Neural Computing and Applications (2020) 32:11621–11636
https://doi.org/10.1007/s00521-019-04649-0

ORIGINAL ARTICLE

Axiomatic fuzzy set theory-based fuzzy oblique decision tree with dynamic mining fuzzy rules

Yuliang Cai[1] · Huaguang Zhang[2] · Shaoxin Sun[1] · Xianchang Wang[3] · Qiang He[4]

Received: 22 April 2019 / Accepted: 22 November 2019 / Published online: 11 December 2019

Abstract

This paper proposes a novel classification technology—fuzzy rule-based oblique decision tree (FRODT). The neighborhood rough sets-based FAST feature selection (NRS_FS_FAST) is first introduced to reduce attributes. In the axiomatic fuzzy set theory framework, the fuzzy rule extraction algorithm is then proposed to dynamically extract fuzzy rules. And these rules are regarded as the decision function during the tree construction. The FRODT is developed by expanding the unique non-leaf node in each layer of the tree, which results in a new tree structure with linguistic interpretation. Moreover, the genetic algorithm is implemented on σ to obtain the balanced results between classification accuracy and tree size. A series of comparative experiments are carried out with five classical classification algorithms (C4.5, BFT, LAD, SC and NBT), and recently proposed decision tree HHCART on 20 UCI data sets. Experiment results show that the FRODT exhibits better classification performance on accuracy and tree size than those of the rival algorithms.

Keywords Fuzzy oblique decision tree · Fuzzy rule extraction · AFS theory · Decision function

1 Introduction

Decision trees have received great attention on account of its significant potential applications, especially in statistics, machine learning and pattern recognition [1–3]. They have been widely used in classification problems due to the following three advantages: (1) the classification performance of the decision trees is close to or even outperforming other classification models, (2) the decision trees can handle different types of attributes, such as numeric and categorical, and (3) the results of decision trees are easy to be comprehended [4–6].

The decision trees grow in a top-down way, and recursively divide the training samples into segments having similar or the same outputs. Until now, there are three types of decision trees: "standard" decision trees [7–10], fuzzy decision trees [11–15], and oblique decision trees [16–22]. "Standard" decision trees are the simplest decision trees. However, they are incapable of addressing uncertainties consistent with human cognitive, such as vagueness and ambiguity. In this case, fuzzy decision trees with fuzzy uncertainty measure came into fashion [11–15]. For example, a novel fuzzy decision tree was introduced for the data mining task [11]. A new type of coherence

✉ Huaguang Zhang
zhanghuaguang@mail.neu.edu.cn

Yuliang Cai
ylcaivv@163.com

Shaoxin Sun
ssx5fd@163.com

Xianchang Wang
wxcixll@sohu.com

Qiang He
zhanhel@163.com

[1] College of Information Science and Engineering, Northeastern University, Shenyang, Liaoning, China

[2] College of Information Science and Engineering, State Key Laboratory of Synthetical Automation for Process Industries, Northeastern University, Shenyang, Liaoning, China

[3] School of Sciences, Dalian Ocean University, Dalian, Liaoning, China

[4] College of Computer Science and Engineering, Northeastern University, Shenyang, Liaoning, China

membership function based on AFS theory was established to describe the fuzzy concepts, and fuzzy rule-based classifier was proposed in [12]. Moreover, based on fuzzy set theory, the new tree construction technology for data classification problem was presented in [14]. The above decision trees [6–15] take one attribute as decision function during the tree construction. However, if these data are more properly segmented by the hyperplanes, they may lead to complex and inaccurate trees. In this event, oblique decision trees are more suitable.

Many profound results on oblique tree induction algorithms have been obtained [16–22]. For example, Erick et al. improved the decision function by combining evolutionary algorithm with genetic algorithm to increase the efficiency of tree building [17]. A new bottom-up oblique decision tree structure framework was presented in [19]. Moreover, Wickramarachchi et al proposed an effective heuristic approach to build oblique decision trees [22]. These evolutionary algorithms greatly improve the efficiency of tree building; however, they lack of semantic interpretation. Fortunately, the AFS theory-based classification methods have received extensive attention because of its significant advantage with semantic interpretation for classification results. Motivated by the above discussion, this paper combines AFS theory with decision tree technology to construct the new fuzzy oblique decision tree endowed with readable linguistic interpretation.

The structure of the FRODT is depicted in Fig. 1. All sample data are contained at the root node of the FRODT. At this node, we adopt the FREA to extract fuzzy rules and the samples that have not been classified by these rules are placed on an additional non-leaf node. At the non-leaf node, we use the FREA again to generate new fuzzy rules.

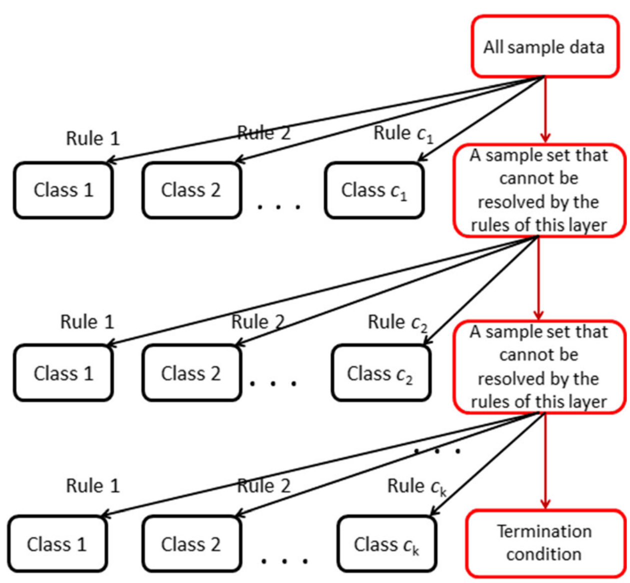

Fig. 1 The overall structure of the FRODT

And the samples that have not been classified by these new rules are also placed on an additional non-leaf node. The growth of the FRODT is repeated until the stopping condition has been satisfied.

The main contributions of the current work are summarized as four aspects:

– The Neighborhood Rough Sets-based FAST Feature Selection (NRS_FS _FAST) algorithm is introduced to reduce data redundancy and improve classification efficiency.
– A new fuzzy rule extraction algorithm (FREA) is proposed to decrease the scale of the tree.
– The AFS theory is adopted to increase semantic interpretation and decrease the human subjectivity in selecting membership functions.
– The genetic algorithm is implemented on σ to balance the results between classification accuracy and tree size.

This paper is arranged as follows: the NRS_FS_FAST algorithm is introduced in Sect. 2. In Sect. 3, the basic notions and properties of the AFS theory are provided. Section 4 describes the construction of the FRODT in detail. In Sect. 5, several comparative experiments are conducted to verify the superiority and interpretability of the FRODT. Section 6 concludes this paper.

2 The neighborhood rough sets-based FAST feature selection (NRS_FS_FAST) algorithm

2.1 The neighborhood rough set

The basic notations of the neighborhood rough set are introduced in this section, and the readers can refer to the details in [23, 24].

The data can be denoted as $NDT = \langle U, A, D \rangle$, where U is a non-empty set of samples $\{x_1, x_2, x_3, \ldots, x_n\}$, A is a condition attribute set and D is a decision attribute. $NDT = \langle U, A, D \rangle$ is called as a neighborhood decision system.

Definition 1 ([24]). Consider a neighborhood decision system $NDT = \langle U, A, D \rangle$, for $\forall x_i \in U$ and $B \subseteq A$, the neighborhood $\delta_B(x_i)$ of x_i is defined as:

$$\delta_B(x_i) = \{x_j \mid x_j \in U, \Delta_B(x_i, x_j) \le \delta\}, 1 \le i, j \le n, \qquad (1)$$

where the metric Δ is a metric function.

Definition 2 ([24]). Let $NDT = \langle U, A, D \rangle$ be a neighborhood decision system, the decision attribute D partitions the domain U into N equivalence classes with $X_1, X_2, \ldots, X_N$. For arbitrary $B \subseteq A$, the lower and upper approximations of the decision attribute D with regard to the condition attribute set B are described as:

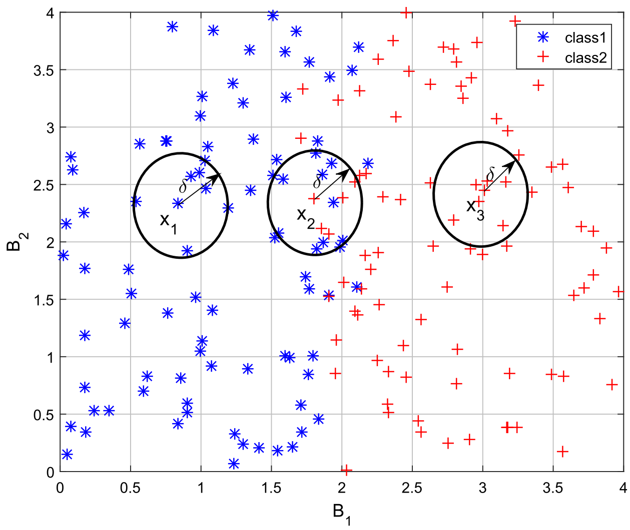

Fig. 2 The binary classification example

$$\underline{N_B}D = \bigcup_{k=1}^{N} \underline{N_B}X_k, \tag{2}$$

$$\overline{N_B}D = \bigcup_{k=1}^{N} \overline{N_B}X_k, \tag{3}$$

where:

$$\underline{N_B}X_k = \{x_i \mid \delta_B(x_i) \subseteq X_k, x_i \in U, 1 \le i \le n\}, \tag{4}$$

$$\overline{N_B}X_k = \{x_i \mid \delta_B(x_i) \cap X_k \ne \varnothing, x_i \in U, 1 \le i \le n\}, \tag{5}$$

and the lower approximation is usually called decision positive region, denoted by $POS_B(D)$.

The decision boundary region of the decision attribute D with regard to the condition attribute set B is defined as:

$$BN(D) = \overline{N_B}D - \underline{N_B}D. \tag{6}$$

Figure 2 shows a binary classification problem with two condition attributes B_1 and B_2. The decision attribute D partitions the domain U into two equivalence classes X_1 and X_2. X_1 is labeled with "*" and X_2 is labeled with "+". Let the metric Δ be a circular neighborhood with radius δ for each condition attribute. Given three samples x_1, x_2 and x_3, on the basis of above definitions, we can get $\delta_{B_1}(x_1) \subseteq X_1$, $\delta_{B_1}(x_3) \subseteq X_2$, $\delta_{B_1}(x_2) \cap X_1 \ne \varnothing$ and $\delta_{B_1}(x_2) \cap X_2 \ne \varnothing$. Therefore, $x_1 \in \underline{N_{B_1}}X_1$, $x_3 \in \underline{N_{B_1}}X_2$ and $x_2 \in B_1N(D)$. Similarly, we can obtain $\delta_{B_2}(x_1) \subseteq X_1$, $\delta_{B_2}(x_3) \subseteq X_2$, $\delta_{B_2}(x_2) \cap X_1 \ne \varnothing$ and $\delta_{B_2}(x_2) \cap X_2 \ne \varnothing$. Therefore, $x_1 \in \underline{N_{B_2}}X_1$, $x_3 \in \underline{N_{B_2}}X_2$ and $x_2 \in B_2N(D)$.

Definition 3 ([24]). The dependence degree of the decision attribute D on the condition attribute set B is defined as:

$$\gamma_B(D) = |POS_B(D)|/|U|. \tag{7}$$

Obviously, $0 \le \gamma_B(D) \le 1$. If $\gamma_B(D) = 1$, the decision attribute D completely depends on the condition attribute set B.

Definition 4 ([24]). Consider a neighborhood decision system $\langle U, A, D \rangle$, for any $B \subseteq A$, $a \in A - B$, the significance of the attribute a to condition attribute set B is given as:

$$\mathrm{SIG}(a, B, D) = \gamma_{B \cup a}(D) - \gamma_B(D). \tag{8}$$

2.2 The NRS_FS_FAST algorithm

As we all know, setting the same neighborhood size for all attributes can affect the results of feature selection, due to the fact that the data distribution of each attribute is often different. To settle this issue, the NRS_FS_FAST algorithm is introduced in this paper, and the pseudo-code is summarized in Algorithm 1.

Algorithm 1: The NRS_FS_FAST algorithm.

Input:
NDT : The neighborhood decision system $NDT = \langle U, A, D \rangle$.
L : The given parameter.
Output:
Red: The attribute subset.

Begin:
Process the sample data by the normalization method.
$\forall a \in A$, get neighborhood relation matrix: $N_a =$ GetNeighborRelation $(NDT,\ L)$.
Initialize the $red = \varnothing, pos = \varnothing$.
    red: the attribute reduction set.
    pos: the positive region of attribute reduction.
Select attributes: $Red = SelectAttributes(NDT, L, N_a)$.
for $i = 1, \ldots, m$ **do**
    (1) each attribute $a_i \in A - red$.
    (2) $SIG(a_i, red, D) = \gamma_{red \cup a_i}(D) - \gamma_{red}(D)$.
    (3) $\underline{N_{red \cup a_i}}(D) = \bigcup_{i=1}^{N} \underline{N_{red \cup a_i}} X_i$. $X_1, X_2, \ldots, X_N$ are equivalence classes.
    (4) Find out the attribute a_k with the largest significance degree and calculate its positive region.
        $SIG(a_k, red, D) = \max_i \{SIG(a_i, red, D)\}$.
        $pos = \underline{N_{red \cup a_k}}(D)$.
end
if $\Delta(SIG(a_k, red, D)) > 0$ **then**
    (i) $red = red \cup a_k$.
    (ii) $a_i \in A - red$, delete the posth rows and the posth columns of N_{a_i}.
    (iii) $i = i + 1$.
else
    $Red = red$.
    break.
end

Definition 5 ([25]). Let $NDT = \langle U, A, D \rangle$ be a neighborhood decision system, $SD = \{SD_1, SD_2, \ldots, SD_m\}$ is a set of standard deviation of each attribute, where m is the number of attributes. The relationship matrix of neighborhood relation N_i of the attribute i on U is defined as:

$$M_{(N_i)} = (r_{p,q})_{n \times n}, \tag{9}$$

where

$$r_{p,q} = \begin{cases} 1, & \Delta(x_p, x_q) \le \delta_i \\ 0, & \text{others} \end{cases}, 1 \le p \le n, 1 \le q \le n, \tag{10}$$

$\delta_i = SD_i / L$ denotes the threshold of neighborhood size, and L is a given parameter to control the size of neighborhood.

3 The introduction of the AFS theory

This section mainly recalls the notations of the AFS theory. The detailed introduction can be referred to [26–31].

Table 1 The description of samples

	Age	Appearance		Wealth		Gender		Hair color		
		Height	Weight	Salary	Estate	Male	Female	Black	White	Yellow
x_1	20	1.9	90	1	0	1	0	6	1	4
x_2	13	1.2	32	0	0	0	1	4	3	1
x_3	50	1.7	67	140	34	0	1	6	1	4
x_4	80	1.8	73	20	80	1	0	3	4	2
x_5	34	1.4	54	15	2	1	0	5	2	2
x_6	37	1.6	80	80	28	0	1	6	1	4
x_7	45	1.7	78	268	90	1	0	1	6	4
x_8	70	1.65	70	30	45	1	0	3	4	2
x_9	60	1.82	83	25	98	0	1	4	3	1
x_{10}	3	1.1	21	0	0	0	1	2	5	3

3.1 AFS algebra

In order to explain the AFS algebra, the following example is given.

Example 1 Consider a set of data with 10 samples $X = \{x_1, \ldots, x_{10}\}$ and the features of data are described by real numbers (age, appearance, and wealth), Boolean values (gender) and the order relations (hair color), shown in Table 1.

In Table 1, the order number i in the "hair color" columns denotes that the hair color of $x \in X$ has ordered according to our perception. Take the order relations of "hair black," for example, $x_7 > x_{10} > x_4 = x_8 > x_2 = x_9 > x_5 > x_6 = x_3 = x_1$, where $x_i > x_j$ means that the hair of x_i is more closer to black color than that of x_j.

Let $M = \{m_1, m_2, \ldots, m_{12}\}$ be a set of simple fuzzy concepts on X and each $m \in M$ associates with one single attribute. These fuzzy concepts can be explained as follows: m_1: "old persons", m_2: "tall persons", m_3: "heavy persons", m_4: "high salary", m_5: "more estate", m_6: "male", m_7: "female", m_8: "black hair persons", m_9: "white hair persons", m_{10}: "yellow hair persons", m_{11}: "young persons", and m_{12}: "the persons about 40 years old". For any $A \subseteq M$, $\prod_{m\in A} m$ indicates the conjunction (i.e., the logic operator "and") of concepts in A. For example, for $A = \{m_1, m_6\} \subseteq M$, $\prod_{m\in A} m$ indicates the new complex fuzzy concept "old males" associated with two attributes "age" and "gender".

$\sum_{i\in I}(\prod_{m\in A_i} m)$ is the formal sum of $\prod_{m\in A_i} m$ ($A_i \subseteq M, i \in I$, and I is a non-empty index set), which denotes the disjunction (i.e., the logic operator "or") of the complex fuzzy concept $\prod_{m\in A_i} m$. For example, $\gamma = m_1m_6 + m_1m_3 + m_2$ can be interpreted as "old males" or "heavy old persons" or "tall persons". By the comparison of the complex fuzzy concepts $m_3m_8 + m_1m_4 + m_1m_6m_7 + m_1m_4m_8$ and $m_3m_8 + m_1m_4 + m_1m_6m_7$, we can get that the fuzzy concepts in the left side and right side are equivalent. This is due to the fact that, for any sample x, the degree of x belonging to the fuzzy concept $m_1m_4m_8$ is always less than or equal to that of m_1m_4. Therefore, the term $m_1m_4m_8$ is redundant in the left side.

Take two complex fuzzy concepts with the form $\alpha : m_1m_4 + m_2m_5m_6$ and $\upsilon : m_5m_6 + m_5m_8$ into consideration, the semantics of "α or υ" and "α and υ" can be explained as:

"α or υ": $m_1m_4 + m_2m_5m_6 + m_5m_6 + m_5m_8 = m_1m_4 + m_5m_6 + m_5m_8$;

"α and υ": $m_1m_4m_5m_6 + m_2m_5m_6 + m_1m_4m_5m_8 + m_2m_5m_6m_8 = m_1m_4m_5m_6 + m_2m_5m_6 + m_1m_4m_5m_8$.

The $\sum_{i\in I}(\prod_{m\in A_i} m)$ forms AFS algebra, and the set EM^* is given as:

$$EM^* = \left\{\sum_{i\in I}\left(\prod_{m\in A_i} m\right) | A_i \subseteq M,\ I \text{ is a non-empty index set}\right\}, \tag{11}$$

where the symbols Σ and Π denote the disjunction and conjunction of fuzzy concepts, respectively.

For any fuzzy concepts m, n, and $h \in M$, the AFS algebra is based on the following assumptions:

(1) "m and m and n" is equivalent to "m and n";
(2) "m and n" is equivalent to "n and m";
(3) " m and n and h" or "n and m" is equivalent to "n and m".

Definition 6 ([27]) Consider a simple fuzzy concept set M, the binary equivalence relation R on EM^* can be described as: for any $\sum_{i\in I}(\prod_{m\in P_i} m)$, and $\sum_{j\in J}(\prod_{m\in Q_j} m) \in EM^*$, $(\sum_{i\in I}(\prod_{m\in P_i} m))R(\sum_{j\in J}(\prod_{m\in Q_j} m)) \Leftrightarrow$

(1) for arbitrary $P_i (i \in I)$, there exists $Q_h (h \in J)$ such that $P_i \supseteq Q_h$;
(2) for arbitrary $Q_j (j \in J)$, there exists $P_k (k \in I)$ such that $Q_j \supseteq P_k$.

The notation $\sum_{i\in I}(\prod_{m\in P_i} m)R\sum_{j\in J}(\prod_{m\in Q_j} m)$ means that $\sum_{i\in I}(\prod_{m\in P_i} m)$ and $\sum_{j\in J}(\prod_{m\in Q_j} m)$ are equivalent under equivalence relation R. For example, $\xi = m_3m_8 + m_1m_4 + m_1m_6m_7 + m_1m_4m_8$, $\zeta = m_3m_8 + m_1m_4 + m_1m_6m_7 \in EM$, by Definition 6, we have $\xi = \zeta$.

Theorem 1 ([27]) *Consider a simple fuzzy concept set M, the $(EM, \vee, \wedge)$ constitutes a completely distributive lattice if $\sum_{i\in I}(\prod_{m\in P_i} m) \in EM$ and $\sum_{j\in J}(\prod_{m\in Q_j} m) \in EM$ satisfy the following conditions:*

$$\sum_{i\in I}\left(\prod_{m\in P_i} m\right) \vee \sum_{j\in J}\left(\prod_{m\in Q_j} m\right) = \sum_{k\in I\bigsqcup J}\left(\prod_{m\in W_k} m\right), \tag{12}$$

$$\sum_{i\in I}\left(\prod_{m\in P_i} m\right)\wedge\sum_{j\in J}\left(\prod_{m\in Q_j} m\right) = \sum_{i\in I, j\in J}\left(\prod_{m\in P_i\cup Q_j} m\right), \tag{13}$$

where $I\bigsqcup J$ denotes the disjoint union of I and J. Therefore, $W_k = P_k$ if $k \in I$, and $W_k = Q_k$ if $k \in J$.

Definition 7 ([30]) Given a simple fuzzy concept m on X, the binary relation R_m is described as: for any $u, v \in X, (u, v) \in R_m \Leftrightarrow u$ belongs to the fuzzy concept m, and the degree of u belonging to m is larger than or equals to that of v; or u belongs to m to some extent while v does not belong to m.

3.2 AFS structure

AFS structure can produce various lattice representations of the fuzzy logic operations and fuzzy membership degrees [27, 30].

Definition 8 ([30]) Given a simple fuzzy concept set M, $\tau(u, v)=\{\xi|\xi \in M, (u, v) \in R_\xi\} \in 2^M$, if τ meets the following conditions:

(1) for any $(u, v) \in X \times X, \tau(u, v) \subseteq \tau(u, u)$,
(2) for any $(u, v), (v, w) \in X \times X, \tau(u, v) \cap \tau(v, w) \subseteq \tau(u, w)$,

where 2^M is the power set of M, and (M, τ, X) is considered as AFS structure.

For instance, when $x = x_4, y = x_1, x_2\ldots, x_{10}$, respectively, one has

$$\begin{aligned}
\tau(x_4, x_1) &= \{m_1, m_4, m_5, m_6, m_8, m_{10}\},\\
\tau(x_4, x_2) &= \{m_1, m_2, m_3, m_4, m_5, m_6, m_8\},\\
\tau(x_4, x_3) &= \{m_1, m_2, m_3, m_5, m_6, m_8, m_{10}\},\\
\tau(x_4, x_4) &= \{m_1, m_2, m_3, m_4, m_5, m_6, m_8, m_9, m_{10}, m_{11}, m_{12}\},\\
\tau(x_4, x_5) &= \{m_1, m_2, m_3, m_4, m_5, m_6, m_8, m_{10}\},\\
\tau(x_4, x_6) &= \{m_1, m_2, m_5, m_6, m_8, m_{10}\},\\
\tau(x_4, x_7) &= \{m_1, m_2, m_6, m_9, m_{10}\},\\
\tau(x_4, x_8) &= \{m_1, m_2, m_3, m_5, m_6, m_8, m_9, m_{10}\},\\
\tau(x_4, x_9) &= \{m_1, m_6, m_8\},\\
\tau(x_4, x_{10}) &= \{m_1, m_2, m_3, m_4, m_5, m_6, m_9, m_{10}\}.
\end{aligned}$$

Definition 9 ([30]) Given a simple fuzzy concept set M, for any $A \subseteq M$ and $x \in X$, $A^\tau(x) \subseteq X$ is defined as:

$$A^\tau(x) = \{y \in X|\tau(x, y) \supseteq A\}. \tag{14}$$

For instance, if $x = x_4$, $A = \{m_1, m_2, m_3\} \subseteq M$, we have $\tau(x_4, x_2) \supseteq A$, $\tau(x_4, x_3) \supseteq A$, $\tau(x_4, x_4) \supseteq A$, $\tau(x_4, x_5) \supseteq A$, $\tau(x_4, x_8) \supseteq A$, and $\tau(x_4, x_{10}) \supseteq A$ (refer to Definition 8). Then $A^\tau(x_4) = \{x_2, x_3, x_4, x_5, x_8, x_{10}\}$ can be obtained. In summary, $A^\tau(x)$ is determined by data distribution and the semantic interpretations of fuzzy sets.

Definition 10 ([27]) Given a simple fuzzy concept ω on X, if $\rho_\omega : X \to R^+$ satisfies the following two conditions:

(1) $\rho_\omega(x) = 0 \Leftrightarrow (x, x) \notin R_\omega, \forall x \in X$;
(2) $\rho_\omega(x) \geq \rho_\omega(y) \Leftrightarrow (x, y) \in R_\omega, \forall x, y \in X$,

ρ_ω is called as the weight function of fuzzy concept ω.

Definition 11 ([27]). Consider a simple fuzzy concept set M and weight function ρ_m of $m \in M$, for any $x \in X$, $A_i \subseteq M$, the membership degree of x belonging to $\zeta = \sum_{i\in I}(\prod_{m\in A_i} m) \in EM$ is given as:

$$\mu_\zeta(x) = \sup_{i\in I} inf_{m\in A_i} \frac{\sum_{u\in A_i^\tau(x)} \rho_m(u) N_u}{\sum_{u\in X} \rho_m(u) N_u}, \tag{15}$$

here N_u is the number that how many times u has been observed.

From Eq. (15), we can see that $\mu_\zeta(x)$ is determined by $A_i^\tau(x)$, simple concept set A_i and weight function $\rho_m(u)$. From Definition 9, $A^\tau(x)$ is determined by data distribution and the semantics of fuzzy sets. Thus, if the weight function has been determined, the membership function is only related to the data distribution and the semantics of fuzzy sets.

4 The construction of the FRODT

4.1 The expression of fuzzy rules

Let $X = [x_{ij}]_{n\times m}$ be a sample data set, here m is the feature numbers and n is the sample numbers. The set of class labels of X is $C = \{1, 2, \ldots, c\}$, and the jth feature of X is $f_j (j = 1, 2, \ldots, m)$. Moreover, $X_l (l = 1, 2, \ldots, c)$ is the set of samples belonging to the lth class.

For simplicity, the fuzzy concepts "small", "medium" and "big" are adopted to describe the characteristics of each feature f_j. We use $f_{j,p}$ to represent the pth fuzzy concept of the jth attribute. The linguistic interpretation of $f_{j,p} (p = 1, 2, 3)$ is that f_j is small, medium, and big, respectively.

Algorithm 2: FREA.

Input:
$X = [x_{i,j}]_{n\times m}$: The training samples set with c classes.
$M = \{1, 2, \ldots, m\}$: The set containing m attributes.
β: The parameter used to keep balance between the number of fuzzy concepts in the corresponding rule and its Fuzzy Confidence level FConf ($0 \le \beta \le 1$, and usually is set to 0.02).
$MaxR$: The preset parameter denoting the maximum length of fuzzy rule, i.e. the number of fuzzy concepts in the corresponding rule.
Output:
R_l : The obtained fuzzy rule to describe the lth class.

Begin:
 Initialize $\tilde{A} = \varnothing$. $\tilde{A}$ represents the set of fuzzy concepts.
 Calculate the fuzzy concepts $f_{j,p}(j \in M, p = 1, 2, 3)$ of root node samples and let $A = \{f_{j,p}\}(j \in M, p = 1, 2, 3)$.
for l=1,...,c **do**
  Calculate the fuzzy confidence degrees of all fuzzy concepts:
   $\forall f \in A$, compute $FConf(f \Rightarrow l)$.
  for t=1,...,$min(MaxR,\ c*m)$ **do**
   Determine the corresponding fuzzy concept with the maximum Fuzzy Confidence level:
    $f_{a_t,b_t} = arg\ max_f FConf(f_{j,p} \Rightarrow l)(j \in M, p = 1, 2, 3)$.
   Compute the complex fuzzy concept f_t of the tth cycle:
    Let $\tilde{A} = \{\tilde{A}, f_{a_t,b_t}\}$, for $\forall f \in \tilde{A}$, $f_t = \prod_{f\in\tilde{A}} f$.
   An index used for determining the antecedent part of the fuzzy rule is computed as follows:
    $FC_t = FConf(f_t \Rightarrow l) + t * \beta$.
   Remove the fuzzy concepts $\{f_{a_t,p}\}(p = 1, 2, 3)$ of the a_tth attribute and obtain the rest attributes:
    $M = M \setminus a_t$.
   The remaining fuzzy concepts are denoted as:
    $\hat{A} = \{f_{j,p}\}(j \in M, p = 1, 2, 3)$.
   if $M \neq \varnothing$ *and the length of* $f_m < min(MaxR, c * m)$ **then**
    $\forall \hat{f} \in \hat{A}$, $let f = \hat{f} \wedge f_t$, where the symbol $\wedge$ stands for "and" logic operator.
    Calculate the fuzzy confidence degrees: $FConf(f \Rightarrow l)(j \in M, p = 1, 2, 3)$.
    $t = t + 1$.
   else
    $d = arg\ max_t FC_t$, so the antecedent part (i.e. IF part) of the fuzzy rule R_l is f_d.
    break.
    return R_l.
   end
  end
end

The main features of data can be well described by the fuzzy IF-THEN rules. In 1997, Zadeh proposed a method of expressing fuzzy rules [32]:

Rule: if x_{i1} is big and x_{i2} is small, then the sample x_i falls into the first class.

By the defined fuzzy concepts $f_{j,p}$, we can rewrite the above rule:

Rule: if the sample x_i is $f_{1,3}$ and $f_{2,1}$, then it is classified as the first class.

According to the definitions of logical operations "and" and "or" in Example 1, we can redescribe the above rule:

Rule: if the sample x_i is $f_{1,3}f_{2,1}$, then it falls into the first class.

4.2 Fuzzy rule extraction

Fuzzy IF-THEN rules are critical for constructing the FRODT. The classical form of fuzzy association rules is $A \Rightarrow B$. It means that an element satisfies the condition A can also satisfy the condition B. The association degrees of association rules were measured by the Support and Confidence indices [33]. Later, Fuzzy Support and Fuzzy Confidence indices were applied in classification problems [34]. Inspired by [34], based on AFS theory, we define the Fuzzy Confidence indice of this paper, as follows:

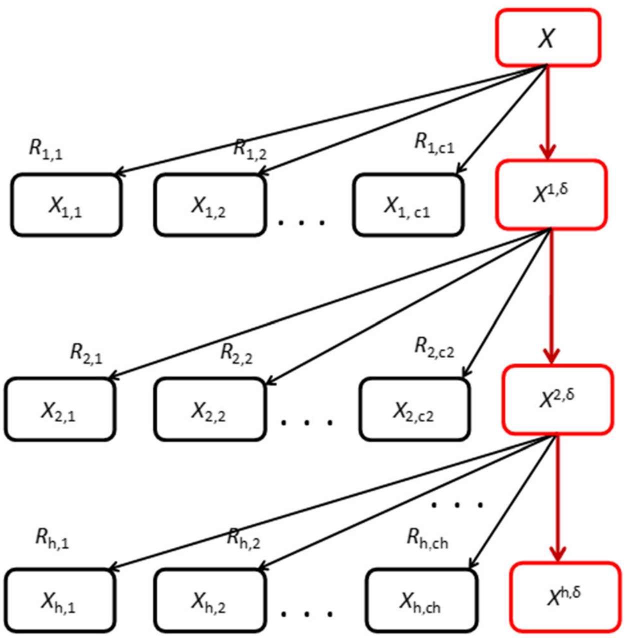

Fig. 3 The overall structure of the FRODT

$$FConf(F \Rightarrow c) = \frac{\sum_{x \in X_c} \mu_F(x)}{\sum_{x \in X} \mu_F(x)}, \quad (16)$$

$$\mu_F(x) = \frac{\sum_{f \in A} \mu_f(x)}{|A|}, \quad (17)$$

where the former part of the fuzzy rule corresponds to the complex fuzzy concept F, and the class label corresponds to c. $\mu_F(x)$ ($\mu_f(x)$) demonstrates the average membership degree (membership degree) of x that is described by the complex fuzzy concept F (simple fuzzy concept f), and |.| represents the cardinality of a set. Moreover, A is the set of all simple fuzzy concepts contained in F, for example, if $F = f_{11}f_{22}$, then $A = \{f_{11}, f_{22}\}, |A| = 2$. The bigger the fuzzy confidence degree is, the more suitable complex fuzzy concept F is for describing the lth class.

The FREA based on Fuzzy Confidence is proposed to extract only one single rule for each class, shown in Algorithm 2.

4.3 The architecture of the FRODT

The architecture of the FRODT is shown in Fig. 3, and the build-up process is as follows.

Firstly, all sample data X is contained at the root node of the FRODT. At this node, we use the FREA to extract fuzzy rules, expressed as $R_{1,l}(l = 1, \ldots, c_1)$ with "1" denoting the first layer and "l" indicating the lth category. Since only one rule is extracted for each class, thus $c_1 = c$. Each class is assigned a leaf node, which contains as many samples that belongs to this class as possible. The samples that cannot be covered by these rules are placed on an non-leaf node $X^{1,\sigma}$. Besides, the threshold σ is applied to judge whether the sample can be covered by these rules or not. In order to balance the accuracy of classification and the size of tree, we propose a new algorithm—Rules Covering Samples Algorithm (RCSA), as shown in Algorithm 3.

Secondly, on the non-leaf node $X^{1,\sigma}$, we use the FREA again to extract fuzzy rules, expressed as $R_{2,l}(l = 1, \ldots, c_2)$. Each class is also assigned a leaf node and the samples that cannot be covered by these rules are placed on the second non-leaf node $X^{2,\sigma}$.

$\ddots$

Finally, on additional non-leaf node $X^{h,\sigma}$, the FRODT continues to grow until it meets one of the three stopping conditions: (1)$X^{h,\sigma} = \varnothing$, (2) $X^{h,\sigma} = X^{h-1,\sigma}$ and (3) $c_h = 1$. They, respectively, represent that, in the hth layer of the FRODT, the rules cover all samples; the rules lose its effect; and the samples belong to the same class. In the second case, we select the class with the largest number of samples as the final category.

The construction process of FRODT is summarized in Algorithm 4.

Algorithm 3: RCSA.

Input:
$R_{h,l}(l = 1, 2, \ldots, c_h)$: The rule to describe the lth category in the hth layer of the tree.
σ : The threshold value, $0 \leq \sigma \leq 1$.
Output:
$X_{h,l}$: The training samples that can be decided by the rule $R_{h,l}$.
$X^{h,\sigma}$: The samples contained in additional non-leaf node in the hth layer of the tree.

Begin:
Calculate the fuzzy membership function, as follows:
$$R_{h,l}(x) = F_{h,l}(x) = \frac{\sum_{f \in F_{h,l}(x)}}{|F_{h,l}|}.$$
%Where $F_{h,l}$ denotes the IF part of $R_{h,l}$.
Determine the samples that the rule $R_{h,l}$ covers:
$$X_{h,l} = \{x | R_{h,l}(x) > \sigma\}(x \in X^{h-1,\sigma}).$$
%It can be seen from the above formula that the threshold σ is the key parameter to determine the sample set $X_{h,l}$.
Optimize the parameter σ:
$$G(\sigma) = |X| * A_s - \sigma * M.$$
%Here, $|X|$ and M represent the sample numbers and leaf nodes numbers, respectively. Moreover, A_s denotes the accuracy of the samples.
Determine the samples in the additional non-leaf node $X^{h,\sigma}$:
$$X^{h,\sigma} = X^{h-1,\sigma} \setminus \bigcup X_{h,l}.$$
return $X_{h,l}$ and $X^{h,\sigma}$.

4.4 Analysis of the time complexity

In this subsection, we analyze the time complexity of Algorithms 1–4. Assuming that the Algorithms 1–4 are conducted on one data set with n training instances, m attributes, and c class label. In general, the number of training instances is greater than that of attributes, i.e., $n > m$. Algorithm 1 is composed of two main phases: GetNeighborRelation and SelectAttributes. By Definition 5 and Algorithm 1, we can get that the time complexity of GetNeighborRelation is $O(m * n * n)$, and the time complexity of SelectAttributes is $O(m * m * n)$. Since $n > m$, thus the time complexity of Algorithm 1 is $O(m * n * n)$. For Algorithm 2, the maximum length of fuzzy rule is $H = min\{MaxR, c * m\}$, thus the time complexity of Algorithm 2 is $O(H * c)$. For Algorithm 3, it is easy to get that the time complexity of Algorithm 3 is $O(1)$. For Algorithm 4, the number of layers of the FRODT is the number of samples in the worst situation, i.e., only one sample of training data is determined on each layer of the tree. Therefore, the time complexity of Algorithm 4 is $O(H * c * n)$ in the worst situation. However, the depth of the tree is on the order of $O(logn)$. Thus, the total time complexity of the Algorithm 4 is $O(H * c * logn)$.

5 Experimental results and analysis

In this section, we compare the FRODT with HHCAT [19] and five classical classification algorithms such as SC [5], C4.5 [7], BFT [35], LAD [36], and NBT [37] under the Waikato environment for knowledge analysis (WEKA) 3.6 framework [38]. We use ten times tenfold cross-validation to estimate the classification accuracy and tree size of the FRODT. In the experiment, the parameter L in the NRS_FS_FAST is set to 2, the number of simple fuzzy concepts on each attribute is set to 3, and the adjustment factors $MaxR$ and β in the FREA are set to 5 and 0.02, respectively. For the sake of simplicity, $N_u = 1$ and $\rho_m(u) = 1$ in this paper.

5.1 The experiment on Iris data

Iris data is one of the most commonly used data in the UCI machine learning database. We apply the proposed algorithm to Iris data set, and the detailed process is as follows.

Algorithm 4: The FRODT algorithm.

Input:
X: The samples set with c classes $X_l (l = 1, \ldots, c)$.
β: The parameter used to keep balance between the number of fuzzy concepts in the corresponding rule and its Fuzzy Confidence level FConf ($0 \le \beta \le 1$, and usually is set to 0.02).
$MaxR$: The preset parameter denoting the maximum length of fuzzy rule, i.e. the number of fuzzy concepts in the corresponding rule.
σ : The threshold value, $0 \le \sigma \le 1$.
Output:
$R_{h,l}$: The obtained fuzzy rule.
$X_{h,l}$: The training samples that can be covered by $R_{h,l}$.
$X^{h,\sigma}$: The training samples that cannot be covered by $R_{h,l}$.

Begin:
  $X^{0,\sigma} = X$, h=1.
while $X^{h-1,\sigma} \neq \varnothing$ **do**
    Compute the class number c_h of $X^{h-1,\sigma}$.
    if $c_h = 1$ **then**
        break.
    end
    Compute $R_{h,l} = FREA(X^{h-1,\sigma}, MaxR, \beta)$.
    Compute $\{X_{h,l}, X^{h,\sigma}\} = RCSA(R_{h,l}, \sigma)$.
    if $X^{h,\sigma} = X^{h-1,\sigma}$ *or* $X^{h,\sigma} = \varnothing$ **then**
        break.
    end
    $h = h + 1$.
    return $R_{h,l}$ and $X_{h,l}, X^{h,\sigma}$.
end

First, the NRS_FS_FAST algorithm is applied to select an attribute subset, which has petal length, petal width, sepal length and sepal width, respectively. That is, the number of attributes is not reduced on iris data.

Second, according to the AFS theory, we can get the fuzzy concepts $f_{i,j}, i = 1, 2, 3, 4, j = 1, 2, 3$ of each attribute. For example, the semantics of $f_{2,2}$ is " the width of sepal is medium ". Moreover, by Eq. (15), we can get the membership functions of these concepts.

Then, we adopt the FREA to generate fuzzy rules that are used for building the FRODT. Given parameter σ =0.6, we can get a tree depicted in Fig. 4a. And the corresponding rules of the FRODT are given as:

IF the sample x is $F_{1,1}$, THEN it belongs to class 1 $\longrightarrow R_{1,1}$,
ELSE IF x is $F_{1,2}$, THEN it belongs to class 2 $\longrightarrow R_{1,2}$,
ELSE IF x is $F_{1,3}$, THEN it belongs to class 3 $\longrightarrow R_{1,3}$,

where $F_{1,1} = f_{3,1}f_{4,1}f_{1,1}f_{2,3}$, $F_{1,2} = f_{3,2}f_{4,2}$, $F_{1,3} = f_{3,3}f_{4,3}f_{1,3}$.

The linguistic interpretations of the fuzzy rules are that " the samples that have short petal length, short petal width, short sepal length and long sepal width belong to class 1; the samples that have medium petal length, medium petal width belong to class 2; and the samples that have long petal length, long petal width and long petal length belong to class 3 ". Moreover, Fig. 4b presents the tree builded by the C4.5 algorithm. It can be observed that the structure of the FRODT is obviously simpler than that

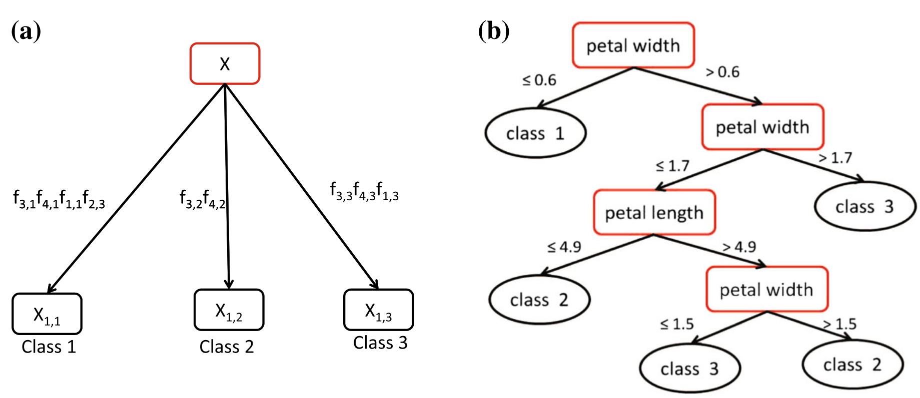

Fig. 4 **a** The structure of the FRODT, **b** the tree obtained by C4.5

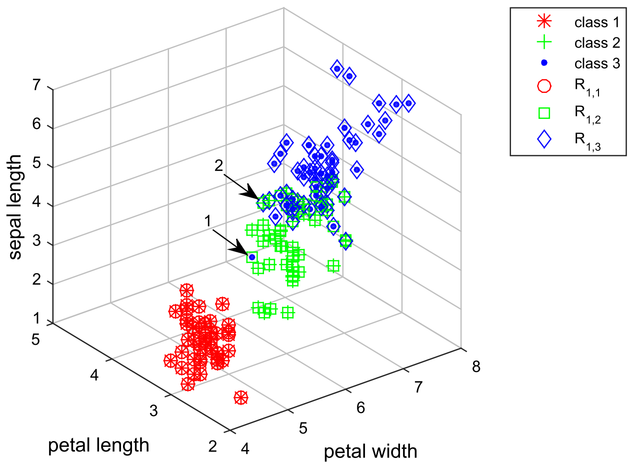

Fig. 5 The three-dimensional classification result of Iris training data by fuzzy rules $R_{1,1}, R_{1,2}$ and $R_{1,3}$

of C4.5 tree. Therefore, the tree size in this paper is less than C4.5.

Moreover, the three-dimensional classification result of Iris data by fuzzy rules $R_{1,1}, R_{1,2}$, and $R_{1,3}$ is shown in Fig. 5. The red circles indicate the samples determined by the rule $R_{1,1}$, the green squares represent the samples determined by the rule $R_{1,2}$, and the blue diamonds stand for the samples determined by the rule $R_{1,3}$. Besides, arrow 1 indicates that the rule $R_{1,2}$ classifies the samples that belong to the third class into the second class, and arrow 2 represents that the rule $R_{1,3}$ classifies the samples that belong to the second class into the third class. In addition, the membership degrees of iris training data on fuzzy rules $R_{1,1}, R_{1,2}$ and $R_{1,3}$ are depicted in Fig. 6. It shows that the samples in the first class originally belong to the rule $R_{1,1}$ with the largest membership degrees, the samples in the

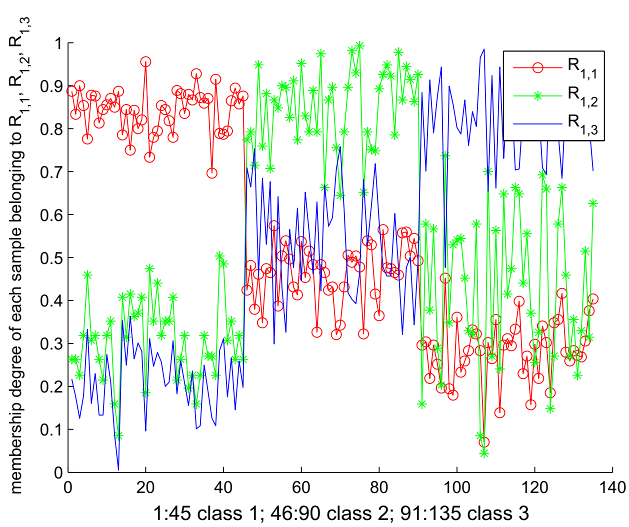

Fig. 6 Membership degrees of Iris training data on fuzzy rules $R_{1,1}, R_{1,2}, R_{1,3}$

second class originally belong to the rule $R_{1,2}$ with the largest membership degrees, and the samples in the third class originally belong to the rule $R_{1,3}$ with the largest membership degrees. That is, we can obtain satisfactory results by using fuzzy rules $R_{1,1}, R_{1,2}$ and $R_{1,3}$ to classify the iris data set.

5.2 Comparison of the FRODT and HHCART

In order to demonstrate the superiority of the proposed method, we compare FRODT with HHCART of [19] in this section.

Table 2 summarizes the results in terms of classification accuracy and tree size along with the respective standard deviations. It shows that the classification accuracy of the FRODT is higher than HHCART tested for all data sets except Boston Housing and Glass. From the "tree size" column, it demonstrates that the FRODT produces fewer leaf nodes than the chosen benchmarks on Boston Housing, Bupa, Glass, Heart and Survival data sets. Moreover, from the last line of Table 2, we can obtain that the average classification accuracy and tree scale of the FRODT are better than those of the rival algorithms. Moreover, Fig. 7 depicts these results. These results advocate that our strategy has more preferable performance than the comparison algorithms.

5.3 Comparison of the FRODT and five conventional decision trees

In order to better verify the superiority of the FRODT, we also compare FRODT with five conventional decision trees such as SC, C4.5, BFT, LAD, and NBT on 20 UCI data sets, shown in Table 3.

The classification accuracies of the FRODT and five state-of-the-art methods on 20 UCI data sets are presented in Table 4. The overall picture conveyed by the results in Table 4 is clearly in favor of the FRODT. The FRODT outperforms the other methods on most data sets. In particular, it is not as good as the traditional decision tree SC only in three data sets and BFT or C4.5 only in four data sets and LAD only in five data sets. Besides, compared with the chosen benchmarks, the FRODT obtains the highest average classification accuracy. Note that the symbol "★" indicates that the NBT is inoperative on AMLALL data set.

Table 5 shows the tree sizes of the FRODT and five chosen benchmarks on 20 UCI data sets. It can be observed that the FRODT is better than BFT, C4.5, LAD, and SC on all data sets. In particular, from the last line of Table 5, we can get that the FRODT has the least average tree size.

Table 2 The classification accuracy (%) and tree size along with the respective standard deviations of the HHCART(A), HHCART(D) and FRODT, and the best scores are indicated in boldface

Dataset	Classification accuracy			Tree size		
	HHCART(A)	HHCART(D)	FRODT	HHCART(A)	HHCART(D)	FRODT
Boston housing	**83.4** ± 1.2	82.0 ± 1.1	81.72 ± 0.07	7.0 ± 2.9	8.0 ± 2.8	**5.3** ± 1.0
Breast cancer	97.0 ± 0.3	97.0 ± 0.3	**97.13** ± 0.24	**2.3** ± 0.4	2.6 ± 1.1	6.5 ± 0.4
BUPA	64.9 ± 3.0	64.8 ± 2.1	**66.58** ± 2.59	7.8 ± 1.5	10.2 ± 3.0	**7.3** ± 0.4
Glass	**61.9** ± 3.1	61.7 ± 3.4	51.53 ± 1.95	8.8 ± 3.1	10.7 ± 2.7	**5.9** ± 0.5
Heart	75.0 ± 2.3	75.2 ± 3.6	**81.56** ± 1.41	5.5 ± 1.9	8.1 ± 3.1	**4.6** ± 0.2
Pima Indian	73.2 ± 1.4	73.7 ± 1.5	**74.96** ± 1.05	11.9 ± 6.5	**11.5** ± 8.4	13.6 ± 0.9
Wine	91.4 ± 1.8	88.3 ± 1.8	**94.68** ± 1.01	**3.4** ± 0.3	4.7 ± 0.7	3.5 ± 0.4
Survival	72.5 ± 1.7	72.2 ± 2.2	**73.61** ± 1.08	6.5 ± 2.6	10.6 ± 5.5	**3.2** ± 0.2
Average	77.41	76.86	**77.72**	6.65	8.30	**6.23**

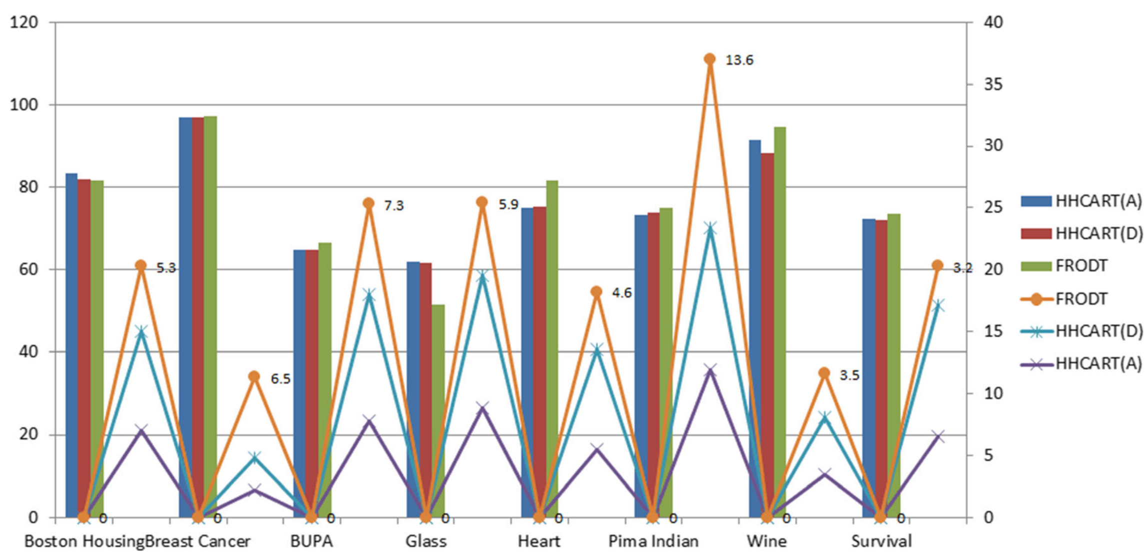

Fig. 7 The classification accuracy (%) and tree size of the HHCART(A), HHCART(D) and FRODT

Therefore, we can conclude that the FRODT can generate more simpler decision trees. Moreover, we plot the results in Tables 4 and 5 as two bar diagrams, as shown in Figs. 8 and 9. It can be seen that the FRODT obtains superior classification performance than the chosen benchmarks on accuracy and tree scale.

Remark 1 The unique features of the FRODT are highlighted in the following four aspects. Firstly, the use of AFS theory can reduce the subjectivity of choosing membership functions. Secondly, the dynamic mining fuzzy rules that are used as the decision functions at each non-leaf node can reduce the size of the tree. Thirdly, FRODT overcomes the shortcoming that the oblique decision trees lack semantic interpretation. Finally, the ideal threshold σ can be obtained by using the genetic algorithm, and the balance between classification accuracy and tree size can be achieved.

The main advantages of the results over others are as follows: (1) the FRODT performs better in both average classification accuracy and tree size than the chosen benchmarks; (2) different from the "traditional" decision trees where only one feature is considered on each node, the FRODT takes one dynamic mining fuzzy rule which involves multiple features at each node to simplify tree structure; and (3) the FRODT is endowed with readable linguistic interpretation.

5.4 The Holm test

The Holm test [39] is applied to analyze whether FRODT is significantly better than other decision trees, and the test

Table 3 Description of data sets in UCI database

No	Data set	Samples	Attributes	Classes
1	Iris	150	4	3
2	Wine	178	13	3
3	Wdbc	569	30	2
4	Credit	690	14	2
5	Heart	270	13	2
6	Haberman	306	3	2
7	Newthyroid	215	5	3
8	Wobc	699	9	2
9	Column	310	6	3
10	AMLALL	72	7129	2
11	Australian	690	14	2
12	Breast cancer	683	9	2
13	Bupa	345	6	2
14	Hepatitis	155	19	2
15	Ionosphere	351	34	2
16	Pima Indian	768	8	2
17	Sonar	208	60	2
18	Tae	151	5	3
19	Transfusion	748	4	2
20	Wpbc	198	32	2

statistic for comparing the jth classifier and the kth classifier is as follows:

$$Z = \frac{\mathrm{Rank}_j - Rank_k}{SE}, \tag{18}$$

$$Rank_j = \frac{1}{N}\sum_{i=1}^{N} r_i^j, \tag{19}$$

here $SE = \sqrt{l(l+1)/(6 \times N)}$, l is the number of classifiers, N is the number of data sets, r_i^j is the ranking of the jth classifier on the ith data set, and $Rank_j$ is the average ranking of the jth classifier on the entire data sets.

Statistic Z follows the standard normal distribution. According to Z value, we can get the corresponding probability p. Moreover, the number of classifiers is l, the number of Z needed to calculate is l, and the number of the corresponding probability p is $l-1$. We sort $p_1 \leq p_2 \leq \ldots \leq p_{l-1}$, and compare p_j with $a/(1-j)$. If $p_1 < a/(l-1)$, the hypothesis (two classifiers have the same performance) should be rejected, and then compare $p_2 < a/(l-2)$ until the last.

According to Table 4 and Eq. (19), the average accuracy ranking of all decision trees can be obtained, $Rank_{LAD} = 4.25$, $Rank_{SC} = 4.05$, $Rank_{BFT} = 4$,

Table 4 The classification accuracy (%) and standard deviation of different decision trees

Data set	BFT	C4.5	LAD	SC	NBT	FRODT
Iris	94.40 ± 0.87	94.73 ± 0.80	94.47 ± 0.87	94.20 ± 1.00	93.47 ± 1.27	**95.06** ± 0.90
Wine	89.55 ± 1.24	93.20 ± 1.35	87.08 ± 1.80	89.49 ± 1.69	**96.07** ± 1.35	94.68 ± 1.01
Wdbc	93.04 ± 0.65	93.76 ± 0.49	90.65 ± 1.21	93.16 ± 0.62	93.95 ± 0.81	**94.04** ± 0.74
Credit	84.61 ± 0.52	83.91 ± 0.71	79.48 ± 1.86	84.68 ± 0.65	84.17 ± 0.59	**85.51** ± 0.02
Heart	77.22 ± 1.70	78.15 ± 2.26	72.37 ± 1.93	78.07 ± 1.63	80.93 ± 1.22	**81.56** ± 1.41
Haberman	72.42 ± 1.47	72.16 ± 1.11	70.52 ± 2.42	73.24 ± 1.21	71.57 ± 1.31	**74.57** ± 0.97
Newthyroid	**92.93** ± 0.79	92.60 ± 0.98	88.88 ± 1.02	91.86 ± 0.93	92.37 ± 1.40	91.52 ± 1.87
Wobc	94.45 ± 0.54	95.01 ± 0.44	93.89 ± 0.76	94.74 ± 0.37	96.37 ± 0.43	**96.78** ± 0.45
Column	80.06 ± 1.52	**81.55** ± 1.19	77.32 ± 1.71	80.87 ± 1.29	80.71 ± 1.52	80.62 ± 0.91
AMLALL	84.10 ± 11.09	81.43 ± 10.99	**95.41** ± 7.40	83.94 ± 10.86	★	84.42 ± 1.84
Australian	84.60 ± 4.38	83.91 ± 3.84	84.88 ± 4.33	84.68 ± 4.09	84.17 ± 4.41	**85.50** ± 0.01
Breast cancer	94.87 ± 2.43	95.43 ± 2.42	95.90 ± 2.13	95.08 ± 2.32	96.50 ± 2.35	**97.13** ± 0.24
Bupa	67.03 ± 7.94	66.19 ± 7.20	**67.40** ± 8.16	66.19 ± 7.62	63.82 ± 8.92	66.58 ± 2.59
Hepatitis	57.10 ± 10.92	60.71 ± 11.54	57.55 ± 12.66	56.97 ± 9.66	63.81 ± 11.18	**66.17** ± 1.17
Ionosphere	89.22 ± 4.52	89.76 ± 4.64	89.36 ± 4.26	88.91 ± 4.28	**90.02** ± 4.74	89.06 ± 0.94
Pima Indian	73.30 ± 4.09	73.99 ± 5.01	74.90 ± 4.05	74.23 ± 4.17	**75.72** ± 4.04	74.96 ± 1.05
Sonar	71.63 ± 0.95	73.61 ± 9.34	75.91 ± 8.92	70.72 ± 9.42	**77.11** ± 10.33	74.58 ± 2.47
Tae	54.46 ± 12.37	**55.13** ± 13.26	52.15 ± 12.97	52.52 ± 11.65	53.84 ± 12.35	50.16 ± 2.76
Transfusion	77.95 ± 4.26	78.10 ± 3.96	77.79 ± 3.65	78.01 ± 3.95	75.44 ± 4.12	**78.18** ± 0.71
Wpbc	74.55 ± 5.59	73.61 ± 8.69	73.63 ± 8.31	74.56 ± 5.37	75.06 ± 4.74	**75.21** ± 1.81
Average	80.37	80.85	79.98	80.31	81.12	**81.81**

The best scores are indicated in boldface

Table 5 The tree size and standard deviation of different decision trees

Data set	BFT	C4.5	LAD	SC	NBT	FRODT
Iris	9.3 ± 2.1	4.6 ± 0.6	7.0 ± 0.3	7.4 ± 2	4.4 ± 2.9	**3.1** ± 0.2
Wine	10.6 ± 2.7	9.6 ± 1.2	13.0 ± 5.1	10.3 ± 3.2	3.9 ± 2.6	**3.5** ± 0.4
Wdbc	16.5 ± 4.6	22.4 ± 3.9	16.2 ± 2.6	12.6 ± 4.4	18.2 ± 3.6	**9.7** ± 0.5
Credit	30.3 ± 23.3	51.7 ± 12.1	8.8 ± 2.2	10.5 ± 10.6	14.2 ± 7.7	**2.0** ± 0
Heart	28.8 ± 11.9	34.6 ± 5.7	15.6 ± 1.2	15.4 ± 8.1	9.6 ± 3.7	**4.6** ± 0.2
Haberman	20.2 ± 22.5	21.8 ± 11.4	8.4 ± 1.9	3.8 ± 3.8	9.9 ± 6.8	**3.1** ± 0.2
Newthyroid	13.6 ± 2.8	14.9 ± 2.1	11.8 ± 1.9	11.8 ± 3.6	7.6 ± 3.2	**5.3** ± 0.4
Wobc	31.0 ± 12.4	23.5 ± 5.5	13.6 ± 3.2	15.9 ± 7.1	**5.7** ± 5.6	6.6 ± 0.4
Column	27.3 ± 11.2	23.2 ± 5.7	9.8 ± 0.9	13.3 ± 8.3	16.0 ± 5.1	**8.1** ± 0.4
AMLALL	3.8 ± 0.9	4.3 ± 0.9	28.2 ± 2.0	3.2 ± 0.6	★	**3.0** ± 0.7
Australian	30.3 ± 23.4	51.7 ± 12.2	30.3 ± 1.3	10.5 ± 10.7	14.2 ± 7.7	**2.0** ± 0
Breast cancer	26.1 ± 9.8	20.6 ± 5.1	30.9 ± 0.3	15.5 ± 6.1	**5.5** ± 5.5	6.5 ± 0.4
Bupa	44.6 ± 26.8	49.6 ± 11.2	31.0 ± 0	26.7 ± 23.1	**7.2** ± 3.3	7.3 ± 0.4
Hepatitis	21.1 ± 12.3	35.1 ± 8.1	31.0 ± 0	9.6 ± 9.3	**3.5** ± 2.7	5.1 ± 0.3
Ionosphere	14.9 ± 6.9	26.8 ± 4.2	31.0 ± 0	9.9 ± 6.9	16.2 ± 3.7	**4.5** ± 0.3
Pima Indian	49.4 ± 42.2	39.3 ± 13.7	31.0 ± 0	20.1 ± 16.0	**4.7** ± 4.6	13.6 ± 0.9
Sonar	17.6 ± 7.3	27.9 ± 3.5	31.0 ± 0	10.5 ± 7.3	13.7 ± 2.6	**6.8** ± 0.9
Tae	45.5 ± 12.0	51.8 ± 8.7	31.0 ± 0	31.2 ± 19.3	7.6 ± 2.1	**4.6** ± 0.4
Transfusion	35.1 ± 34.1	12.1 ± 4.4	31.0 ± 0	12.6 ± 6.3	**2.4** ± 1.6	4.6 ± 0.2
Wpbc	11.3 ± 9.6	22.1 ± 6.3	31.0 ± 0	4.1 ± 5.9	4.4 ± 3.2	**3.9** ± 0.6
Average	24.37	27.38	22.08	12.75	8.89	**5.39**

The best scores are indicated in boldface

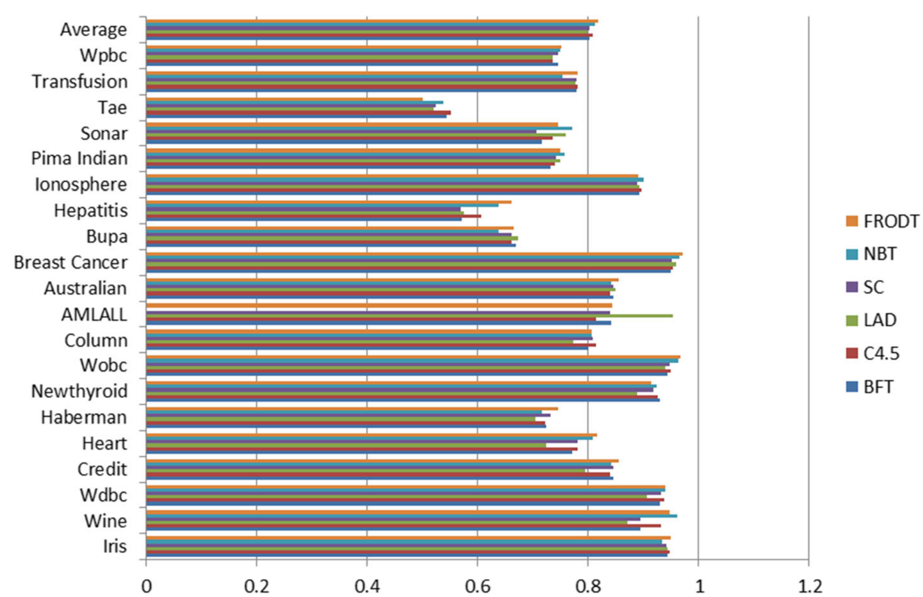

Fig. 8 The classification accuracy of different decision trees

$Rank_{C4.5} = 3.40$, $Rank_{NBT} = 3.15$, and $Rank_{FRODT} = 2.15$. The confidence level a is set to 0.05 and the results of the Holm test are presented in Table 6. It shows that the first three hypotheses are rejected and the last two hypotheses are accepted. This means that the classification accuracy of the FRODT is significantly better than that of traditional decision trees LAD, SC, and BFT. Although the FRODT is not significantly higher than C4.5 and NBT, the average classification accuracy of the FRODT is higher than that of C4.5 and NBT.

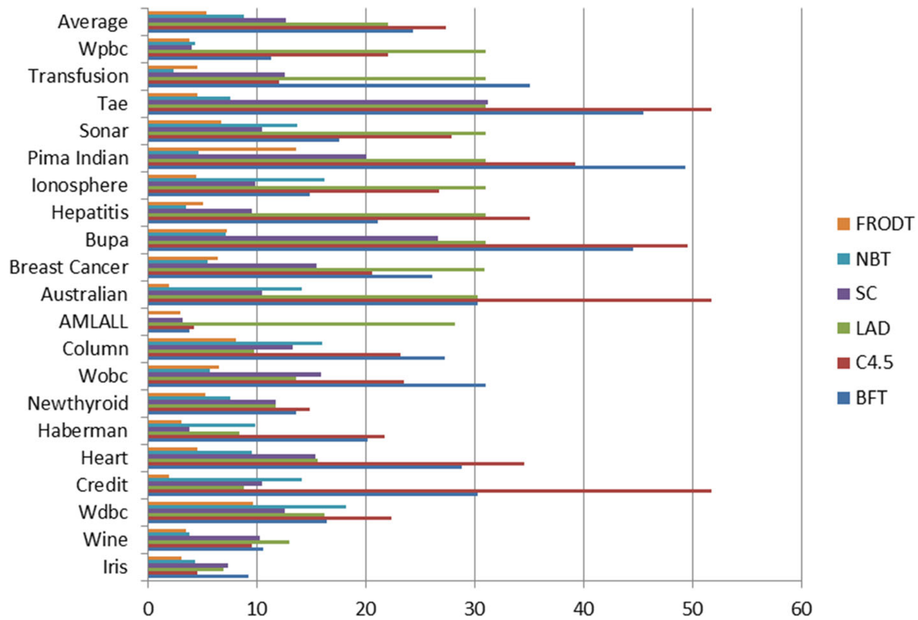

Fig. 9 The tree size of different decision trees

Table 6 The Holm test

No	Classifier	$\frac{Rank_j - Rank_{FRODT}}{SE}$	Z	p	$\frac{a}{l-j}$
1	LAD	(4.25–2.15)/0.5916	3.5497	7.3244e-04	0.01
2	SC	(4.05–2.15)/0.5916	3.2116	0.0023	0.0125
3	BFT	(4–2.15)/0.5916	3.1271	0.0030	0.0167
4	C4.5	(3.40–2.15)/0.5916	2.1129	0.0428	0.025
5	NBT	(3.15–2.15)/0.5916	1.6903	0.0956	0.05

6 Conclusion

This paper proposes a new architecture of the FRODT based on dynamic mining fuzzy rules. It is different from the traditional tree construction methods that take one single attribute or the combination of several attributes as decision function. In order to eliminate data redundancy and improve classification efficiency, the NRS_FS_FAST algorithm is first introduced. And the AFS theory is adopted to increase semantic interpretation and decrease human subjectivity in selecting membership functions. In the AFS theory framework, the FREA is proposed to dynamically extract fuzzy rules. And in each layer of the tree, the build-up of the FRODT is achieved by developing the only one non-leaf node. Moreover, the genetic algorithm is adopted to optimize the threshold σ that can affect the scale of the tree. Finally, a series of comparative experiments have been carried out to show the superiority of our algorithm.

It should be noted that the main disadvantage of the FRODT is that the FREA extracts only one fuzzy rule for each class. When dealing with large data sets, it is difficult to discover more potential knowledge by leveraging one rule extracted for each class. Therefore, how many rules are extracted for each class is a direction of future research.

Moreover, the application of decision trees is often related to the analysis goal and scenario. For example, financial industry can use decision tree to evaluate loan risk, insurance industry can use decision tree to forecast the promotion of insurance products, medical industry can use decision tree to generate assistant diagnostic disposal model and so on. The decision trees are largely used in all area of real life. Several typical applications on this topic were discussed in [40] for business, in [41] for power systems, in [42] for medical diagnosis, in [43] for intrusion detection, and in [44] for energy modeling. Our paper also deals with the data sets from different application domains, such as the data sets including Wdbc, Heart, Haberman, Column, Breast Cancer, Bupa, Pima Indian and Wpbc in medical diagnosis area, and the data sets including Credit and Australian in business area. How to apply the proposed method to other application scenarios is also our future research direction.

Acknowledgements We are very grateful to all the anonymous editors and reviewers, as well as to all the co-authors for their contributions. Moreover, we would like to acknowledge the National Natural Science Foundation of China (61433004, 61627809, 61621004), and the Liaoning Revitalization Talents Program (XLYC1801005).

Compliance with ethical standards

Conflict of interest The authors declare that there are no financial and personal relationships with other people or organizations that can inappropriately influence our work. And there are no potential conflicts of interest with respect to this work.

References

1. López-Chau A, Cervantes J, López-Garca L, Lamont FG (2013) Fisher's decision tree. Expert Syst Appl 40(16):6283–6291
2. Mirzamomen Z, Kangavari MR (2017) A framework to induce more stable decision trees for pattern classification. Pattern Anal Appl 20(4):991–1004
3. Manwani N, Sastry PS (2012) Geometric decision tree. IEEE Trans Syst Man Cybernet Part B (Cybernet) 42(1):181–192
4. Kevric J, Jukic S, Subasi A (2017) An effective combining classifier approach using tree algorithms for network intrusion detection. Neural Comput Appl 28(1):1051–1058
5. Breiman L (2017) Classification and regression trees. Routledge, Abingdon
6. Azar AT, El-Metwally SM (2013) Decision tree classifiers for automated medical diagnosis. Neural Comput Appl 23(7–8):2387–2403
7. Quinlan JR (2014) C4.5: programs for machine learning. Elsevier, Amsterdam
8. Sok HK, Ooi MP, Kuang YC (2016) Multivariate alternating decision trees. Pattern Recogn 50:195–209
9. Kumar PS, Yung Y, Huan TL (2017) Neural network based decision trees using machine learning for alzheimer's diagnosis. Int J Comput Inf Sci 4(11):63–72
10. Wu CC, Chen YL, Liu YH (2016) Decision tree induction with a constrained number of leaf nodes. Appl Intell 45(3):673–685
11. Shukla SK, Tiwari MK (2012) GA guided cluster based fuzzy decision tree for reactive ion etching modeling: a data mining approach. IEEE Trans Semicond Manuf 25(1):45–56
12. Liu X, Feng X, Pedrycz W (2013) Extraction of fuzzy rules from fuzzy decision trees: an axiomatic fuzzy sets (AFS) approach. Data Knowl Eng 84:1–25
13. Segatori A, Marcelloni F, Pedrycz W (2018) On distributed fuzzy decision trees for big data. IEEE Trans Fuzzy Syst 26(1):174–192
14. Han NM, Hao NC (2016) An algorithm to building a fuzzy decision tree for data classification problem based on the fuzziness intervals matching. J Comput Sci Cybernet 32(4):367–380
15. Sardari S, Eftekhari M, Afsari F (2017) Hesitant fuzzy decision tree approach for highly imbalanced data classification. Appl Soft Comput 61:727–741
16. Tan PJ, Dowe DL (2006) Decision forests with oblique decision trees. In: Mexican international conference on artificial intelligence, Springer, Berlin, Heidelberg, pp 593–603
17. Cantu-Paz E, Kamath C (2003) Inducing oblique decision trees with evolutionary algorithms. IEEE Trans Evol Comput 7(1):54–68
18. Do TN, Lenca P, Lallich S (2015) Classifying many-class high-dimensional fingerprint datasets using random forest of oblique decision trees. Vietnam J Comput Sci 2(1):3–12
19. Barros RC, Jaskowiak PA, Cerri R (2014) A framework for bottom-up induction of oblique decision trees. Neurocomputing 135:3–12
20. Patil SP, Badhe SV (2015) Geometric approach for induction of oblique decision tree. Int J Comput Sci Inf Technol 5(1):197–201
21. Rivera-Lopez R, Canul-Reich J (2017) A global search approach for inducing oblique decision trees using differential evolution. In: Canadian conference on artificial intelligence, Springer, Cham, pp 27–38
22. Wickramarachchi DC, Robertson BL, Reale M et al (2016) HHCART: an oblique decision tree. Comput Stat Data Anal 96:12–23
23. Wang C, Shao M, He Q, Qian Y, Qi Y (2016) Feature subset selection based on fuzzy neighborhood rough sets. Knowl-Based Syst 111:173–179
24. He Q, Xie Z, Hu Q, Wu C (2011) Neighborhood based sample and feature selection for svm classification learning. Neurocomputing 74(10):1585–1594
25. Zhang DW, Wang P, Qiu JQ, Jiang Y (2010) An improved approach to feature selection. In: International conference on machine learning and cybernetics, pp 488–493
26. Liu X (1998) The fuzzy sets and systems based on AFS structure, EI algebra and EII algebra. Fuzzy Sets Syst 95(2):179–188
27. Liu X, Chai T, Wang W, Liu W (2007) Approaches to the representations and logic operations of fuzzy concepts in the framework of axiomatic fuzzy set theory i. Inf Sci 177(4):1007–1026
28. Wang B, Liu XD, Wang LD (2015) Mining fuzzy association rules in the framework of AFS theory. Ann Data Sci 2(3):261–270
29. Menga E, Dan A, Lu J, Liu X (2015) Ranking alternative strategies by SWOT analysis in the framework of the axiomatic fuzzy set theory and the ER approach. J Intell Fuzzy Syst 28(4):1775–1784
30. Burra LR, Poosapati P (2016) A study of notations and illustrations of axiomatic fuzzy set theory. Int J Comput Appl 134(11):7–12
31. Li Z, Duan X, Zhang Q, Wang C, Wang Y, Liu W (2017) Multi-ethnic facial features extraction based on axiomatic fuzzy set theory. Neurocomputing 242:161–177
32. Zadeh LA (1997) Toward a theory of fuzzy information granulation and its centrality in human reasoning and fuzzy logic. Fuzzy Sets Syst 90(2):111–127
33. Agrawal R, Imielinski T, Swami A (1993) Database mining: a performance perspective. IEEE Trans Knowl Data Eng 5(6):914–925
34. Wang X, Liu X, Pedrycz W, Zhu X, Hu G (2012) Mining axiomatic fuzzy set association rules for classification problems. Eur J Oper Res 218(1):202–210
35. Shi H (2007) Best-first decision tree learning. University of Waikato, Hamilton
36. Holmes G, Pfahringer B, Kirkby R, Frank E, Hall M (2002) Multiclass alternating decision trees. Springer, Berlin
37. Kohavi R (1996) Scaling up the accuracy of naive-bayes classifiers:a decision-tree hybrid. In: Second international conference on knowledge discovery and data mining
38. Witten IH, Frank E, Hall MA, Pal CJ (2016) Data mining: practical machine learning tools and techniques. Morgan Kaufmann, Burlington
39. Ar J (2006) Statistical comparisons of classifiers over multiple data sets. J Mach Learn Res 7(1):1–30
40. Creamer G, Freund Y (2010) Using boosting for financial analysis and performance prediction: application to s&p 500 companies, latin american adrs and banks. Comput Econ 36(2):133–151
41. Liu C, Sun K, Rather ZH, Chen Z, Bak CL, Thøgersen P, Lund P (2013) A systematic approach for dynamic security assessment and the corresponding preventive control scheme based on decision trees. IEEE Trans Power Syst 29(2):717–730
42. Al Snousy MB, El-Deeb HM, Badran K, Al Khlil IA (2011) Suite of decision tree-based classification algorithms on cancer gene expression data. Egypt Inform J 12(2):73–82
43. Sindhu SSS, Geetha S, Kannan A (2012) Decision tree based light weight intrusion detection using a wrapper approach. Expert Syst Appl 39(1):129–141
44. Yu Z, Haghighat F, Fung BC, Yoshino H (2010) A decision tree method for building energy demand modeling. Energy Build 42(10):1637–1646