achieved good outcomes. For example, randomized C4.5 ensemble techniques [10], adaptive boosting negative correlation (AdaBoost.NC) learning extension with C4.5 decision tree as base classifier [34] and random feature weights (RFW) decision tree ensemble construction method [21].

* Corresponding authors.
E-mail addresses: viedma@decsai.ugr.es (E. Herrera-Viedma), jamoren@decsai.ugr.es (J.A. Morente-Molinera).

https://doi.org/10.1016/j.ins.2023.119252
Received 22 October 2022; Received in revised form 24 January 2023; Accepted 24 May 2023

Available online 30 May 2023

The classical decision tree is also considered to be a classifier that is very unstable to small changes in the training data, in other words, a method with high variance. With the elasticity of the fuzzy set formalism, fuzzy logic can improve this problem. A variant of classical decision tree inductive learning using fuzzy set theory is proposed, called fuzzy decision trees or soft decision trees [39,22]. After that, some advanced fuzzy decision trees are introduced, such as optimization of fuzzy decision trees [35], flexible fuzzy decision trees (FlexDT) [18], multi flexible fuzzy decision tree (MFlexDT) [11], Chi-square-based multi flexible fuzzy decision trees (Chi-MFlexDT) [19] and fuzzy rule-based decision trees (FRDT) [36].

The FRDT is a special architecture of fuzzy decision trees, which is an oblique tree assembled by layers. Each node is a fuzzy rule with a unique consequent class, which is written as:

IF input sample is antecedent fuzzy numbers THEN consequent class.

Each class corresponds to only one fuzzy rule in a layer in order to minimize the size of the decision tree. If a sample cannot be classified in the fuzzy rules of this layer, it will be sent to the next layer. Compared to other decision tree algorithms, the FRDT exhibits superior performance in terms of both accuracy and the size of produced trees. However, there are also two points of FRDT can be improved. First, the initial fuzzy numbers with respect to each feature are determined by the average values of classes and will not change. The antecedent fuzzy numbers of fuzzy rules for a layer are selected from initial fuzzy numbers. The intersection region between adjacent initial fuzzy numbers is equally divided. This method does not fit the training data well if it is unbalanced. Second, the features selection approach cannot select the most appropriate features. A fuzzy rule does not consider all features but uses fuzzy confidence to select a few suitable ones from them. Apparently, as the increase of number of considered features, the fuzzy confidence of the fuzzy rule will be decreased. Thus, the authors have to set a penalty coefficient to select the most suitable features. When one of the two fuzzy rules considers more features than the other, and their fuzzy confidence difference is greater than this coefficient, the fuzzy rule with the larger confidence will be used. However, fuzzy confidence has not been proved to be directly related to the performance of decision trees.

For the first point about the initial fuzzy numbers, the concept of Z-number can be used to extend the fuzzy rule. A Z-number is an ordered pair of fuzzy numbers, where the first one restricts the values of the variable, and the second one measures the reliability of the first [40]. Because of the special form of the Z-number, it can describe information with both uncertainty and partial reliability. Uncertain and partially reliable information is the most common information in the real world. Since the introduction of Z-number, it has demonstrated its powerful representative ability in many fields, such as decision making [4,23], information fusion [30], failure analysis [38,13], linguistic information processing [7,32] and game theory [14]. In the light of the definition of Z-number, the base value of the second fuzzy number is in $[0,1]$ and the larger it is, the more reliable the first one is. For classification problems, we can assume that the more reliable a fuzzy number is, the more sample space it can cover. Conversely, a less reliable fuzzy number should cover a smaller sample space. Thus, we can adjust the initial fuzzy number in the training procedure to cover more positive training samples by using the second fuzzy number. In this paper, we define the form of the Z-number-valued if-then rule written as:

IF input sample is antecedent Z-numbers THEN consequent class.

If the fuzzy rules are replaced by Z-number-valued rules, the FRDT is extended to a new decision tree, named the Z-number-valued rule-based decision tree (ZRDT). We also provide the learning algorithm of the ZRDT: Firstly, a series of initial fuzzy numbers are obtained by the average values from training data as the first fuzzy numbers of the antecedent Z-numbers. Then the second fuzzy numbers are structured by negative samples covered by the rule. If a rule does not cover any negative samples, we believe that this rule is totally reliable, and its second fuzzy numbers should be 1 (crisp numbers are a special form of fuzzy numbers). The second fuzzy number will be used to adjust the first one. If the second one is 1, then the first one is still in its original shape. If the second one is less than 1, then the sample space covered by the first one will be scaled down.

For the second point about feature selection, information gain can replace the fuzzy confidence in FRDT to select features for a rule. Information gain is a common index to measure the impurity of a node in decision tree algorithms [28,37]. In addition information gain, GINI index [31,20], gain ratio [5,15], and misclassification rate [1] are common in decision tree algorithms. Information gain is based on entropy, which is the difference between the entropy of a class and the conditional entropy of a class for the selected feature. In this paper, features are chosen so as to maximize the FOIL's information gain [24,25], which is a measure of improvement of the rule in comparison with the default rule for the target class. In this case, we can discard the penalty coefficient parameter for comparing the fuzzy confidences. The main contributions of this paper can be summarized as follows.

(1) Proposed the Z-number adjustment method to make the Z-numbers fit training data better than fuzzy numbers.
(2) Defined the structure of the Z-number-valued rule used in rule-based decision trees.
(3) Using information gain instead of fuzzy confidence for feature selection further improved the classification effect and reduced the size of decision tree.
(4) Designed the Z-number-valued rule-based decision tree (ZRDT) and provided its learning algorithm, the Z-number-valued rule extraction algorithm.

In order to evaluate the effectiveness of the proposed ZRDT method, an experiment is conducted to compare ZRDT to FRDT with different parameters and two classical decision tree methods, PUBLIC [26] and C4.5 [9], and a decision tree ensemble method, AdaBoost.NC [34]. First we compare the classification performances of each method over thirteen benchmark datasets used in [36].

Table 1
Descriptions of some notations used in this paper.

Notation	Description
$X = \{x_i\}_{i=1}^{\chi}$	The dataset X with χ samples, where x_i is the i-th sample.
$x_i = \{x_{ij}\}_{j=1}^{n}$	The i-th sample x_i with n features, where x_{ij} is the value of sample with respect to j-th feature.
$Y = \{y_w\}_{w=1}^{c}$	The set of classes in dataset.
R	The fuzzy rule.
y_R	The consequent class of rule R.
$y_{l,w}$	The w-th class of layer l.
$A_{\rho(k)}$	The k-th antecedent fuzzy number of the fuzzy rule, where $\rho(k) \in \{1, \cdots, n\}$.
$A_{jw} = (a_{jw}, b_{jw}, c_{jw}, d_{jw})$	The w-th fuzzy number for j-th feature.
s_{jw}	The mean value of class y_w on j-th feature.
$infi$	The positive infinite value.
$-infi$	The negative infinite value.
m	The number of considered features of the rule.
$m_{l,w}$	The number of considered features of rule $R_{Zl,w}$.
$FConf(R)$	The fuzzy confidence of fuzzy rule R.
$\phi_R(x)$	The average membership degree of sample x for rule R.
$\mu_{A_{\rho(k)}}(x_{i\rho(k)})$	The membership function of $x_{i\rho(k)}$ belonging to fuzzy number $A_{\rho(k)}$.
X_{y_R}	The set of samples belonging to class y_R.
X^{imp}	The additional impure node contains a set of uncertain samples.
α	The penalty coefficient parameter used to select features.
δ	The threshold parameter.
$l = 1, \cdots, L$	The layer index, where L is the number of layers.
R_Z	The Z-number-valued rule.
$R_{Zl,w}$	The Z-number-valued rule in layer l with consequent class $y_{l,w}$.
$Z_{\rho(k)} = (A_{\rho(k)}, B_{\rho(k)})$	The k-th antecedent Z-number of the rule.
$A'_{\rho(k)}$	The adjusted fuzzy number of $Z_{\rho(k)} = (A_{\rho(k)}, B_{\rho(k)})$.
$IG(R)$	The FOIL information gain of rule R.
$Z_{l,w\rho_{l,w}(k_{l,w})}$	The $k_{l,w}$-th antecedent Z-number of rule $R_{Zl,w}$.
c_l	The number of classes of layer l.
$X_{l,w}$	The set of samples belonging to class $y_{l,w}$ in layer l.

Besides, we investigate the size of generated decision trees of all methods. The experiments show that the proposed ZRDT has the highest classification performance and the smallest decision tree size.

This study is set out as follows: Section 2 recalls some background knowledge of FRDT and the definition of Z-numbers. Section 3 details the algorithms of the ZRDT, including the second fuzzy number calculation algorithm, the Z-number-valued rules extraction algorithm, and the ZRDT overall algorithm. In Section 4, an experiment using thirteen well-known benchmark datasets is performed to analyze and compare the classification performances and size of the produced decision trees of the FRDT with three different parameters and two decision tree methods, PUBLIC and C4.5, and a decision tree ensemble method, AdaBoost.NC. Finally, we concluded this study in Section 5.

2. Preliminaries: fuzzy rule-based decision trees and Z-numbers theory

In this section, we review some fundamentals of the FRDT and the Z-numbers theory. Furthermore, in order to help understand the meaning of some notations and terminologies used in this paper, we briefly describe a few in Table 1.

The FRDT is a special architecture of fuzzy decision trees. Unlike the tree structure of a traditional decision tree, FRDT is an oblique tree structure assembled by layers as shown in Fig. 1 [36]. Each layer includes one and only one pure leaf node for a class. It also includes one additional impure node used to contain the samples that are not covered by the current layer. If the impure node is not empty and different from the impure node of the previous layer, the next layer will be developed to classify the containing samples until the new impure node is empty or the same as the impure node of the previous layer. A pure leaf node of each layer corresponds to a fuzzy rule, which involves multiple features (variables) rather than a single feature considered at a node in traditional decision trees. The feature selection of a fuzzy rule is based on the index of fuzzy confidence, which measures the validity of a fuzzy rule. Some of the definitions and algorithms in FRDT are described in detail below.

A dataset X has χ samples and n features, $X = \{x_1, \cdots, x_i, \cdots, x_\chi\}$, where $x_i = \{x_{i1}, \cdots, x_{ij}, \cdots, x_{in}\}$, and x_{ij} is the value of x_i with respect to j-th feature. The dataset X contains c different classes, denoted as a set $Y = \{y_1, \cdots, y_c\}$.

Definition 1. For a input sample $x_i = \{x_{i1}, \cdots, x_{in}\}$, the formulation of a fuzzy rule R considered m features can be written as Eq. (1) [36].

$$R: \text{ IF } x_{i\rho(1)} \text{ is } A_{\rho(1)} \text{ and } \cdots \text{ and } x_{i\rho(m)} \text{ is } A_{\rho(m)} \text{ THEN } x_i \text{ belong to class } y_R, \tag{1}$$

where $\rho(1), \cdots, \rho(m)$ are elements of $\{1, \cdots, n\}$ and indicate the considered features; $A_{\rho(1)}, \cdots, A_{\rho(m)}$ are the antecedent fuzzy numbers in $\rho(1)$-th, $\cdots$, $\rho(m)$-th features respectively; $y_R \in Y$ is the consequent class of the fuzzy rule.

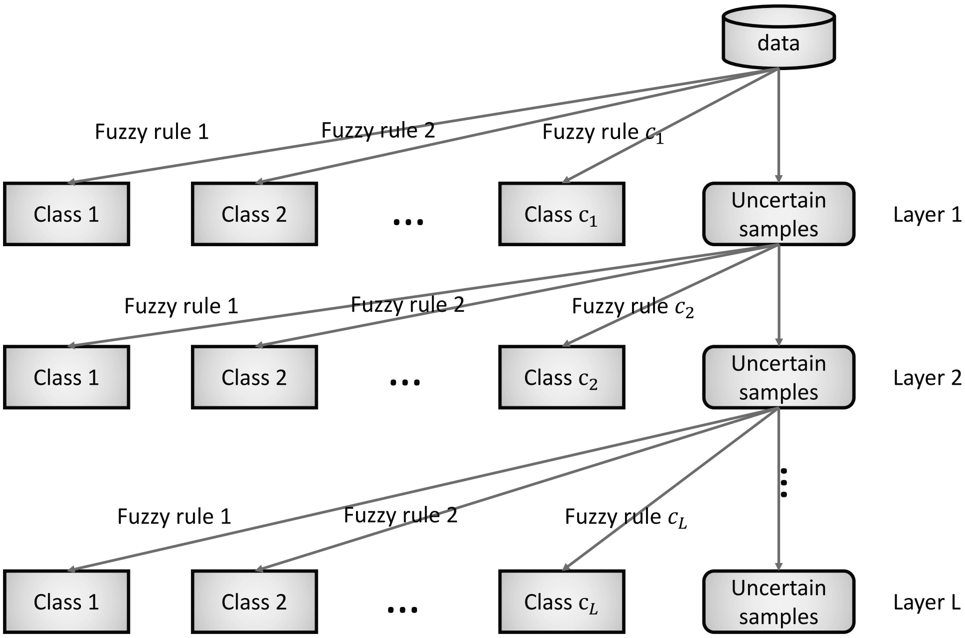

Fig. 1. The overall structure of FRDT.

Definition 2. The fuzzy confidence of fuzzy rule R is defined as Eq. (2) [36].

$$FConf(R) = \frac{\sum_{x \in X_{y_R}} \phi_R(x)}{\sum_{x \in X} \phi_R(x)}, \tag{2}$$

$$\phi_R(x) = \frac{\sum_{k=1}^{m} \mu_{A_{\rho(k)}}(x_{i\rho(k)})}{m}, \tag{3}$$

where $\phi_R(x)$ is the average membership degree of a sample x for all antecedent fuzzy numbers in rule R; $\mu_{A_{\rho(k)}}(x_{i\rho(k)})$ is the membership degree of $x_{i\rho(k)}$ belonging to fuzzy number $A_{\rho(k)}$; X_{y_R} is the set of all samples belonging to class y_R.

The FRDT also provides an algorithm to generate the initial fuzzy numbers used in fuzzy rules. For j-th feature, the number of initial fuzzy numbers equals the numbers of classes c of data to be processed. Let A_{jw} be the w-th ($w \in \{1, \cdots, c\}$) fuzzy number defined for the j-th feature. A_{jw} is a trapezoidal fuzzy number, denoted as Eq. (4) [36].

$$A_{jw} = (a_{jw}, b_{jw}, c_{jw}, d_{jw}), \tag{4}$$

where $a_{jw}, b_{jw}, c_{jw}, d_{jw}$ are real numbers. These parameters are given by the mean value of all samples belonging to class y_w about j-th feature, ($w = 1, \cdots, c$). Let the set $\{s_{j1}, \cdots, s_{jc}\}$ is the mean values set of all classes on j-th feature, where s_{jw} is the mean value of class y_w on j-th feature calculated by Eq. (5).

$$s_{jw} = \frac{\sum_{x_i \in X_{y_w}} x_{ij}}{|X_{y_w}|}. \tag{5}$$

Then the parameters of fuzzy numbers can be given by Eqs. (6)-(9).

$$a_{jw} = \begin{cases} -infi, & w = 1, \\ s_{j\sigma(w-1)}, & \text{otherwise}; \end{cases} \tag{6}$$

$$b_{jw} = \begin{cases} -infi, & w = 1, \\ s_{j\sigma(w)}, & \text{otherwise}; \end{cases} \tag{7}$$

$$c_{jw} = \begin{cases} infi, & w = c, \\ s_{j\sigma(w)}, & \text{otherwise}; \end{cases} \tag{8}$$

$$d_{jw} = \begin{cases} infi, & w = c, \\ s_{j\sigma(w+1)}, & \text{otherwise}. \end{cases} \tag{9}$$

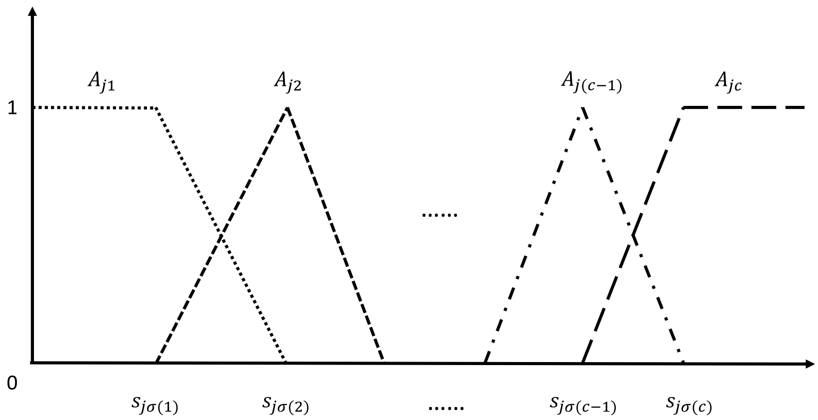

Fig. 2. The generated initial fuzzy numbers on j-th feature.

Where $\sigma(w)$ is the index function that means $s_{j\sigma(w)}$ is the w-th smallest element of $\{s_{j1}, \cdots, s_{jc}\}$; $infi$ is the positive infinite value and $-infi$ is the negative infinite value. The generated fuzzy numbers of $\{s_{j\sigma(1)}, \cdots, s_{j\sigma(c)}\}$ are portrayed in Fig. 2.

For building the layers of the tree, FRDT provides the association rules extraction algorithm (AREA) to extract only one fuzzy rule for each class [36]. The fuzzy confidence is used as a criterion to select several promising features with high fuzzy confidence value. The specific algorithm processes are shown in Algorithm 1. Where the penalty coefficient $\alpha = 0.02$ is a certain parameter.

Algorithm 1 The association rules extraction algorithm ($AREA(X, \Lambda, MaxL, \alpha)$).

Require: The c-classes training data: X;
The initial fuzzy numbers of features on X: $\Lambda = \{A_{jw}\}$, $j = 1, \cdots, n$, $w = 1, \cdots, c$;
Maximum number of features used: $MaxL$;
The penalty coefficient: α.
Ensure: The fuzzy rules of all classes: $R_1, \cdots, R_c$.
for each class y_w **do**
  $y_{R_w} \leftarrow y_w$
  $v \leftarrow \min(MaxL, n)$
  for $u = 1, 2, \cdots, \min(MaxL, n)$ **do**
    Compute r_u and r_{u+1} {r_u is the fuzzy rule that has maximum fuzzy confidence when considering u features and its consequent class is y_w.}
    if $FConf(r_u) - FConf(r_{u+1}) > \alpha$ **then**
      $v \leftarrow u$
      break
    end if
  end for
  $R_w \leftarrow r_v$
end for

After constructing the pure leaf nodes of a layer, the additional impure node contains a set of uncertain samples of that layer, denoted as X^{imp}, satisfied $\forall\ x \in X^{imp}$, $\forall\ R_w \in \{R_1, \cdots, R_c\}$, $\phi_{R_w}(x) < \delta \in [0, 1]$. To summarize, the complete algorithm of FRDT is as Algorithm 2 [36].

Definition 3. A Z-number is an ordered pair of fuzzy number,

$$Z = (A, B), \tag{10}$$

where A is the first fuzzy number with the membership function $\mu_A(u) \to [0, 1]$ and $u \in \mathcal{R}$ that constrains the values that the variable can take; B is the second fuzzy number with the membership function $\mu_B(v) \to [0, 1]$ and $v \in [0, 1]$ that measures the reliability of the first fuzzy number.

3. Z-number-valued rule based decision tree (ZRDT)

The structure of the proposed ZRDT is the same as FRDT, which is an oblique tree assembled of layers. The fuzzy rules in FRDT are replaced with Z-number-valued rules. The overall structure of ZRDT is shown in Fig. 3.

3.1. Z-numbers adjustment

In the FRDT, the fuzzy numbers used in fuzzy rules are generated by mean values. Their covered sample space is fixed. When the distribution ranges of samples from different classes are relatively different, this method causes the fuzzy numbers of classes with smaller distribution ranges to cover more negative samples.

Algorithm 2 Fuzzy rule based decision tree algorithm ($FRDT(X, MaxL, \alpha, \delta)$).

Require: The c-classes training data: X;
Maximum number of features used: $MaxL$;
The penalty coefficient: α;
The threshold: δ.
Ensure: The layers of decision tree: $\Gamma_l = \{R_{l1}, \cdots, R_{lc}, X_l^{imp}\}$, $l = 1, 2, \cdots, L$ is the l-th layer of the tree.
$l \leftarrow 0$, $X_0^{imp} \leftarrow X$
while $X_l^{imp} \neq \emptyset$ **do**
  $\Lambda \leftarrow$ the initial fuzzy numbers of features for each class in X_l^{imp}
  $R_{(l+1)1}, \cdots, R_{(l+1)c_{l+1}} \leftarrow AREA(X_l^{imp}, \Lambda, MaxL, \alpha)$
  for Each sample x in X_l^{imp} **do**
    if $\forall\ R_{(l+1)w} \in \{R_{(l+1)1}, \cdots, R_{(l+1)c_{l+1}}\}$, $\phi_{R_{(l+1)w}}(x) < \delta$ **then**
      Add x to X_{l+1}^{imp}
    end if
  end for
  if $X_{l+1}^{imp} = \emptyset$ or $X_{l+1}^{imp} = X_l^{imp}$ **then**
    break
  end if
  $l \leftarrow l + 1$
end while

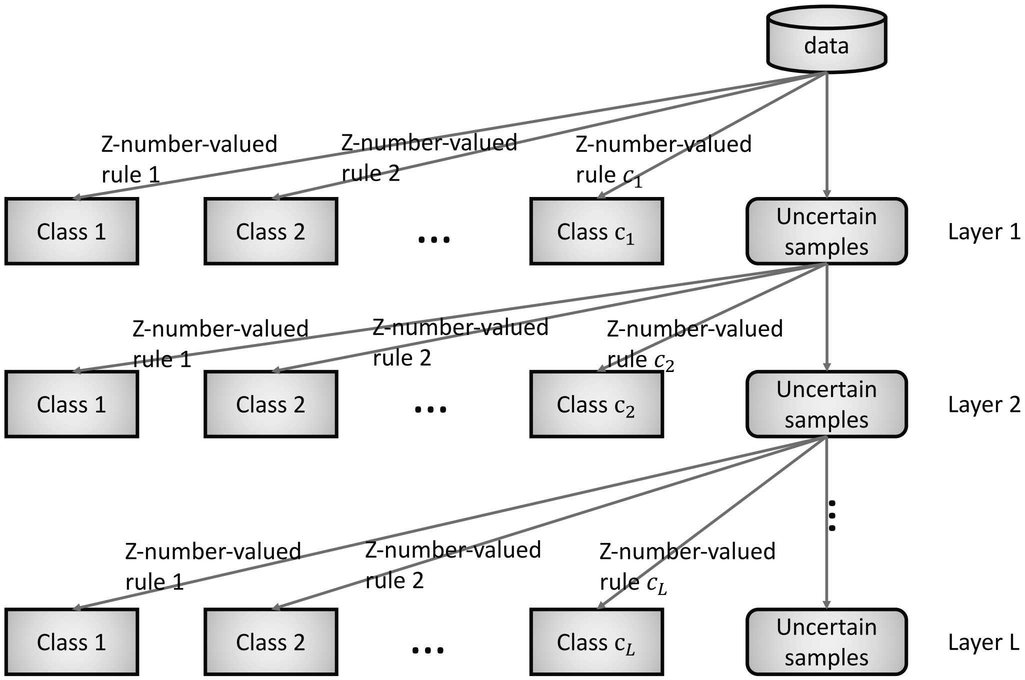

Fig. 3. The overall structure of ZRDT.

Example 1. In a two-class classification problem shown in Fig. 4, the mean values in this feature of class 1 and class 2 are 4.2292 and 8.7033, respectively. Then the fuzzy number generated by class 1 is $A_1 = (-infi, -infi, 4.2292, 8.7033)$ and the fuzzy number generated by class 2 is $A_2 = (4.2292, 8.7033, infi, infi)$. The area between the mean values of the two classes is equally divided into two fuzzy numbers. But in fact, class 1 covers a wider area than class 2. This results in two samples of class 1 being misclassified as class 2.

We use the concept of Z-number to solve this problem. For a Z-number, the second fuzzy number measures the reliability of the first fuzzy number. For a classification problem, we can say that the more reliable a fuzzy number is, the more sample space it can cover. Conversely, a less reliable fuzzy number should cover a smaller sample space. If a fuzzy number is totally reliable, then we assume that it can be in its original shape. If a fuzzy number is partially reliable, then its covered area should be scaled down. Then we can use the second fuzzy number to adjust the first fuzzy number and control its covered area.

Definition 4. For a Z-number $Z = (A, B)$, $A = (a_1, a_2, a_3, a_4)$ and $B = (b_1, b_2, b_3, b_4)$ are two trapezoidal fuzzy numbers, $a_1, a_2, a_3, a_4, b_1, b_2, b_3, b_4 \in \mathcal{R}$. After using B to adjust A the result is a fuzzy number $A' = (a'_1, a'_2, a'_3, a'_4)$, defined as Eq. (11).

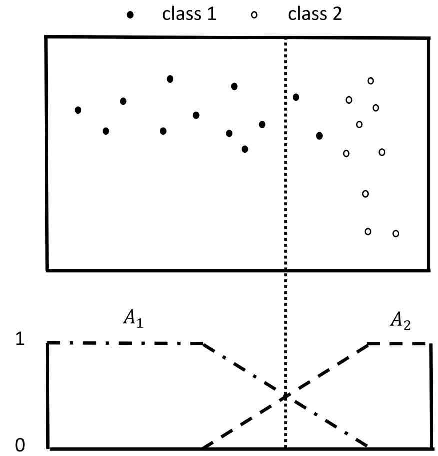

Fig. 4. A two-class classification problem with fuzzy numbers.

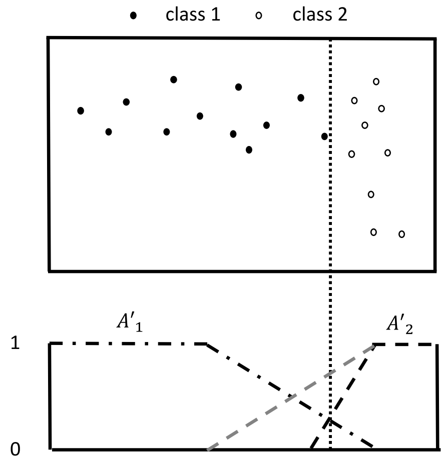

Fig. 5. A two-class classification problem with adjusted fuzzy numbers.

$$\begin{aligned} a'_1 &= a_2 - (a_2 - a_1) \times Rel(B), \\ a'_2 &= a_2, \\ a'_3 &= a_3, \\ a'_4 &= a_3 + (a_4 - a_3) \times Rel(B), \end{aligned} \tag{11}$$

where $Rel(B)$ is the reliability value of A, defined as Eq. (12).

$$\begin{aligned} Rel(B) &= \frac{y^*_B(b_3 + b_2) + (b_4 + b_1)(1 - y^*_B)}{2}, \\ y^*_B &= \frac{(b_3 - b_2)/(b_4 - b_2) + 2}{6}. \end{aligned} \tag{12}$$

When $b_1 = b_4$, $Rel(B) = b_1$; when $b_1 \neq b_2 = b_3 \neq b_4$, B is a triangular fuzzy number and $Rel(B) = (b_1 + b_2 + b_4)/3$.

Example 2. In the two-class classification problem in Example 1, we assume the reliability value of the fuzzy number about class 1 is 1 and the reliability value of the fuzzy number about class 2 is 0.4. After adjustment, the fuzzy numbers become $A'_1 = (-inf i, -inf i, 4.2292, 8.7033)$ and $A'_2 = (6.9136, 8.7033, inf i, inf i)$ shown in Fig. 5. All samples are correctly classified.

3.2. Structure of Z-number-valued if-then rule

When the fuzzy numbers of Eq. (1) are replaced by Z-numbers, the structure of Z-number-valued if-then rule can be defined as follows.

Definition 5. For a input sample $x_i = \{x_{i1}, \cdots, x_{in}\}$, the formulation of a Z-number-valued rule R_Z considered m features can be written as Eq. (13).

$$R_Z: \text{ IF } x_{i\rho(1)} \text{ is } Z_{\rho(1)} \text{ and } \cdots \text{ and } x_{i\rho(m)} \text{ is } Z_{\rho(m)} \text{ THEN } x_i \text{ belong to class } y_{R_Z}, \tag{13}$$

where $\rho(1), \cdots, \rho(m) \in \{1, 2, \cdots, n\}$ and indicate the considered features; $Z_{\rho(1)}, \cdots, Z_{\rho(m)}$ are the antecedent Z-numbers in $\rho(1)$-th, $\cdots$, $\rho(m)$-th features, respectively; $Z_{\rho(k)} = (A_{\rho(k)}, B_{\rho(k)})$, $A_{\rho(k)}$ and $B_{\rho(k)}$ are fuzzy numbers, $k = 1, 2, \cdots, m$; $y_{R_Z} \in Y$ is the consequent class of the Z-number-valued rule.

Definition 6. The average membership degree of a sample x for all antecedent Z-numbers in rule R_Z can be calculated by Eq. (14).

$$\phi_{R_Z}(x) = \frac{\sum_{k=1}^{m} \mu_{A'_{\rho(k)}}(x_{i\rho(k)})}{m}, \tag{14}$$

where $A'_{\rho(k)}$ is the fuzzy number of $A_{\rho(k)}$ adjusted by $B_{\rho(k)}$.

3.3. Z-number-valued rule extraction algorithm

The selection of features in the FRDT is based on the fuzzy confidence. A rule considered u features will only be used if its maximum fuzzy confidence value is larger than that of a fuzzy rule considered $u+1$ features by $\alpha = 0.02$. Although the authors state that α is a certain parameter, changes in the value of this parameter have an impact on the results. The determination of the value of α, 0.02, is also unsupported. In the proposed ZRDT algorithm, the FOIL information gain [24,25] is used to select features defined as follows.

Definition 7. The FOIL information gain is a measure of improvement of the rule in comparison with the default rule for the target class, and can be calculated by Eq. (15) [24,25].

$$IG(R) = p_R \times (\log_2(\frac{p_R}{p_R + n_R}) - \log_2(\frac{p}{p+n})), \tag{15}$$

where p_R and n_R are the numbers of positive and negative samples covered by rule R, respectively; p and n are the numbers of positive and negative samples covered by default rule.

The more positive samples covered by the rule, the greater the value of information gain. And the value of information gain is not related to the number of features considered in this rule. Therefore, we can discard the parameter α. That is, a rule with u features can be used if its maximum information gain is larger than that of a rule with $u+1$ features.

Before giving the complete Z-number-valued rule extraction algorithm, the second fuzzy number generation algorithm must be mentioned. For a Z-number-valued rule in the ZRDT, the first numbers of antecedent Z-numbers are calculated by Eqs. (6)-(9). The algorithm of calculating second fuzzy numbers of antecedent Z-numbers $SeFC(X, R, \delta)$ is shown as Algorithm 3.

In the algorithm $SeFC(X, R, \delta)$, N_X is the set of negative samples covered by rule R. The formula $\gamma_{\rho(k)} \leftarrow 1 + \delta - \epsilon - \mu_{A_{\rho(k)}}(x^N_{\rho(k)})$ calculates the distance of the membership degree of $A_{\rho(k)}$ about the negative sample with respect to $\rho(k)$-th feature from the threshold. The closer the value of $\mu_{A_{\rho(k)}}(x^N_{\rho(k)})$ is to δ, the closer the value of $\gamma_{\rho(k)}$ is to $1 - \epsilon$, the smaller the adjustment needed for the fuzzy number $A_{\rho(k)}$.

Example 3. In a two-class classification problem shown in Fig. 6, the circled negative sample is only a bit over the threshold. A_1 only needs to be adjusted τ_1 to A'_1 in order not to cover this sample. In contrast, the circled negative sample in Fig. 7 is much over the threshold, A_2 should be adjusted τ_2 to A'_2 in order to exclude the sample. Thus, we believe that the closer the negative samples are to the threshold, the higher the reliability of the fuzzy number.

In order to build the pure leaf node of the tree, we design the Z-number-valued rule extraction algorithm $ZREA(X, \Lambda, MaxL, \delta)$ based on information gain, which is shown as Algorithm 4. Where r_{Zu} is the Z-number-valued rule that has maximum information gain when considering u features. If $u = 1$, the Z-number-valued rule has only one antecedent Z-number. We will generate the Z-numbers corresponding to all features, and then choose the one with the highest information gain as the antecedent Z-number of this rule. If $u = 2$, we will choose another feature on the basis of r_{Z1} so that the information gain is maximized. Use this way until the increase in features does not contribute to the increase in information gain.

3.4. The overall algorithm of ZRDT

The additional impure node of each layer in ZRDT is generated in the same way as FRDT, relying on a threshold δ to collect the samples whose average membership degree for all antecedent Z-numbers in rule R_Z is below the threshold. The pseudocode of ZRDT $ZRDT(X, MaxL, \delta)$ is shown in Algorithm 5. The mission of each layer of ZRDT is to classify the uncertain samples of the previous layer, where new fuzzy numbers will be generated based on the uncertain samples, and then new Z-number-valued rules will be generated. The initial uncertain samples set is the entire training data set. Afterward, the structure of the ZRDT in Fig. 3 can

Algorithm 3 The second fuzzy number calculation algorithm ($SeFC(X, R, \delta)$).

Require: The c-classes training data: X;
The fuzzy rule: R, with antecedent fuzzy numbers: $A_{\rho(k)}$, $k = 1, \cdots, m$;
The threshold: δ.
Ensure: The Z-number-valued rule: R_Z, with antecedent Z-numbers: $Z_{\rho(k)} = (A_{\rho(k)}, B_{\rho(k)})$, $k = 1, \cdots, m$.
$N_X \leftarrow \emptyset$
for Each sample x in X **do**
  if $\phi_R(x) \geq \delta$ and $y_x \neq y_R$ **then**
    Add x to N_X
  end if
end for
for Each antecedent fuzzy number $A_{\rho(k)}$ in R **do**
  $\min_{\rho(k)} \leftarrow 1$, $mean_{\rho(k)} \leftarrow 0$, $\max_{\rho(k)} \leftarrow 0$
  for Each sample x^N in N_X **do**
    if $\mu_{A_{\rho(k)}}(x^N_{\rho(k)}) > \delta$ **then**
      $\gamma_{\rho(k)} \leftarrow 1 + \delta - \epsilon - \mu_{A_{\rho(k)}}(x^N_{\rho(k)})$ {ϵ is the infinitesimals.}
      $mean_{\rho(k)} \leftarrow mean_{\rho(k)} + \gamma_{\rho(k)}/|N_X|$
      if $\gamma_{\rho(k)} > \max_{\rho(k)}$ **then**
        $\max_{\rho(k)} \leftarrow \gamma_{\rho(k)}$
      end if
      if $\gamma_{\rho(k)} < \min_{\rho(k)}$ **then**
        $\min_{\rho(k)} \leftarrow \gamma_{\rho(k)}$
      end if
    end if
  end for
  $B_{\rho(k)} = (\min_{\rho(k)}, mean_{\rho(k)}, mean_{\rho(k)}, \max_{\rho(k)})$
end for
Compute the information gain of R and R_Z: $IG(R)$, $IG(R_Z)$.
if $IG(R) \geq IG(R_Z)$ **then**
  $B_{\rho(k)} = (1, 1, 1, 1)$ for $k = 1, \cdots, m$
  Reutrn R_Z.
else
  Return R_Z.
end if

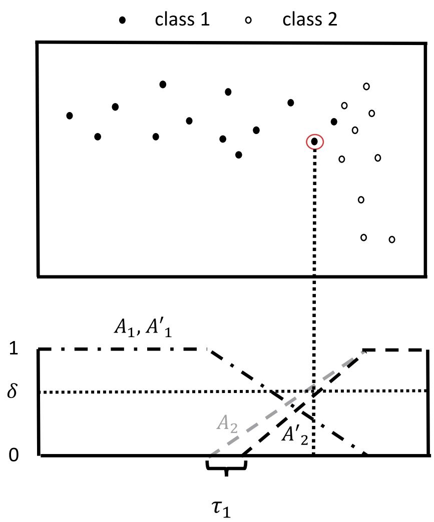

Fig. 6. A two-class classification problem with fuzzy numbers.

be visualized in Fig. 8. Where $R_{Zl,w}$, $l = 1, \cdots, L$, $w = 1, \cdots, c_l$, is the Z-number-valued rule in the l-th layer whose consequent class is $y_{l,w}$. $X_{l,w}$ is a set of samples that are classified as class $y_{l,w}$ by rule $R_{Zl,w}$.

3.5. New samples classification

After training we can obtain a model consisting of Z-number-valued rules as Fig. 9. For a new input sample $x = \{x_1, \cdots, x_n\}$, $R_{Zl,w}$, $l = 1, \cdots, L$, $w = 1, \cdots, c_l$, is the Z-number-valued rule in the l-th layer with consequent class $y_{l,w}$:

$$R_{Zl,w}: \text{ IF } x_{\rho_{l,w}(1)} \text{ is } Z_{l,w\rho_{l,w}(1)} \text{ and } \cdots \text{ and } x_{\rho_{l,w}(m_{l,w})} \text{ is } Z_{l,w\rho_{l,w}(m_{l,w})} \text{THEN } x \text{ belong to class } y_{l,w}, \tag{16}$$

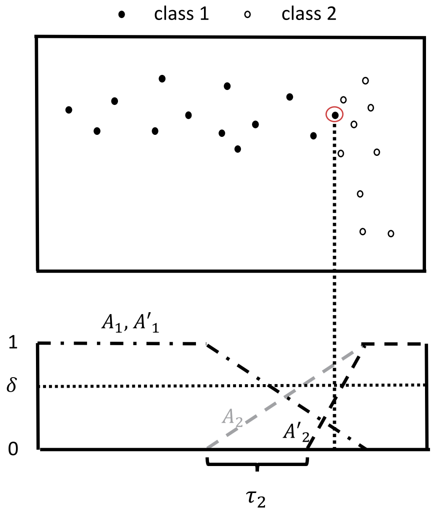

Fig. 7. A two-class classification problem with adjusted fuzzy numbers.

Algorithm 4 The Z-number-valued rules extraction algorithm ($ZREA(X, \Lambda,\ MaxL, \delta)$).

Require: The c-classes training data: X;
The initial fuzzy numbers of features on X: $\Lambda = \{A_{jw}\}$, $j = 1, \cdots, n$, $w = 1, \cdots, c$;
Maximum number of features used: $MaxL$;
The threshold: δ.
Ensure: The Z-number-valued rules of all classes: $R_{Z1}, \cdots, R_{Zc}$.

```
for each class y_w do
    y_{R_{Zw}} ← y_w
    v ← min(MaxL, n)
    for u = 1, 2, ⋯ , min(MaxL, n) do
        Compute r_{Zu} and r_{Zu+1} {r_{Zu} is the Z-number-valued rule that has maximum information gain when considering u features and its consequent class is y_w.}
        if IG(r_{Zu}) > IG(r_{Zu+1}) then
            v ← u
            break
        end if
    end for
    R_{Zw} ← r_{Zv}
end for
```

Algorithm 5 Z-number-valued rule based decision tree algorithm ($ZRDT(X, MaxL, \delta)$).

Require: The c-classes training data: X;
Maximum number of features used: $MaxL$;
The threshold: δ.
Ensure: The layers of decision tree: $\Gamma_l = \{R_{Zl,1}, \cdots, R_{Zl,c_l}, X_l^{imp}\}$, $l = 1, 2, \cdots, L$ is the l-th layer of the tree.

```
l ← 0, X_0^{imp} ← X
while X_l^{imp} ≠ ∅ do
    Λ ← the fuzzy numbers of feature for each class in X_l^{imp}
    R_{Z(l+1),1}, ⋯ , R_{Z(l+1),c_{l+1}} ← ZREA(X_l^{imp}, Λ, MaxL, δ)
    for Each sample x in X_l^{imp} do
        if ∀ R_{Z(l+1),w} ∈ {R_{Z(l+1),1}, ⋯ , R_{Z(l+1),c_{l+1}}}, φ_{R_{Z(l+1),w}}(x) < δ then
            Add x to X_{l+1}^{imp}
        end if
    end for
    if X_{l+1}^{imp} = ∅ or X_{l+1}^{imp} = X_l^{imp} then
        break
    end if
    l ← l + 1
end while
```

where $\rho_{l,w}(1), \cdots, \rho_{l,w}(m_{l,w}) \in \{1, \cdots, n\}$ and indicate the considered features in rule $R_{Zl,w}$; c_l is the number of classes of layer l; $m_{l,w}$ is the number of considered features in rule $R_{Zl,w}$; $y_{l,w} \in \{y_1, \cdots, y_c\}$. The new sample can be classified by rule R_z^*, if $\phi_{R_z^*}(x) \geq \delta$ and

$$\phi_{R_Z^*}(x) = \max(\phi_{R_{Z_{l,1}}}(x), \phi_{R_{Z_{l,2}}}(x), \cdots, \phi_{R_{Z_{l,c_l}}}(x)). \tag{17}$$

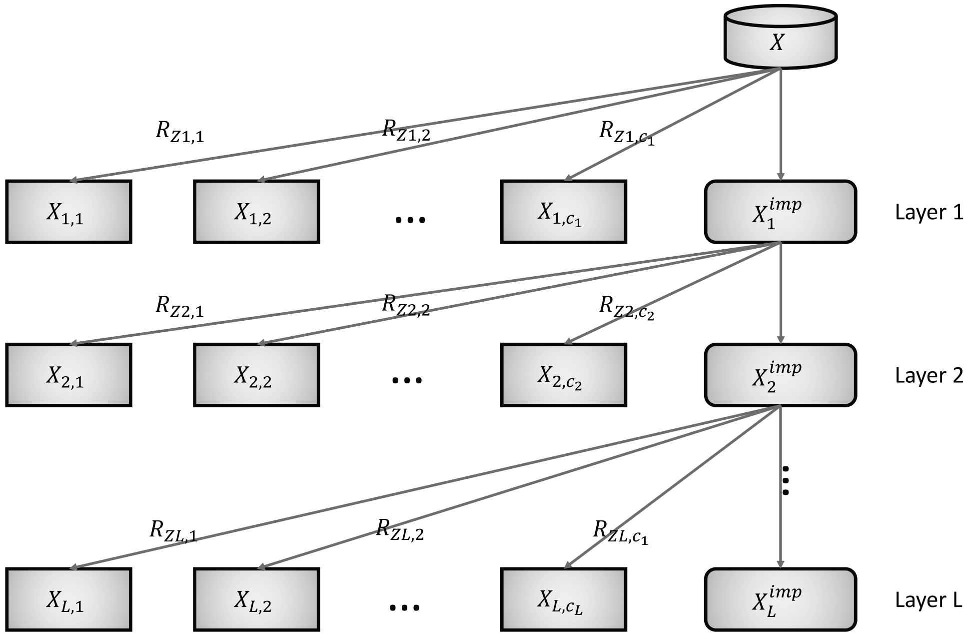

Fig. 8. The specific Z-number-valued rules classification tree.

IF $x_{\rho_{1,1}(1)}$ is $(Z_{1,1\rho_{1,1}(1)})$ and … and $x_{\rho_{1,1}(m_{1,1})}$ is $(Z_{1,1\rho_{1,1}(m_{1,1})})$ THEN x belongs to class $y_{1,1}$
ELSE IF $x_{\rho_{1,2}(1)}$ is $(Z_{1,2\rho_{1,2}(1)})$ and … and $x_{\rho_{1,2}(m_{1,2})}$ is $(Z_{1,2\rho_{1,2}(m_{1,2})})$ THEN x belongs to class $y_{1,2}$
…
ELSE IF $x_{\rho_{1,c_1}(1)}$ is $(Z_{1,c_1\rho_{1,c_1}(1)})$ and … and $x_{\rho_{1,c_1}(m_{1,c_1})}$ is $(Z_{1,c_1\rho_{1,c_1}(m_{1,c_1})})$ THEN x belongs to class y_{1,c_1}
ELSE
IF $x_{\rho_{2,1}(1)}$ is $(Z_{2,1\rho_{2,1}(1)})$ and … and $x_{\rho_{2,1}(m_{2,1})}$ is $(Z_{2,1\rho_{2,1}(m_{2,1})})$ THEN x belongs to class $y_{2,1}$
ELSE IF $x_{\rho_{2,2}(1)}$ is $(Z_{2,2\rho_{2,2}(1)})$ and … and $x_{\rho_{2,2}(m_{2,2})}$ is $(Z_{2,2\rho_{2,2}(m_{2,2})})$ THEN x belongs to class $y_{2,1}$
…
ELSE IF $x_{\rho_{2,c_2}(1)}$ is $(Z_{2,c_2\rho_{2,c_2}(1)})$ and … and $x_{\rho_{2,c_2}(m_{2,c_2})}$ is $(Z_{2,c_2\rho_{2,c_2}(m_{2,c_2})})$ THEN x belongs to class y_{2,c_2}
ELSE
…
ELSE
IF $x_{\rho_{L,1}(1)}$ is $(Z_{L,1\rho_{L,1}(1)})$ and … and $x_{\rho_{L,1}(m_{L,1})}$ is $(Z_{L,1\rho_{L,1}(m_{L,1})})$ THEN x belongs to class $y_{L,1}$
ELSE IF $x_{\rho_{L,2}(1)}$ is $(Z_{L,2\rho_{L,2}(1)})$ and … and $x_{\rho_{L,2}(m_{L,2})}$ is $(Z_{L,2\rho_{L,2}(m_{L,2})})$ THEN x belongs to class $y_{L,2}$
….
ELSE IF $x_{\rho_{L,c_L}(1)}$ is $(Z_{L,c_L\rho_{L,c_L}(1)})$ and … and $x_{\rho_{L,c_L}(m_{L,c_L})}$ is $(Z_{L,c_L\rho_{L,c_L}(m_{L,c_L})})$ THEN x belongs to class y_{L,c_L}

Fig. 9. The obtained Z-number-valued rules from ZRDT.

The average membership degree for this new sample will be calculated under each rule starting from the first layer, $l = 1$. If the maximum average membership degree of rules in this layer is over the set threshold, the new sample will be classified as the class that is the consequent class of the rule which has the maximum average membership degree. If the maximum average membership degree of rules in this layer is below the set threshold, the new sample will be sent to the next layer, $l = l + 1$.

3.6. Time complexity

The time complexity of FRDT is $O(MaxL * c * \chi * 10)$ analyzed in [36], where 10 is the maximum number of layers. For the proposed ZRDT, the maximum length of fuzzy rules for each class is $MaxL$. We need to iterate all samples to compute the second fuzzy numbers and information gain values for all fuzzy rules. Then the time complexity of each layer is $O(MaxL * c * \chi^2)$. And like FRDT, the maximum number of layers is 10. Therefore, the time complexity of ZRDT is greater than FRDT, which is $O(10 * MaxL * c * \chi^2)$.

4. Experimental studies

In this section, an experiment is used to evaluate the classification performance and the structural advantage of the proposed ZRDT. The FRDT is an effective fuzzy rule-based model for supervised classification problems, but it needs to preset two parameters: the penalty coefficient: α and the threshold value: δ. The proposed ZRDT uses information gain to reduce the parameter α. The

Table 2
The description of used datasets.

Dataset	#Samples	#Features	#Classes
authorship	841	70	4
column_3C	310	6	3
credit	690	13	2
haberman	306	3	2
heart	270	13	2
iris	150	4	3
new-thyroid	215	5	3
waveform1	5000	21	3
waveform2	5000	40	3
wdbc	569	30	2
wine	178	12	3
wisconsin	699	9	2
wodc	699	9	2

experiment focus on two aspects: the improvement of the proposed ZRDT in terms of classification effect and the size of the decision tree.

4.1. Experimental setup

Thirteen well-known benchmark datasets from the UCI repository are used to evaluate the performance of the ZRDT. Most of these datasets are the same as datasets used in [36], except for the authorship dataset. These datasets differ in the number of classes, features, and samples. The details of these datasets are described in Table 2, where '#Samples' is the number of samples, '#Features' is the number of features and '#Classes' is the number of classes.

In order to evaluate generalization ability, 10-fold cross-validation (10CV) is used. Samples are divided into 10 folds, each containing 10% samples. The model is trained with nine folds and is tested with another. The average accuracy of 10 folds can provide insight into the model's performance. There are seven algorithms used to compare their classification effect and the size of the produced decision trees (all algorithms are implemented on software KEEL [2]):

(1) FRDT (0.01): The FRDT with parameters $\alpha = 0.01$ and $\delta = 0.6$. The value of α is smaller than that recommended in [36].
(2) FRDT (0.02): The FRDT with parameters $\alpha = 0.02$ and $\delta = 0.6$. The value of α is recommended in [36].
(3) FRDT (0.03): The FRDT with parameters $\alpha = 0.03$ and $\delta = 0.6$. The value of α is larger than that recommended in [36].
(4) PUBLIC: PrUning and BuiLding Integrated in Classification decision tree with default parameters in KEEL software [26].
(5) C4.5: C4.5 decision tree with default parameters in KEEL software [3].
(6) AdaBoost.NC: Adaptive Boosting Negative Correlation learning extension with C4.5 decision tree as base classifier with default parameters in KEEL software [34].
(7) ZRDT: The proposed ZRDT with parameter $\delta = 0.6$. The Z-numbers make the model fit training data better and the information gain can find the most suitable features.

4.2. Performance comparison

The classification accuracy rates of seven algorithms over testing data are listed in Table 3. The results in bold show that the current method is the most superior. In most datasets, the ZRDT outperforms other algorithms except for the credit, heart, iris, wine, wisconsin and wodc datasets. To analyze the differences between the classifiers more statistically, the Holm test is used to compare a classifier with the superior performance classifier. The performance of two classifiers is significantly different if the hypothesis is rejected.

The Holm test is a nonparametric test which is based on the relative performance of classifiers in terms of their ranks: The better the performance, the higher the rank (in case of ties, average ranks are assigned); see Table 4. Let K be the number of classifiers and N be the number of datasets. Let r_k^n be the rank of classifier k on dataset n, $k = 1, 2, \cdots, K$ and $n = 1, 2, \cdots, N$. The test z-value for compared classifier k is calculated by:

$$z = \frac{aveR_k - \min(aveR_1, aveR_2, \cdots, aveR_K)}{\sqrt{K(K+1)/(6N)}}, \tag{18}$$

where $aveR_k = \frac{\sum_{n=1}^{N} r_k^n}{N}$ is the average rank of classifier j; $\min(aveR_1, aveR_2, \cdots, aveR_K)$ is the average rank of the highest performance classifier; $k \neq argmin_i(aveR_i)$. The z-values are used to find the corresponding probability values (p-values) of all classifiers: $p_1, p_2, \cdots, p_{K-1}$. Then we store the p-values in order lowest-to-highest: $p_{\rho(1)}, p_{\rho(2)}, \cdots, p_{\rho(K-1)}$, where $\rho(k)$ is the index function satisfying $p_{\rho(k-1)} \leq p_{\rho(k)} \leq p_{\rho(k+1)}$. If $p_{\rho(k)} < \frac{\alpha}{K-k}$, reject the null hypothesis that the performance of classifier $\rho(k)$ and the best performance classifier are not significantly different and continue to the next step, otherwise end.

Table 3
Classification accuracy rate (%) of seven algorithms over testing data.

Dataset	FRDT (0.01)	FRDT (0.02)	FRDT (0.03)	PUBLIC	C4.5	AdaBoost.NC	ZRDT
authorship	87.15%	87.99%	88.48%	92.39%	93.58%	70.87%	**95.02%**
column_3C	80.65%	81.29%	81.29%	79.68%	81.61%	58.71%	**82.26%**
credit	85.36%	**86.09%**	**86.09%**	83.48%	84.06%	85.22%	85.65%
haberman	**73.47%**	73.15%	70.84%	72.86%	73.16%	72.48%	**73.47%**
heart	73.70%	83.70%	**84.81%**	74.44%	75.19%	80.74%	81.48%
iris	**97.33%**	94.67%	94.67%	92.00%	96.00%	66.00%	95.33%
new-thyroid	91.17%	93.53%	93.51%	90.22%	93.98%	81.88%	**94.42%**
waveform1	78.50%	77.30%	78.20%	76.38%	77.24%	55.48%	**79.90%**
waveform2	73.36%	73.36%	73.36%	75.96%	74.94%	55.86%	**79.38%**
wdbc	94.02%	94.89%	94.89%	93.32%	92.25%	95.25%	**95.60%**
wine	94.93%	**96.60%**	93.24%	92.65%	94.90%	71.37%	96.08%
wisconsin	93.56%	94.13%	94.56%	95.56%	94.85%	**97.68%**	95.85%
wodc	93.28%	93.71%	93.56%	95.71%	93.28%	**96.61%**	96.57%
average	85.88%	86.95%	86.73%	85.74%	86.54%	76.01%	**88.54%**

Table 4
Ranks of seven algorithms.

Dataset	FRDT (0.01)	FRDT (0.02)	FRDT (0.03)	PUBLIC	C4.5	AdaBoost.NC	ZRDT
authorship	6	5	4	3	2	7	1
column_3C	5	3.5	3.5	6	7	2	1
credit	4	1.5	1.5	7	6	5	3
haberman	1.5	4	7	5	3	6	1.5
heart	7	2	1	6	5	4	3
iris	1	4.5	4.5	6	2	7	3
new-thyroid	5	3	4	6	2	7	1
waveform1	2	4	3	6	5	7	1
waveform2	5	5	5	2	3	7	1
wdbc	5	3.5	3.5	6	7	2	1
wine	3	1	5	6	4	7	2
wisconsin	7	6	5	3	4	1	2
wodc	6	4	5	3	7	1	2
average	4.42	3.62	4	5	4.38	4.85	1.73

Table 5
The Holm test on performance.

k	Classifier	z-value	p-value	critical value $\alpha/(K-k)$	Hypothesis
1	PUBLIC	$(5-1.73)/0.8473=3.8584$	0.000233436	0.0083	rejected
2	AdaBoost.NC	$(4.85-1.73)/0.8473=3.6768$	0.000462665	0.001	rejected
3	FRDT (0.01)	$(4.42-1.73)/0.8473=3.1775$	0.00256131	0.0125	rejected
4	C4.5	$(4.38-1.73)/0.8473=3.1321$	0.002955671	0.0167	rejected
5	FRDT(0.03)	$(4-1.73)/0.8473=2.6782$	0.011050379	0.025	rejected
6	FRDT(0.02)	$(3.62-1.73)/0.8473=2.2243$	0.033620997	0.05	rejected

In this study, the average ranks of classifiers are: $aveR_1=4.42$, $aveR_2=3.62$, $aveR_3=4.0$, $aveR_4=5.0$, $aveR_5=4.38$, $aveR_6=4.85$ and $aveR_7=1.73$. The number of classifiers is $K=7$ and the number of datasets is $N=13$. Then $\sqrt{K(K+1)/(6N)}=0.8473$. The certain pre-specified significance level is $\alpha=0.05$. In the Table 5, we list the z-value and p-value of each classifier. We can see that the Holm procedure rejects all hypotheses. We can conclude that the ZRDT performs significantly better than other algorithms at the significance level 0.05.

4.3. Decision tree size comparison

We summarized the average number of rules (nodes) of decision trees of 10 folds in the previous experiment in Table 6. In which, both PUBLIC and ZRDT achieved the smallest decision tree in five out of thirteen datasets. But in terms of the average values, we can see that the size of the decision trees generated by ZRDT is consistently smaller than those generated by other algorithms. For high-dimensional data, both ZRDT and FRDT can significantly reduce the number of rules. Among them, FRDT has better effect than FRDT. But for low-dimensional data, the effect of reducing decision trees is not obvious.

Table 6
The average size of decision trees.

Dataset	FRDT (0.01)	FRDT (0.02)	FRDT (0.03)	PUBLIC	C4.5	AdaBoost.NC	ZRDT
authorship	11.80	22.10	43.20	16.30	18.90	220.60	**11.40**
column_3C	9.90	10.80	11.40	**4.40**	8.80	99.40	11.40
credit	15.20	15.00	14.20	**5.00**	26.70	204.80	5.40
haberman	4.30	4.90	7.00	**1.00**	1.90	87.50	4.00
heart	9.10	11.70	12.30	6.50	17.30	96.30	**6.30**
iris	6.30	5.30	5.60	5.90	**3.70**	39.80	5.00
new-thyroid	10.40	12.20	11.70	8.00	7.30	61.40	**6.10**
waveform1	21.50	23.70	24.00	50.70	275.20	2014.50	**14.30**
waveform2	9.60	9.60	**9.50**	45.00	291.50	2452.30	14.00
wdbc	7.70	8.50	8.00	**4.50**	10.70	47.70	6.50
wine	6.80	8.20	9.40	4.50	**4.10**	48.70	6.50
wisconsin	8.40	**7.80**	**7.80**	**7.80**	11.70	87.10	8.00
wodc	8.40	8.20	8.20	8.20	24.80	81.80	**7.70**
average	9.95	11.38	13.25	12.91	54.05	426.30	**8.20**

5. Conclusions

The concept of Z-number is a more adequate mathematical form for descripting both uncertain and partially reliable real-world information. In this study, we extend the fuzzy rule-based decision tree in a Z-number-valued framework to propose the Z-number-valued rule-based decision tree called ZRDT. We pioneer the use of negative samples to generate the second fuzzy number of Z-numbers. The second fuzzy number can be used to adjust the first one to fit the training data better. We also use information gain to replace the fuzzy confidence to select features for each rule. Thus, not only can ZRDT select the most appropriate features for each rule, but it also reduces one parameter in FRDT. An experiment on thirteen datasets shows that the proposed ZRDT significantly outperforms the FRDT with three different parameters and other two classical decision trees, PUBLIC and C4.5, and a decision tree ensemble method, AdaBoost.NC, in classification performance. For the size of the produced decision tree, the decision tree generated by ZRDT also is the smallest. Therefore, the proposed ZRDT has the highest classification performance while the size of the decision tree is smaller.

CRediT authorship contribution statement

Yangxue Li: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Software, Validation, Visualization, Writing – original draft. **Enrique Herrera-Viedma:** Funding acquisition, Supervision, Writing – review & editing. **Gang Kou:** Funding acquisition, Supervision, Visualization. **Juan Antonio Morente-Molinera:** Funding acquisition, Supervision, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgements

This work was supported by the project B-TIC-590-UGR20 co-funded by the Programa Operativo FEDER 2014-2020 and the Regional Ministry of Economy, Knowledge, Enterprise and Universities (CECEU) of Andalusia, by the China Scholarship Council (CSC) (202106070037), by the project PID2019-103880RB-I00 funded by MCIN/AEI/10.13039/501100011033 and by the Andalusian government through project P20_00673.

References

[1] S. Aghaei, M.J. Azizi, P. Vayanos, Learning optimal and fair decision trees for non-discriminative decision-making, in: Proceedings of the AAAI Conference on Artificial Intelligence, 2019, pp. 1418–1426.
[2] J. Alcalá-Fdez, L. Sanchez, S. Garcia, M.J. del Jesus, S. Ventura, J.M. Garrell, J. Otero, C. Romero, J. Bacardit, V.M. Rivas, et al., Keel: a software tool to assess evolutionary algorithms for data mining problems, Soft Comput. 13 (2009) 307–318.
[3] A.A. Aldino, H. Sulistiani, Decision tree C4.5 algorithm for tuition aid grant program classification (case study: department of information system, universitas teknokrat Indonesia), J. Ilmiah Edutic, Pendidik. Inform. 7 (2020) 40–50.

[4] R.A. Aliev, B.G. Guirimov, O.H. Huseynov, R.R. Aliyev, Z-relation equation-based decision making, Expert Syst. Appl. 184 (2021) 115387.
[5] N. Bhargava, G. Sharma, R. Bhargava, M. Mathuria, Decision tree analysis on j48 algorithm for data mining, in: Proceedings of International Journal of Advanced Research in Computer Science and Software Engineering, vol. 3, 2013.
[6] C.E. Brodley, P.E. Utgoff, Multivariate decision trees, Mach. Learn. 19 (1995) 45–77.
[7] J. Chai, Y. Su, S. Lu, Linguistic Z-number preference relation for group decision making and its application in digital transformation assessment of SMEs, Expert Syst. Appl. (2022) 118749.
[8] B. Charbuty, A. Abdulazeez, Classification based on decision tree algorithm for machine learning, J. Appl. Sci. Technol. Trends 2 (2021) 20–28.
[9] I.S. Damanik, A.P. Windarto, A. Wanto, S.R. Andani, W. Saputra, et al., Decision Tree Optimization in C4.5 Algorithm Using Genetic Algorithm, Journal of Physics: Conference Series, IOP Publishing, 2019, p. 012012.
[10] T.G. Dietterich, An experimental comparison of three methods for constructing ensembles of decision trees: bagging, boosting, and randomization, Mach. Learn. 40 (2000) 139–157.
[11] A. Isazadeh, F. Mahan, W. Pedrycz, Mflexdt: multi flexible fuzzy decision tree for data stream classification, Soft Comput. 20 (2016) 3719–3733.
[12] C. Jin, L. De-Lin, M. Fen-Xiang, An improved id3 decision tree algorithm, in: 2009 4th International Conference on Computer Science & Education, IEEE, 2009, pp. 127–130.
[13] I.M. Jiskani, F. Yasli, S. Hosseini, A.U. Rehman, S. Uddin, Improved z-number based fuzzy fault tree approach to analyze health and safety risks in surface mines, Resour. Policy 76 (2022) 102591.
[14] B. Kang, G. Chhipi-Shrestha, Y. Deng, K. Hewage, R. Sadiq, Stable strategies analysis based on the utility of z-number in the evolutionary games, Appl. Math. Comput. 324 (2018) 202–217.
[15] R. Kavitha, E. Kannan, An efficient framework for heart disease classification using feature extraction and feature selection technique in data mining, in: 2016 International Conference on Emerging Trends in Engineering, Technology and Science (Icetets), IEEE, 2016, pp. 1–5.
[16] C.S. Lee, P.Y.S. Cheang, M. Moslehpour, Predictive analytics in business analytics: decision tree, Adv. Decis. Sci. 26 (2022) 1–29.
[17] T.S. Lim, W.Y. Loh, Y.S. Shih, A comparison of prediction accuracy, complexity, and training time of thirty-three old and new classification algorithms, Mach. Learn. 40 (2000) 203–228.
[18] X. Liu, W. Pedrycz, The development of fuzzy decision trees in the framework of axiomatic fuzzy set logic, Appl. Soft Comput. 7 (2007) 325–342.
[19] F. Mahan, M. Mohammadzad, S.M. Rozekhani, W. Pedrycz, Chi-mflexdt: Chi-square-based multi flexible fuzzy decision tree for data stream classification, Appl. Soft Comput. 105 (2021) 107301.
[20] K. Mathan, P.M. Kumar, P. Panchatcharam, G. Manogaran, R. Varadharajan, A novel Gini index decision tree data mining method with neural network classifiers for prediction of heart disease, Des. Autom. Embed. Syst. 22 (2018) 225–242.
[21] J. Maudes, J.J. Rodríguez, C. García-Osorio, N. García-Pedrajas, Random feature weights for decision tree ensemble construction, Inf. Fusion 13 (2012) 20–30.
[22] C. Olaru, L. Wehenkel, A complete fuzzy decision tree technique, Fuzzy Sets Syst. 138 (2003) 221–254.
[23] H.g. Peng, X.k. Wang, H.Y. Zhang, J.q. Wang, Group decision-making based on the aggregation of z-numbers with Archimedean t-norms and t-conorms, Inf. Sci. 569 (2021) 264–286.
[24] J.R. Quinlan, Learning logical definitions from relations, Mach. Learn. 5 (1990) 239–266.
[25] J.R. Quinlan, R.M. Cameron-Jones, Foil: a midterm report, in: European Conference on Machine Learning, Springer, 1993, pp. 1–20.
[26] R. Rastogi, K. Shim, Public: a decision tree classifier that integrates building and pruning, Data Min. Knowl. Discov. 4 (2000) 315–344.
[27] L. Rokach, O. Maimon, Top-down induction of decision trees classifiers-a survey, IEEE Trans. Syst. Man Cybern., Part C, Appl. Rev. 35 (2005) 476–487.
[28] K. Sanjay, A. Danti, Detection of fake opinions on online products using decision tree and information gain, in: 2019 3rd International Conference on Computing Methodologies and Communication (ICCMC), IEEE, 2019, pp. 372–375.
[29] S. Singh, P. Gupta, Comparative study ID3, cart and C4.5 decision tree algorithm: a survey, Int. J. Adv. Sci. Eng. Inf. Technol. (IJAIST) 27 (2014) 97–103.
[30] X. Song, B. Qin, F. Xiao, Fr–kde: a hybrid fuzzy rule-based information fusion method with its application in biomedical classification, Int. J. Fuzzy Syst. 23 (2021) 392–404.
[31] S. Tangirala, Evaluating the impact of Gini index and information gain on classification using decision tree classifier algorithm, Int. J. Adv. Comput. Sci. Appl. 11 (2020) 612–619.
[32] Z.p. Tian, R.x. Nie, J.q. Wang, H. Luo, L. Li, A prospect theory-based QUALIFLEX for uncertain linguistic Z-number multi-criteria decision-making with unknown weight information, J. Intell. Fuzzy Syst. 38 (2020) 1775–1787.
[33] G.K. Tso, K.K. Yau, Predicting electricity energy consumption: a comparison of regression analysis, decision tree and neural networks, Energy 32 (2007) 1761–1768.
[34] S. Wang, X. Yao, Multiclass imbalance problems: analysis and potential solutions, IEEE Trans. Syst. Man Cybern., Part B, Cybern. 42 (2012) 1119–1130.
[35] X. Wang, B. Chen, G. Qian, F. Ye, On the optimization of fuzzy decision trees, Fuzzy Sets Syst. 112 (2000) 117–125.
[36] X. Wang, X. Liu, W. Pedrycz, L. Zhang, Fuzzy rule based decision trees, Pattern Recognit. 48 (2015) 50–59.
[37] F.J. Yang, An extended idea about decision trees, in: 2019 International Conference on Computational Science and Computational Intelligence (CSCI), IEEE, 2019, pp. 349–354.
[38] S. Yousefi, M. Valipour, M. Gul, Systems failure analysis using z-number theory-based combined compromise solution and full consistency method, Appl. Soft Comput. 113 (2021) 107902.
[39] Y. Yuan, M.J. Shaw, Induction of fuzzy decision trees, Fuzzy Sets Syst. 69 (1995) 125–139.
[40] L.A. Zadeh, A note on z-numbers, Inf. Sci. 181 (2011) 2923–2932.
[41] H. Zhou, J. Zhang, Y. Zhou, X. Guo, Y. Ma, A feature selection algorithm of decision tree based on feature weight, Expert Syst. Appl. 164 (2021) 113842.