Fusing Fuzzy Monotonic Decision Trees

Jieting Wang, *Student Member, IEEE*, Yuhua Qian*, *Member, IEEE*, Feijiang Li, *Student Member, IEEE*, Jiye Liang and Weiping Ding, *Senior Member, IEEE*

***Abstract*—Ordinal classification is an important classification task, in which there exists a monotonic constraint between features and the decision class. In this paper, we aim at developing a method of fusing ordinal decision trees with fuzzy rough set based attribute reduction. Most of the existing attribute reduction methods for ordinal decision tables are based on the dominance rough set theory or significance measures. However, the crisp dominance relation is difficult in making full use of the information of attribute values; and the reducts based on significance measures are poor in interpretability and may contain unnecessary attributes. In this paper, we firstly define a discernibility matrix with fuzzy dominance rough set. With this discernibility matrix, multiple reducts can be found, which provide multiple complementary feature subspaces with original information. Then diverse ordinal trees can be established from these feature subspaces, and finally, the trees are fused by majority voting. The experimental results show that the proposed fusion method performs significantly better than other fusion methods using dominance rough set or significance measures.**

***Index Terms*—Ordinal classification, ensemble learning, attribute reduction, fuzzy dominance rough set, discernibility matrix**

I. Introduction

FUZZY rough set theory is known as a powerful model for analyzing uncertainty in big data [1]–[5]. In the rough case, a crisp set is provided with a lower approximation and an upper approximation, which allow for a granular representation of knowledge and an excellent description of the uncertain region. In the fuzzy case, relations between objects and sets or relations among objects are characterized by degrees of membership. This allows for great flexibility in dealing with imprecise information. Fuzzy rough set theory combines the advantages of fuzzy sets and rough sets. It can handle uncertainty in nominal or real-valued attributes and has been successfully applied to machine learning, logical reasoning, pattern recognition, intelligent information processing, and other fields [6]–[10].

An important achievement of rough set theory is that of attribute reduction in the decision table. Attribute reduction removes the redundant attributes and preserves the necessary attributes that can maintain the same discrimination information as the original decision table. Great progress has made on attribute reduction in the past few decades. Ślezak [11] reviewed the attribute reduction methods that kept the entropy invariant. Tsang et al. [12] developed a granularity and reduction theory of fuzzy rough sets. Chen et al. [13] proposed a reduction method based on fuzzy rough sets with Gaussian kernel. Wang et al. [14] proposed a task fitting selection model with fuzzy rough sets to guarantee that the membership degree of a sample reaches the maximal value in its class. Ding and Lin et al. proposed a series of efficient reduction methods for large scale databases with fuzzy rough sets [15]–[17].

Ordinal classification problems widely exist in practical scenarios such as credit risk assessment, product performance evaluation, university ranking and so on [18], [19]. In ordinal decision tables, the values of attributes are characterized by preference relationships. In addition, there is a monotonic relationship between the feature attribute and the decision attribute. Generally, the objects with better feature values should not belong to a worse decision class, such as the students with higher scores should have higher grades. Sai and Yao et al. [20] gave the precise definition of general ordinal tables and proposed a framework to transform the ordinal table into a traditional one. Another effective approach to solve ordinal classification is the dominance rough set based methods [21]. Dominance rough set replaces the equivalence relation in the classical rough set with the dominance relations. Based on dominance rough set, Hu et al. [7] defined the monotonic rank information entropy to measure the quality of attributes in the ordinal decision tables, and then proposed a monotonic decision tree REMT to solve ordinal classification problem. Just as CART for traditional classification, REMT plays an important role in obtaining monotone consistent, noise robust and highly interpretable rules.

To further improve the performance of REMT, Qian et al. [8] and Wang et al. [22] employed ensemble learning method. Ensemble learning is one of the most promising methods to improve the generalization performance of learning systems, especially for decision tree systems [23]–[30]. The main idea of ensemble learning is to fuse multiple different classifiers with weak performance. The diversity and accuracy of the base classifiers are two crucial factors in determining the ensemble performance. To guarantee the performance of ensemble, Qian et al. [8] and Wang et al. [22] employed attribute reduction based on dominance rough set to generate multiple feature subsets, and then constructed multiple monotonic decision trees on these subsets. The fusion of the monotonic trees performances better than the individual one. This inspires us that the strategy of attribute reduction can be well applied to ensemble learning. In this paper, we aim to use ensemble learning with attribute reduction technique to solve the ordinal classification problem.

Most of the existing researches for ordinal attribute re-

*: The corresponding author.

J.T. Wang, Y.H. Qian and F.J. Li are with the Institute of Big Data Science and Industry, Shanxi University, Taiyuan 030006, Shanxi Province, China; Y.H. Qian and Jiye Liang are with the Key Laboratory of Computational Intelligence and Chinese Information Processing of Ministry of Education, Shanxi University, Taiyuan 030006, Shanxi Province, China; Weiping Ding is with the School of Information Science and Technology, Nantong University, Nantong 226019, China. (e-mail: jietingwang@email.sxu.edu.cn; jinchengqyh@126.com; feijiangli@email.sxu.edu.cn; ljy@sxu.edu.cn; dwp9988@163.com.)

duction are based on the dominance rough set theory. Qian et al. [8] defined a variable monotonic dependency measure based on dominance relation to conduct attribute reduction. Wang et al. [22] defined a discernibility matrix preserving the monotonic consistency based on dominance rough set theory. However, the dominance rough set is an extension of the classical rough set theory. It inherits the limitation of the rough set theory: poor ability to deal with real-valued attributes, such as having a bias to the attributes with more values. Fuzzy dominance rough set theory uses fuzzy dominance relation instead of crisp dominance relation to measure the relationship between objects. Fuzzy dominance relation can not only get a preference relation between attribute values but also get a preference degree [31]. In the area of fuzzy rough set, Hu et al. [32] defined three significance measures and then used the forward heuristic algorithm for ordinal attribute reduction. However, a reduction method based on the heuristic algorithms may contain unnecessary attributes [33]. Besides, the heuristic algorithms usually get one reduct and the result is easily affected by the reading order of attributes.

In addition to the reduction methods based on the significance measures, Skowron and Rauszer put forward a reduction method with a discernibility matrix and a discernibility function [34]–[36]. Each element of the discernibility matrix stores an attribute subset which distinguishes a pair of objects in some specified sense. The discernibility function performs conjunction and disjunction operation on the elements of the discernibility matrix to get multiple attribute subsets. The reduction result based on the discernibility function is usually called as a complete reduct because this kind of result contains all possible attribute subsets defined in the discernibility matrix. In addition, the attribute subsets generated by the discernibility matrix maintain the discrimination information of the original attribute set from different dimensions, which provides a good foundation for establishing multiple classifiers with certain accuracy and diversity. Thus, the result of attribute reduction based on a discernibility matrix is more suitable to the requirements of ensemble learning than that based on the significance measures.

According to the above analysis, we focus on using ensemble learning with discernibility-matrix-based attribute reduction technique to improve the performance of ordinal classifiers. Particularly, we firstly propose the definition of upward and downward local positive region of the ordinal decision table. Based on the definition of lower approximation in [12], we propose a discernibility matrix that preserves the upward local positive region invariant. Then, based on the discernibility matrix, multiple reducts can be obtained. Finally, each attribute subset is used to construct a fuzzy rank entropy monotonic decision tree, and all of them are fused to obtain the final predicted results.

The rest of the paper is organized as follows. In Section II, we introduce the definitions of rough set theory and present the existing methods of attribute reduction for ordinal decision table. In Section III, we give the definition of discernibility matrix preserving the upward local positive region invariant. We also present the algorithms for fusing fuzzy monotonic decision trees in this section. The experimental results and analyses are presented in Section IV. In the end, we conclude and put forward the future work in Section V.

II. Preliminaries

In this section, we give the fundamental definitions in dominance rough set theory and introduce some discernibility matrixes and some significance measures, which were proposed for attribute reduction for ordinal decision table.

A. Dominance Based Rough Sets

Let $DT = (U, A \cup \{d\})$ be an ordinal decision table, where $U = \{x_1, x_2, \cdots, x_n\}$ is a set of objects, $A = \{a_1, a_2, \cdots, a_m\}$ is a set of feature attributes and d is the decision class with values $\{d_1, d_2, ..., d_K\}$. Let $v(x, a)$ be the value of x with respect to the attribute $a \in A$, we consider that $v(x, a) \in \mathcal{Z}$ or $v(x, a) \in \mathcal{R}$, where $\mathcal{Z}$ or $\mathcal{R}$ are the integer or real number area, respectively. Without loss of generality, we assume that $d_1 \leqslant d_2 \leqslant ... \leqslant d_K$. Preference relations exist in decision table: $\geqslant_a$, $\geqslant_d$, $\leqslant_a$ and $\leqslant_d$, which signify the relation of no worse than or no better than with respect to a or d, respectively. For a subset of attributes $B \subseteq A$, we define $x \geqslant_B y$ as $x \geqslant_a y$ for all $a \in B$. We say DT is monotone increasing consistent with respect to B, if $\forall x, y \in U$, $B \subseteq A$, $x \geqslant_B y$, then $x \geqslant_d y$; otherwise, DT is monotone increasing inconsistent with respect to B. Based on the preference relations, the x-dominated or x-dominating sets in terms of D and B are defined. For $x, y \in U$, $d \subseteq D$, $a \in A$, $B \subseteq A$,

$$[x]_D^{\geqslant} = \{y : y \geqslant_d x\}, \quad [x]_D^{\leqslant} = \{y : y \leqslant_d x\}, \tag{1}$$

$$[x]_B^{\geqslant} = \{y : y \geqslant_B x\}, \quad [x]_B^{\leqslant} = \{y : y \leqslant_B x\}. \tag{2}$$

For decision d_i, let $d_i^{\geqslant} = \cup_{j \geqslant i}^{K} d_j$ and $d_i^{\leqslant} = \cup_{j \leqslant i}^{K} d_j$, where d_j is the set of objects with decision d_j. We have that $d_i^{\geqslant} \supseteq d_j^{\geqslant}$ and $d_i^{\leqslant} \subseteq d_j^{\leqslant}$ for $i \leqslant j$. The lower approximations of $d_i^{\geqslant}, d_i^{\leqslant} \subseteq U$ with respect to attribute subset B are defined as:

$$\underline{B}^{\geqslant} d_i^{\geqslant} = \{y : [x]_B^{\geqslant} \subseteq d_i^{\geqslant}\}, \quad \underline{B}^{\leqslant} d_i^{\leqslant} = \{y : [x]_B^{\leqslant} \subseteq d_i^{\leqslant}\}. \tag{3}$$

Based on the dominance rough set theory, Qian et al. [8] presented a discernibility matrix for monotone increasing consistency decision table:

Definition 1: A monotone consistent discernibility matrix is defined as:

$$c_{ij} = \begin{cases} \{a \in \mathbf{A} : x_j \in [x_i]_a^{\geqslant}\}, & \text{if} \quad x_j \in [x_i]_D^{\geqslant}; \\ \varnothing, & \text{otherwise.} \end{cases} \tag{4}$$

Definition 2: The discernibility function with respect to a discernibility matrix $M_A(D) = \{c_{ij}\}$ is defined as:

$$f(M_A(D)) = \wedge\{\vee(c_{ij}) | \forall i, j = 1, ..., n, c_{ij} \neq \varnothing\}. \tag{5}$$

where $\vee$ and $\wedge$ are the disjunction and conjunction operator, respectively.

By transforming the disjunction norm form into the conjunction norm form through the absorption law and the distribution law, one can obtain multiple reducts, that is, the terms of the conjunction norm form.

Considering the influence of noisy objects, Qian et al. [8] also proposed β-variable ($\beta \in [0, 0.5]$) upward lower and upper approximations of the decision class:

$$\underline{R_B^{\geqslant}}(d_i^{\geqslant}) = \{x \in U \mid \frac{|[x]_B^{\leqslant} \cap d_i^{\geqslant}|}{|[x]_B^{\leqslant}|} \geqslant 1-\beta\}, \tag{6}$$

$$\overline{R_B^{\geqslant}}(d_i^{\geqslant}) = \{x \in U \mid \frac{|[x]_B^{\leqslant} \cap d_i^{\geqslant}|}{|[x]_B^{\leqslant}|} \geqslant \beta\}, \tag{7}$$

where $|S|$ is the cardinality of the set S. Based on these, the significance measures of the attribute a w.r.t B are defined as:

$$Sig_{inner}^{\beta}(a, B, d) = \gamma_B^{\beta}(d) - \gamma_{B-\{a\}}^{\beta}(d), \tag{8}$$

$$Sig_{outer}^{\beta}(a, B, d) = \gamma_{B\cup\{a\}}^{\beta}(d) - \gamma_B^{\beta}(d), \tag{9}$$

where $\gamma_B^{\beta}(d)$ is defined as:

$$\gamma_B^{\beta}(d) = \frac{\left|U - \bigcup_{i=1}^{K} \left(\overline{R_B^{\geqslant}}(d_i^{\geqslant}) - \underline{R_B^{\geqslant}}(d_i^{\geqslant})\right)\right|}{|U|}.$$

Actually, $\gamma_B^{\beta}(d)$ measures the monotonic consistency between B and d, and the parameter β reflects the tolerance degree to noise, so as to control the relaxation degree of the model.

B. Fuzzy Dominance Based Rough Sets

Let $DT = (U, A \cup \{d\})$ be an ordinal decision table. We define $U/d^{\geqslant} = \{D_i^{\geqslant}, i = 1, 2, ..., K\}$ and $U/d^{\leqslant} = \{D_i^{\leqslant}, i = 1, 2, ..., K\}$, where $D_i^{\geqslant}$ or $D_i^{\leqslant}$ is the membership function of set $d_i^{\geqslant}$ or $d_i^{\leqslant}$ satisfying $D_i^{\geqslant}(x) \supseteq D_j^{\geqslant}(x)$ or $D_i^{\leqslant}(x) \subseteq D_j^{\leqslant}(x)$ for $i \leqslant j$, respectively. We introduce the sigmoid function to measure the fuzzy preference degree [32]. For $x, y \in U$, the fuzzy dominating relation and fuzzy dominated relation between x and y in terms of a are defined as:

$$R_a^{>}(x, y) = \frac{1}{1 + e^{-k(v(x,a)-v(y,a))}}, \tag{10}$$

$$R_a^{<}(x, y) = \frac{1}{1 + e^{k(v(x,a)-v(y,a))}}. \tag{11}$$

It is easy to observe that $R_a^{>}(x, y) + R_a^{>}(y, x) = 1$. $R_a^{>}(x, y)$ measures the preference degree of x over y: $R_a^{>}(x, y) = \frac{1}{2}$ indicates $v(a, x) = v(a, y)$, $R_a^{>}(x, y) > \frac{1}{2}$ shows $v(a, x) > v(a, y)$, and $R_a^{>}(x, y) < \frac{1}{2}$ shows $v(a, x) < v(a, y)$. The fuzzy dominance relation based on the sigmoid function does not satisfy the properties of reflexivity and symmetry, but it is sup-min transitive. That is:

Property 1: Let $R_a^{>}$ be a fuzzy dominance relation based on sigmoid function, then

(1) $R_a^{>}(x, x) \neq 1$;

(2) $R_a^{>}(x, y) \neq R_a^{>}(y, x)$;

(3) $\sup_z \min\{R_a^{>}(x, z), R_a^{>}(z, y)\} \leqslant R_a^{>}(x, y)$.

The properties are also applied to $R_a^{<}$.

The fuzzy dominance relation degrades into the dominance relation when $k \to \infty$. If $R_a^{>}(x, y)$ and $R_a^{<}(x, y)$ are not participated into the calculation process of learning models, the value of k dose not make any difference. The reason is that they are just used to compare relatively in the same dimension.

Based on $R_a^{>}$, the fuzzy dominating class and fuzzy dominated class for x in terms of a are defined as:

$$\widetilde{[x]}_a^{>} = \frac{R_a^{>}(x_1, x)}{x_1} + \frac{R_a^{>}(x_2, x)}{x_2} + ... + \frac{R_a^{>}(x_n, x)}{x_n}, \tag{12}$$

$$\widetilde{[x]}_a^{<} = \frac{R_a^{<}(x_1, x)}{x_1} + \frac{R_a^{<}(x_2, x)}{x_2} + ... + \frac{R_a^{<}(x_n, x)}{x_n}. \tag{13}$$

If the intersection operator is used to integrate the preference relations generated by multiple attributes, then for $B \subseteq A$, we have:

$$R_B^{>}(x, y) = \cap_{a\in B} R_a^{>}(x, y) = \min_{a\in B} R_a^{>}(x, y), \tag{14}$$

$$R_B^{<}(x, y) = \cap_{a\in B} R_a^{<}(x, y) = \min_{a\in B} R_a^{<}(x, y). \tag{15}$$

Obviously, $R_B^{>}(x, y) \geqslant R_A^{>}(x, y)$ and $R_B^{<}(x, y) \geqslant R_A^{<}(x, y)$.

The upward and downward fuzzy lower approximation in terms of a are defined as:

$$\underline{R_a^{>}}(D_i^{\geqslant})(x) = \inf_{z\in U} \max\{1 - R_a^{>}(z, x), D_i^{\geqslant}(z)\}, \tag{16}$$

$$\underline{R_a^{<}}(D_i^{\leqslant})(x) = \inf_{z\in U} \max\{1 - R_a^{<}(z, x), D_i^{\leqslant}(z)\}. \tag{17}$$

Specifically, if $D_i^{\geqslant}$ is the crisp set with $D_i^{\geqslant}(y) = 1$ for $y \in d_i^{\geqslant}$ and otherwise $D_i^{\geqslant}(y) = 0$. then the upward fuzzy lower approximation is degraded into:

$$\underline{R_a^{>}}(D_i^{\geqslant})(x) = \inf_{z\notin d_i^{\geqslant}} 1 - R_a^{>}(z, x). \tag{18}$$

In this case, the membership of x to the lower approximation of $d_i^{\geqslant}$ is larger if x is better than the greatest objects from the inferior classes. The larger the membership degree, the greater the dependence of decision attribute on feature attribute or the larger the consistency degree between them. In this paper, we consider the crisp ordinal decision class.

With fuzzy rough set, Hu et al. [32] considered the summation of all objects' lower approximation degrees to all decision classes as the approximation quality of feature attributes to the decision attribute. Based on the idea, the upward, downward and global significance of attribute a relative to B are defined as Eq. 19-Eq.21. Then incorporating the significance measures into Algorithm 1, an upward reduct, a downward reduct and a global reduct can be obtained. However, the summation operator is not robust to noise objects, and Algorithm 1 usually gets a reduct with some unnecessary attributes [33].

C. FREMT: Fuzzy Rank Entropy based Monotonic tree

For ordinal classification, ranking mutual information (RMI) and fuzzy ranking mutual information (FRMI) are proposed in [37]. Both RMI and FRMI are important measures of attributes, which can be used as heuristic criterions to evaluate and to select features.

Definition 3: Given $DT = (U, A \cup D)$, where $D = \{d\}$, for $B \subseteq A \cup D$, the upward ranking mutual information of the set U in terms of B and D is defined as:

$$RMI^{>}(B, D) = - \frac{1}{|U|} \sum_{i=1}^{n} \log \frac{\left|[x_i]_B^{>}\right| \times \left|[x_i]_d^{>}\right|}{|U| \times \left|[x_i]_B^{>} \cap [x_i]_d^{>}\right|}. \tag{22}$$

$$sig^{>}(a,B,D)=\frac{\sum_i\sum_{x\in d_i^{\geqslant}}\left(\underline{R^{>}_{\{B\cup a\}}}D_i^{\geqslant}-\underline{R^{>}_{B}}D_i^{\geqslant}\right)}{\sum_i|d_i^{\geqslant}|}, \tag{19}$$

$$sig^{<}(a,B,D)=\frac{\sum_i\sum_{x\in d_i^{\leqslant}}\left(\underline{R^{<}_{\{B\cup a\}}}D_i^{\leqslant}-\underline{R^{<}_{B}}D_i^{\leqslant}\right)}{\sum_i|d_i^{\leqslant}|}, \tag{20}$$

$$sig(a,B,D)=\frac{\sum_i|d_i^{\geqslant}|sig^{>}(a,B,D)+\sum_i|d_i^{\leqslant}|sig^{<}(a,B,D)}{\sum_i|d_i^{\geqslant}+\sum_i|d_i^{\leqslant}|}. \tag{21}$$

Algorithm 1 Forward greedy search for a reduct

Require: Ordinal decision table $DT=(U,A\cup\{d\})$
Ensure: An ordinal reduct of DT: B
1: $B=\varnothing$
2: **for** each feature $a_i\in A-B$, **do**
3: Compute significance degree of attribute a by significance measures $SIG(a_i,B,D)$ \\ such as Eq. 7, Eq.8 and Eq. 19-Eq.21
4: **end for**
5: $SIG(a_k,B,D)=\max_i SIG(a_i,B,D)$
6: **if** $SIG(a_k,B,D)>0$ **then**
7: $B=B\cup a_k$;
8: Go to Step 2;
9: **end if**
10: END

Definition 4: Given $DT=(U,A\cup D)$, where $D=\{d\}$, for $B\subseteq A\cup D$, the upward fuzzy ranking mutual information of the set U in terms of B and D is defined as:

$$FRMI^{>}(B,D)= \tag{23}$$
$$-\frac{1}{|U|}\sum_{i=1}^{n}\log\frac{|\widetilde{[x_i]}_B^{>}|\times|\widetilde{[x_i]}_d^{>}|}{|U|\times|\widetilde{[x_i]}_B^{>}\cap\widetilde{[x_i]}_d^{>}|}.$$

A monotonic decision tree based on RMI, the so-called REMT, is established in [7]. REMT is capable to capture the monotonic structure in ordinal classification. Replacing RMI with FRMI in REMT, we obtain the fuzzy rank entropy based monotonic decision tree (FREMT), which is suitable for ordinal classification and captures more information than REMT. The algorithm of FREMT is shown as Algorithm 2. FREMT is a binary tree. The structure of FREMT is the same as that of the famous CART algorithm, while the difference is that CART adopts the Gini index as the heuristic criterion.

Algorithm 2 FREMT

Require: Ordinal decision table $OD=(U,A\cup d)$; stoping criterion of FREMT: ε
Ensure: A monotonic decision tree T.
1: If the number of objects is 1 or all objects are from the same class, then the branch stops growing.
2: otherwise,
3: **for** each feature a_i, **do**
4: **for** each $c_j\in V_{a_i}$ (V_{a_i} is the domain of value of a_i), **do**
5: Divide objects into two subsets according to c_j,
6: **if** $f(a_i,x)\leq c_j$ **then**
7: Put x into one subset, and set $f(a_i,x)=1$,
8: **else**
9: Put x into the other subset, and set $f(a_i,x)=2$.
10: **end if**
11: Denote the binarized feature in terms of a_i and c_j as $a_i(c_j)$ and compute $FRMI_{c_j}=FRMI^{\geq}(\{a_i(c_j)\},\{d\})$.
12: **end for** j
13: $c_j^*=arg\max_j FRMI_{c_j}$.
14: **end for** i
15: Select the best feature a^* and the corresponding point c^*: $(a^*,c^*)=arg\max_i\max_j FRMI^{\geq}(\{a_i(c_j)\},\{d\})$.
16: If $FRMI^{\geq}(\{a^*\},\{d\})<\varepsilon$, then stop.
17: Build a new node and split objects with a^* and c*.
18: Recursively produce new splits according to the above procedure until stopping criterion is satisfied.

III. Fusing Fuzzy Monotonic Decision Trees Using Fuzzy Dominance Rough Set

In this section, we define a discernibility matrix for ordinal decision table using fuzzy dominance rough set and present the fusion algorithm for ordinal classification.

A. Discernibility Matrix with Fuzzy Dominance Rough Set

Although some significance measures of attributes have been proposed, the reduction methods with them are based on the forward heuristic algorithm. Generally, the heuristic reduction algorithm can only get a reduct and is easily affected by the reading order of attributes. To get multiple reducts, researchers often permute the attributes [38]. This strategy makes the relation of the multiple attributes subsets unclear and lack of interpretability. In this subsection, we define a discernibility matrix for ordinal attribute reduction. The subsets of attributes based on the discernibility matrix are complete and diverse.

In the classical rough set, the union of the lower approximation to each decision class, the so-called positive region, is often used to define the discernibility matrix. However, in the ordinal decision table, there is an inclusion relationship between the hierarchical decision classes ($d_i^{\geqslant}$ or $d_i^{\leqslant}$), which leads to the union of the lower approximations is equal to the lower approximation of the universal set. Firstly, we give the definition of the local positive region, then based on

the definition of lower approximation in [12], we obtain a discernibility matrix.

Definition 5: For each object x, the upward local positive region with respect to $R_A^{\geqslant}$ and and the downward local positive region with respect to $R_A^{\leqslant}$ are defined as:

$$LPos_{R_A^{\geqslant}}(D)(x) = \cup_{D_i^{\geqslant} \in U/d^{\geqslant}, d_i \geqslant d_x} \underline{R_A^{\geqslant}}(D_i^{\geqslant})(x), \quad (24)$$

$$LPos_{R_A^{\leqslant}}(D)(x) = \cup_{D_i^{\leqslant} \in U/d^{\leqslant}, d_i \leqslant d_x} \underline{R_A^{\leqslant}}(D_i^{\leqslant})(x). \quad (25)$$

By definition, we have that $LPos_{R_A^{\geqslant}}(D)(x)$ takes its maximum at $\underline{R_A^{\geqslant}}(D_x^{\geqslant})(x)$ and $LPos_{R_A^{\leqslant}}(D)(x)$ takes its maximum at $\underline{R_A^{\leqslant}}(D_x^{\leqslant})(x)$:

$$LPos_{R_A^{\geqslant}}(D)(x) = \underline{R_A^{\geqslant}}(D_x^{\geqslant})(x), \quad (26)$$

$$LPos_{R_A^{\leqslant}}(D)(x) = \underline{R_A^{\leqslant}}(D_x^{\leqslant})(x). \quad (27)$$

This is consistent with the situation in the fuzzy rough set that the positive region of an object is equal to the lower approximation of its own decision class [12], [14]. Compared with taking the summation over the lower approximations as a significance measure, considering each object's lower approximation is expected to be robust to noisy object.

Next, we will introduce another way to define the lower approximation, which is helpful to find the discernibility attributes for each pair of objects. Without loss of generality, we only use the upward relation in the following.

Definition 6: The fuzzy x-dominated class in terms of $R_a^{\geqslant}$ is defined as:

$$(x_\lambda)_{R_a^{\geqslant}}(z) = \begin{cases} 0, & 1 - R_a^{\geqslant}(z,x) \geqslant \lambda, \\ \lambda, & 1 - R_a^{\geqslant}(z,x) < \lambda. \end{cases} \quad (28)$$

where $\lambda \in (0,1]$.

In fact, $(x_\lambda)_{R_a^{\geqslant}}$ is a λ-level cut set with respect to $1 - R_a^{\geqslant}$. It satisfies the following properties:

1) $(x_{\lambda_1})_{R_a^{\geqslant}} \subseteq (x_{\lambda_2})_{R_a^{\geqslant}}$ if $\lambda_1 < \lambda_2$;
2) $1 - R_A^{\geqslant}(z,x) \geqslant 1 - R_B^{\geqslant}(z,x)$ and $(x_\lambda)_{R_A^{\geqslant}} \subseteq (x_\lambda)_{R_B^{\geqslant}}$ if $B \subseteq A$.

Theorem 1: The lower approximation to $D_i^{\geqslant}$ in terms of $R_a^{\geqslant}$ is defined as:

$$\underline{R_a^{\geqslant}}(D_i^{\geqslant}) = \cup_\lambda \{(x_\lambda)_{R_a^{\geqslant}} : (x_\lambda)_{R_a^{\geqslant}} \subseteq D_i^{\geqslant}, \lambda \in (0,1]\}. \quad (29)$$

Proof: Suppose $\underline{R_a^{\geqslant}}(D_i^{\geqslant})(z) = \lambda^*$, then we prove that $\cup_\lambda \{(x_\lambda)_{R_a^{\geqslant}} : (x_\lambda)_{R_a^{\geqslant}} \subseteq D_i^{\geqslant}, \lambda \in (0,1]\} = \lambda^*$. By definition, we have that $\forall z \in d_i^{<}$, $1 - R_a^{\geqslant}(z,x) \geqslant \lambda^*$. Besides,

$$(x_\lambda)_{R_a^{\geqslant}} \subseteq D_i^{\geqslant} \Leftrightarrow \forall z \in d_i^{<}, (x_\lambda)_{R_a^{\geqslant}}(z) = 0 \quad (30)$$

$$\Leftrightarrow \forall z \in d_i^{<}, 1 - R_a^{\geqslant}(z,x) \geqslant \lambda. \quad (31)$$

Hence, $(x_\lambda)_{R_a^{\geqslant}} \subseteq D_i^{\geqslant}$ requires that $\lambda \leqslant \lambda^*$. Thus, for $(x_\lambda)_{R_a^{\geqslant}} \subseteq D_i^{\geqslant}$, $\cup_\lambda (x_\lambda)_{R_a^{\geqslant}} = (x_{\lambda^*})_{R_a^{\geqslant}}$.

In rough set and fuzzy rough set, the lower approximation of a set can be represented by the union of multiple equivalence classes or fuzzy equivalence classes contained in the set and different equivalence classes are mutually disjoint [12]. In dominance rough set, the lower approximation can also be expressed as an union of fuzzy classes, while $(x_\lambda)_{R_a^{\geqslant}} \neq (y_\lambda)_{R_a^{\geqslant}} \Rightarrow (x_\lambda)_{R_a^{\geqslant}} \cap (y_\lambda)_{R_a^{\geqslant}} = \varnothing$ is not established. The reason is that the transitivity of fuzzy relation is directional.

Definition 7: For a fuzzy ordinal decision table $DT = (U, A \cup \{d\})$, $B \subseteq A$ is a local positive region reduct relative to A if it satisfies:

$$1) \quad LPos_{R_B^{\geqslant}} D = LPos_{R_A^{\geqslant}} D; \quad (32)$$

$$2) \quad \forall a \in B, LPos_{R_{\{B-a\}}^{\geqslant}} D \neq LPos_{R_B^{\geqslant}} D. \quad (33)$$

Theorem 2: Suppose $B \subseteq A$ is a local positive region reduct of fuzzy ordinal decision table if and only if for each x, B satisfies $(x_\lambda)_{R_B^{\geqslant}} \subseteq D_x^{\geqslant}$, $\lambda = \underline{R_A^{\geqslant}}(D_x^{\geqslant})(x)$.

Proof: $\Leftarrow$: It is clear by using $(x_\lambda)_{R_A^{\geqslant}} \subseteq (x_\lambda)_{R_B^{\geqslant}}$.

Theorem 3: Suppose $B \subseteq A$ is a local positive region reduct of fuzzy ordinal decision table if and only if for every x with $\lambda = \underline{R_A^{\geqslant}}(D_x^{\geqslant})(x)$, when $y \in D_x^{<}$, we have $1 - R_B^{\geqslant}(y,x) \geqslant \lambda$.

Proof: If $B \subseteq A$ is a local positive region reduct of fuzzy ordinal decision table, then for $\lambda = \underline{R_A^{\geqslant}}(D_x^{\geqslant})(x)$, we have

$$\forall x \in U, \quad (x_\lambda)_{R_B^{\geqslant}} \subseteq D_x^{\geqslant}, \quad (34)$$

$$\Leftrightarrow \quad \text{for} \quad y \in D_x^{<}, \quad (x_\lambda)_{R_B^{\geqslant}}(y) = 0, \quad (35)$$

$$\Leftrightarrow \quad \text{for} \quad y \in D_x^{<}, \quad 1 - R_B^{\geqslant}(y,x) \geqslant \lambda. \quad (36)$$

According to the analysis, we can define the discernibility matrix $M_A(D) = (c_{ij})_{n \times n}$ as follows:

Definition 8: A discernibility matrix keeping the local positive region invariant is defined as:

$$c_{ij} = \begin{cases} \{a \in A : 1 - R_a^{\geqslant}(x_j, x_i) \geqslant \lambda_i\}, & \text{if} \quad x_j \in D_{x_i}^{<}; \\ \varnothing, & \text{otherwise}, \end{cases} \quad (37)$$

where $\lambda_i = \underline{R_A^{\geqslant}}(D_{x_i}^{\geqslant})(x_i)$.

The above discernibility matrix makes sure of keeping the local positive region of each object invariant, because for $\forall a \in c_{ij}$ satisfies $1 - R_a^{\geqslant}(x_j, x_i) \geqslant \lambda_i$, the final reduction B by discernibility function also satisfies $1 - R_B^{\geqslant}(x_j, x_i) = \max_{a \in B} 1 - R_a^{\geqslant}(x_j, x_i) \geqslant \lambda_i$, which meets the requirement of Theorem 3.

Next, an example is employed for showing the efficiency of the proposed reduction method. Table I is an ordinal decision table selected from the Bankruptcy risk dataset. Bankruptcy risk records the experience of a Greek industrial development bank financing industrial and commercial firms. Suppose $U = \{x_1, x_2, x_3, x_4, x_4, x_5, x_6\}$ is a set of objects, each object is characterize by 12 ordinal feature attributes. The decision class is shown in the last row.

TABLE I: Ordinal decision table from Bankruptcy risk

U/A	1	2	3	4	5	6	7	8	9	10	11	12	D
x_1	3	2	1	1	1	4	4	4	4	4	2	4	3
x_2	2	1	3	1	1	3	5	2	4	2	1	3	3
x_3	2	1	1	1	1	3	2	2	4	4	2	3	2
x_4	2	1	2	1	1	2	4	3	3	2	1	2	2
x_5	3	2	1	1	1	1	3	3	3	4	3	4	1
x_6	3	1	1	1	1	1	2	2	3	4	3	4	1

Based on Definition 1, the obtained discernibility matrix is shown in Table II. According to the discernibility function Eq. (4), the reduction result is

TABLE II: Discernibility matrix based on Definition 1

U^2	x_1	x_2	x_3	x_4	x_5	x_6
x_1	$\emptyset$	$\emptyset$	$\{1,2,6,7,8,12\}$	$\{1,2,6,8,9,10,11,12\}$	$\{6,7,8,9\}$	$\{2,6,7,8,9\}$
x_2	$\emptyset$	$\emptyset$	$\{3,7\}$	$\{3,6,7,9,12\}$	$\{3,6,7,9\}$	$\{3,6,7,9\}$
x_3	$\emptyset$	$\emptyset$	$\emptyset$	$\emptyset$	$\{6,9\}$	$\{6,9\}$
x_4	$\emptyset$	$\emptyset$	$\emptyset$	$\emptyset$	$\{3,6,7\}$	$\{3,6,7,8\}$
x_5	$\emptyset$	$\emptyset$	$\emptyset$	$\emptyset$	$\emptyset$	$\emptyset$
x_6	$\emptyset$	$\emptyset$	$\emptyset$	$\emptyset$	$\emptyset$	$\emptyset$

TABLE III: Discernibility matrix based on Definition 8

U^2	x_1	x_2	x_3	x_4	x_5	x_6
x_1	$\emptyset$	$\emptyset$	$\{7,8\}$	$\{6,10,12\}$	$\{6\}$	$\{6,7,8\}$
x_2	$\emptyset$	$\emptyset$	$\{3,7\}$	$\{3,6,7,9,12\}$	$\{3,6,7,9\}$	$\{3,6,7,9\}$
x_3	$\emptyset$	$\emptyset$	$\emptyset$	$\emptyset$	$\{6\}$	$\{6\}$
x_4	$\emptyset$	$\emptyset$	$\emptyset$	$\emptyset$	$\{3,6,7\}$	$\{3,6,7,8\}$
x_5	$\emptyset$	$\emptyset$	$\emptyset$	$\emptyset$	$\emptyset$	$\emptyset$
x_6	$\emptyset$	$\emptyset$	$\emptyset$	$\emptyset$	$\emptyset$	$\emptyset$

$\{\{7,9\},\{6,7\},\{3,9,12\},\{3,8,9\},\{3,6\},\{2,3,9\},\{1,3,9\}\}$. The less monotone consistency attributes $\{1\}$ and $\{12\}$ are included in the result by sample pair (x_1,x_3), (x_1,x_4) and (x_2,x_4). While in the proposed discernibility matrix (Definition 8), each attribute is judged by λ_i, which is the value with respect to the most monotone consistent attribute:

$$\lambda_i = \underline{R_A^{>}}(D_{x_i}^{\geqslant})(x_i) = \inf_{z\in d_{x_i}^{<}} \max_{a\in A} 1 - R_a^{>}(z,x_i). \tag{38}$$

Thus, the attributes will make a comparison with each other before selection. Our discernibility matrix is shown in Table III and the reduction result is $\{\{6,7\},\{3,7,8\}\}$, shown as Eq. (39). This signifies that the proposed discernibility matrix makes full use of the global information of attributes.

Besides, the reduction result of Algorithm 1 with Eq. (19) as the significance measure is $\{6,1,7\}$. Obviously, the attribute $\{1\}$ should not be included in. However, after $\{6\}$ is selected into the reduction set, $\{1\}$ increases the membership degree of x_5 and x_6 to $d_3^{\geqslant}$, which leads to the increase of the total degree. This signifies that the summation operator is not robust to the noise objects, which may produce an inappropriate reduct.

B. A More General Discernibility Matrix for Ordinal Classification

Because the preference relationship induced by an attribute set is determined by the intersection of the relationships induced by its elements, as shown as Eq. (15) and Eq. (16), the local lower approximation for each object λ_i is the distance in terms of the most discriminative attribute with the best objects from its inferior class. Thus in Definition 8, the attributes which are not inferior to the best attribute can be selected into the discernibility matrix. This requirement may be strict for real-world classification problems. For classification, even if an attribute produces a small gap for objects, it may provide some discrimination information. To contain the weak monotonicity of attributes and the inconsistency of decision table, we use the quantile operator $1-p$ instead of the intersection operator to integrate the multiple attributes. Thus, the maximization operator in λ_i is generalized to the p-th maximization.

Definition 9: A generalized discernibility matrix for ordinal classification is defined as:

$$c_{ij} = \begin{cases} \{a \in A : 1 - R_a^{>}(x_j,x_i) \geqslant \lambda_i\}, & \text{if} \quad x_j \in D_{x_i}^{<}; \\ \varnothing, & \text{otherwise}, \end{cases} \tag{40}$$

where $\lambda_i = \inf_{z\in d_{x_i}^{<}} [1 - R_a^{>}(z,x_i), a \in A]_p$ and $[\cdot]_p$ denotes the p-th maximal value in set.

In practice, we use Definition 9 to build the discernibility matrix. The parameter p controls the number of attributes selected into c_{ij}, and thus influences the number of attributes in the reducts. When p value is 1, Definition 9 is degraded to Definition 8. This is the most rigorous way to filter attributes. In this case, the number of attributes in c_{ij} is less, and the probability that the intersection of c_{ij} is empty is larger, thus the probability that the reducts have more attributes is larger. When the p value approaches the number of original features, all the attributes will be selected for each pair of objects, and the reducts are singleton sets.

C. Algorithms for Fusing Complete Fuzzy Monotonic Decision Trees

After building the discernibility matrix, we will find reducts based on it. Skowron and Rauszer proposed a reduction method with a discernibility function [34], which can find all the reducts simultaneously. The essential of this method is the laws of absorption and distribution. Here, we present efficient algorithms to implement the laws of absorption and distribution. Next, following a discernibility matrix simplification methodology for constructing attribute reducts proposed in [39], we offer an efficient procedure to find a reduct based on the discernibility matrix. Finally, we give our fusion algorithm.

Based on the absorption law in discernibility function, only the minimum elements $M_A^*(D)$ that cannot be contained by other elements in the discernibility matrix are sufficient and necessary to find the final reduction [40]. For example, in Table II, only $\{7 \vee 8\}$, $\{6\}$ and $\{3 \vee 7\}$ are required. Putting all the elements into the discernibility function definitely increases the time complexity and the computational load.

To compress discernibility matrix, Chen et.al. proposed the SPS method [41], which selects the sample pairs that determine the minimum elements. The SPS is a point-based algorithm and its time complexity is $|U|^4|A|$. Here, we present a vector-based algorithm to implement the absorption law. Firstly, we represent the inverse of the discernibility matrix

$$\begin{aligned} f(M_A(D)) =& \{7 \vee 8\} \wedge \{6 \vee 10 \vee 12\} \wedge \{6\} \wedge \{6 \vee 7 \vee 8\} \wedge \{3 \vee 7\} \wedge \{3 \vee 6 \vee 7 \vee 9 \vee 12\} \\ & \wedge \{3 \vee 6 \vee 7 \vee 9\} \wedge \{3 \vee 6 \vee 7\} \wedge \{3 \vee 6 \vee 7 \vee 8\} \\ =& \{7 \vee 8\} \wedge \{6\} \wedge \{3 \vee 7\} = \{6 \wedge 7\} \vee \{3 \wedge 6 \wedge 8\}. \end{aligned} \tag{39}$$

as a binary matrix. Here, the inverse refers to the De Morgan law. For example, the inverse of $\{6 \vee 7\} \wedge \{3 \vee 6 \vee 8\}$ is $\{6 \wedge 7\} \vee \{3 \wedge 6 \wedge 8\}$. Let the rows and the columns in the binary matrix correspond to the disjunction operator $\vee$ and the conjunction operator $\wedge$, respectively. Thus the inverse of $M_A(D)$ is represented as:

$$\overline{M}_A(D)_{ij} = \begin{cases} 1, \text{the } i\text{th term contains the } j\text{th attribute;} \\ 0, \text{otherwise.} \end{cases}$$

$\overline{M}_A(D)$ is a matrix with $|A|$ columns and at most $|U|^2$ rows because there are empty sets in the discernibility matrix. Algorithm 3 is designed to execute the absorption law by deleting the other rows that contain the current one. The time complexity of Algorithm 3 is $O(L^*L|A|)$, where L^* and L are the numbers of rows in $\overline{M}^*_A(D)$ and $\overline{M}_A(D)$, respectively and $L^* \leq L \leq |U|^2$. The worst case time complexity of Algorithm 3 is the same as the SPS method, while Algorithm 3 runs faster than SPS in the real application because there are not too many counting, intersection and union operations in Algorithm 3.

Algorithm 3 Matrix absorption law

Require: The inverse of the original discernibility matrix $\overline{M}_A(D)$
Ensure: The inverse of the absorbed discernibility matrix $\overline{M}^*_A(D)$

```
Find the unique rows of M̄_A(D);
Sort the rows of M̄_A(D) in ascending order in terms of the number of 1's in each row and represent the sorted matrix as M̄*_A(D);
Let L* be the number of rows in M̄*_A(D) and i = 1;
while i ≤ L* do
    Delete the other rows that contain the ith row in M*_A(D), that is, the other rows that satisfy M̄*_A(D)C_i^T = l_i, where C_i is the ith row in M̄*_A(D) and l_i is the number of 1's in C_i;
    Update L*
    i = i + 1
end while
```

Next, we implement the distribution law in the discernibility function based on Shannon's expansion [42]. Shannon's expansion shows that any matrix or disjunction normal form $\overline{M}$ can be represented by a combination of two sub-functions of the original:

$$\overline{M} = m_j \overline{M}_{m_j} + \overline{M}_{\overline{m}_j},$$

where m_j is the variable in j-th column, $\overline{M}_{m_j}$ is obtained by setting the j-th column as 0 in M, $M_{\overline{m}_j}$ is obtained by deleting the rows that the j-th column is 1 in M, and $\overline{M}_{\overline{m}_j}$ is the inverse of $M_{\overline{m}_j}$. By recursively using Shannon's expansion, the conjunction normal $M_A(D)$ can be transformed to the disjunction normal form, and the reducts can be found. The detailed implementation processes are shown in Algorithm 4. Actually, the right branch of and the left branch of Algorithm 4 implement the $m_j \overline{M}_{m_j}$ and $\overline{M}_{\overline{m}_j}$, respectively. One can refer to Ref. [42] for a vivid example. The time complexity of Algorithm 4 is $O(L^*V_1)$, where V_1 is the number of nodes and L^* is the time complexity of each node (the number of rows in $\overline{M}^*_A(D)$).

Algorithm 4 Matrix distribution law: Shannon's expansion for $M^*_A(D)$

Require: The inverse of the absorbed discernibility matrix $\overline{M}^*_A(D)$
Ensure: A set of reducts RED

```
Choose the rows in M̄*_A(D) which have the fewest number of 1's and put the features that appear in these rows into feature subset A_0;
Choose the attribute from A_0 that appears most often in the other rows in M̄*_A(D) (suppose it is the jth attribute);
Build the matrix of the right branch Right_j with the rows in M̄*_A(D) whose jth column are 0;
Build the matrix of the left branch Left_j by setting the jth column of M̄*_A(D) as 0 and finding the unique of the rows;
if There is a row of all 0's in Left_j then
    Left'_j = ∅;
else
    Left'_j ← Algorithm 4 (Left_j);
end if
if Right_j = ∅ then
    Right'_j is equal to the zero vector with |A| columns;
else if Right_j has only one row then
    Let l be the number of 1's in Right_j;
    Initialize Right'_j as the zero matrix with l rows and |A| columns;
    for i = 1 to l do
        Denote h as the column index corresponding to the ith 1 of Right_j;
        Set the ith row and hth column in Right'_j as 1;
    end for
else
    Right'_j ← Algorithm 4 (Right_j);
end if
Set the jth column in Right'_j as 1;
RED = [Left'_j; Right'_j].
```

Algorithm 4 finds all reducts simultaneously. Generally, for data sets with a high dimension, the number of reducts found by Algorithm 4 is quite large. However, in ensemble learning,

fewer base classifiers can achieve satisfactory performance. Thus, sometimes we need to manually specify the number of base classifiers for fusion. Based on these two considerations, it is very meaningful to find a reduct by discernibility matrix. In [39], Yao and Zhao proposed a discernibility matrix simplification methodology for constructing attribute reducts. Following their work, we give a procedure to find a reduct based on discernibility matrix, which is shown as Algorithm 5. This algorithm executes the absorption and deletion operations row by row to simplify the elements in discernibility matrix to singleton sets, and the union of all the singleton sets is a reduct. The deletion operation complies with the fact that the attributes set A can be deleted if none of the subsets of A corresponds to a row in $\overline{M}_A^*(D)$. The time complexity of Algorithm 5 is $O(tL^*|A|)$, where t is the number of iterations and $t \ll L^*$. Running Algorithm 5 multiple times can generate a set of reducts for learning multiple monotonic decision trees.

Algorithm 5 Finding a reduct based on discernibility matrix

Require: The inverse of the absorbed discernibility matrix $\overline{M}_A^*(D)$
Ensure: A reduct B
1: Sort the rows of $\overline{M}_A^*(D)$ randomly;
2: Let L^* be the number of rows in $\overline{M}_A^*(D)$;
3: **while** $i \leq L^*$ **do**
4: Select the ith row form $\overline{M}_A^*(D)$ and note as m_i;
5: Put the features that appear in m_i into features set A;
6: Randomly select a feature from A, note as a;
7: Set the ath column of m_i to be 1 and other columns to be 0;
8: Set $A_0 = A - a$
9: **for** $j = i + 1$ to L^* **do**
10: Let m_j be the jth row in $\overline{M}_A^*(D)$;
11: **if** The ath column of m_j is 1 **then**
12: Delete m_j;
13: **else if** The features that appear in m_j are not a subset of A_0 **then**
14: Set the A_0 column of m_j as 0;
15: **end if**
16: **end for**
17: Update L^* and set $i = i + 1$;
18: **end while**
19: Output B as the union of attribute that appears in $\overline{M}_A^*(D)$.

Based on the above two methods to find reducts, we develop two fusion methods. One is for using all reducts (Algorithm 4), and the other is for fusion with a specified number of base classifiers (Algorithm 5). Using each reduct, we grow a FREMT and fuse all FREMT trees by majority voting. That is, an object is classified into the class that is assigned by the majority of the FREMT trees. For the fusion strategy, we can also use the median of the label as the prediction label, while the vote strategy is a more robust and lower-complexity approach [28]. The detailed algorithm is shown in Algorithm 6, named as FFREMT: Fusing Fuzzy Rank Entropy Based Monotonic Tree. The FFREMT mainly contains two aspects: generating attribute reduction subsets and fusing multiple REMT trees. As for the first aspect, the time complexity has been given above. The time complexity of building L' FREMT trees is $O(L'bV_2|U||A|^2)$, where b is the number of nodes in the tree and V_2 is the number of different values in the features.

Algorithm 6 FFREMT

Require: Ordinal decision table $OD = (U, A \cup d)$; stoping criterion of FREMT: ε; Parameter of discernibility matrix p; Sample to be predicted: x
Ensure: The decision of x
1: Generate discernibility matrix $M_A(D)$ by Definition 9 with parameter p and express it in binary matrix form;
2: Simplify $\overline{M}_A(D)$ as $\overline{M}_A^*(D)$ by Algorithm 3;
3: Find multiple reducts RED with $\overline{M}_A^*(D)$ by Algorithm 4 or by running Algorithm 5 L' times; \\ Algorithm 4 is designed for finding all reducts and Algorithm 5 for a reduct.
4: Transform each row of RED into the collection form and denote as $\{B_1, ..., B_{L'}\}$, where L' is the rows number of RED;
5: **for** B_1 to $B_{L'}$ **do**
6: learn a monotonic decision tree T_l with Algorithm 2.
7: make a prediction on x using T_l: $d_l(x) = T_l(x)$.
8: **end for**
9: Return the predict class by voting rule: $\hat{d}(x) = arg\max_k \big(\sum_{j=1}^{L'} I\{d_l = k\} \big)$ ($I\{\cdot\}$ is the counting function).

IV. Experimental Analysis

In this section, the effectiveness of FFREMT is shown by comparing with several other ensemble methods for ordinal classification.

A. Benchmark Methods and Data Sets

FREMT [7] is used as the base classifier for all ensemble methods. Four significance measures for ordinal attributes have been defined in [8] and [32], as introduced in Eq. (8)-(9), (19), (20) and (21) in Section II. The algorithm for searching multiple reduction subsets with significance measures have been applied in [8], that is, Algorithm 3 in [8]. We embed the four significance measures into the Algorithm 3 in [8] to obtain multiple reductions and use them construct multiple FREMT, and the ensemble methods based on the four significance measures are represented as β-red [8], up-sig [32], down-sig [32], and global-sig [32], respectively. In addition, there are two discernibility matrixes defined for ordinal classification. One is the monotonic consistency matrix MCM in Definition 1. The other is FCMT based on the monotonic consistency set [22]. We embed MCM and FCMT into the Step 1 of Algorithm 6 to fuse FREMT. We also use bagging to promote FREMT. In bagging, we form 21 new training sets, construct FREMT on them and ensemble the trees by voting rules. We will compare FFREMT with these methods.

TABLE IV: Data sets in the experimental analysis

ID	Data sets	objects	features	classes	source
1	Breast cancer	699	9	2	UCI
2	Wine-red	1599	12	6	KEEL
3	Wine	178	13	3	KEEL
4	Boston housing	506	13	4	WEKA
5	Australian	690	14	2	UCI
6	German	1000	19	2	KEEL
7	Student score	512	25	3	[8]
8	Statlog Landsat Satellite	6435	36	7	UCI
9	Waveform Database Generator40	5000	40	3	UCI
10	Molecular Biology	106	57	2	UCI
11	Residential-Building	372	103	5	UCI
12	Urban Land Cover	168	147	9	UCI
13	Parkinson's Disease	756	752	2	UCI
14	DrivFace	606	6400	3	UCI
15	Arcene	100	10000	2	UCI

Fifteen real-world classification tasks are used as benchmark data sets. The detail information of them is described in Table IV. The missing values in data sets are imputed by the nearest-neighbor method, while the rows or columns with too many missing values will be deleted. The predict variable of Residential-Building is continuous. We discretize it by the bin method with a ratio of 144:105:90:19:14. For the first seven data sets, we find all the reducts by Algorithm 4, and compare FFREMT with all of the above mentioned benchmark methods. For the last eight data sets, we find 21 reducts by Algorithm 5, and because the significance measure based methods are computationally infeasible, we only compare FFREMT with Bagging and the other two discernibility matrix based methods.

B. Data Pre-processing

To ensure that the preference degree generated by each attribute is in the same dimension, a data set is normalized through training set by the range method. That is, each attribute is normalized by the following transformation:

$$v(x,a) = \frac{v(x,a) - \min_{y\in U_1} v(y,a)}{\max_{y\in U_1} v(y,a) - \min_{y\in U_1} v(y,a)}, \tag{41}$$

where U_1 is the training set.

In practice, the data set may present decreasing monotonicity: the worse feature value gets the better decision class. In this case, we need to preprocess each attribute in data sets to satisfy the assumption of increasing monotonicity. There are several solutions to this problem. Here, we choose the Spearman rank correlation coefficient to measure the monotonicity consistency between attribute and decision. In particular, let a_i and d_i $(i = 1, 2, \cdots, n)$ be the ranking results on n objects according to attribute and decision, respectively. The rank correlation coefficient is calculated as:

$$\textit{Spearman}(a,d) = 1 - \frac{6}{n(n^2-1)}\sum_{i=1}^{n}(a_i - d_i)^2. \tag{42}$$

If the Spearman rank correlation coefficient between it with the decision is negative, we use one minus the attribute to replace the original one.

Besides, label noise is also widespread and makes impacts on the decision-making. In ordinal decision table with increasing monotonic constraint, the objects with a better attribute but a worse decision class will be considered as a noisy object. Although we have preprocessed the data to satisfy increasing monotonic, this process is totally based on a global Spearman rank correlation coefficient value. Here, we reset the label of training objects which belong to a monotone inconsistent set. In particular, let

$$U_M^{\beta} = \begin{cases} \{x | \frac{|[x]_A^{\leqslant} \cap [x]_d^{\leqslant}|}{|[x]_A^{\leqslant}|} \geq 1-\beta\}, & \text{if } |[x]_A^{\geqslant}| < n_0; \\ \{x | \frac{|[x]_A^{\geqslant} \cap [x]_d^{\geqslant}|}{|[x]_A^{\geqslant}|} \geq 1-\beta\}, & \text{if } |[x]_A^{\leqslant}| < n_0; \\ \{x | \frac{|[x]_A^{\geqslant} \cap [x]_d^{\geqslant}|}{|[x]_A^{\geqslant}|} \geq 1-\beta, \frac{|[x]_A^{\leqslant} \cap [x]_d^{\leqslant}|}{|[x]_A^{\leqslant}|} \geq 1-\beta\}, & \text{otherwise} \end{cases} \tag{43}$$

be the monotone consistent set, where $\beta \in [0, 0.5]$ and n_0 is a constant that is much smaller than the number of the training objects. The probability that an object belongs to d_k can be estimated by:

$$P(d(x) = d_k) = P(d(x) \geq d_k) - P(d(x) \geq d_{k+1}), \tag{44}$$

where

$$P\big(d(x) \geqslant d_k\big) = \frac{\big|[x]_A^{\geqslant} \cap d_k^{\geqslant}\big|}{\big|[x]_A^{\geqslant}\big|}, \tag{45}$$

and $P\big(d(x) \geqslant d_{K+1}\big) = 0$. For $x \in U_1 - U_M^{\beta}$, the label of it should be reset as:

$$d(x)^* = arg\max_k P(d(x) = d_k), \quad k = 1, ..., K. \tag{46}$$

C. Evaluation Measures and Parameter Selection

Here, the classification accuracy CA and the mean absolute error MAE are employed to evaluate the performance of the methods. CA is computed as:

$$CA = \frac{1}{n}\sum_{i=1}^{n} I\{\widehat{d}(x_i) = d(x_i)\}, \tag{47}$$

where $I\{\cdot\}$ is the counting function: if $\widehat{d}(x_i) = d(x_i)$, then $I\{\widehat{d}(x_i) = d(x_i)\} = 1$, otherwise $I\{\widehat{d}(x_i) = d(x_i)\} = 0$. MAE is computed as:

$$MAE = \frac{1}{n}\sum_{i=1}^{n} |\widehat{d}(x_i) - d(x_i)|, \tag{48}$$

where n is the number of objects, $\widehat{d}(x_i)$ is the predict decision and $d(x_i)$ is the true decision.

For the first seven data sets, we randomly divide it into U_1 and a test set U_2 at a ratio of 7:3, and for the last eight bigger ones, we divide at a ratio of 1:9. We conduct the division 24 times to obtain the average performance. All methods are run under the same U_1-U_2 division. For the methods without parameters to tune, we use U_1 as the training set. For the methods with parameters to tune, we further divide U_1 into a training set U_{11} and a validation set U_{12} at a ratio of

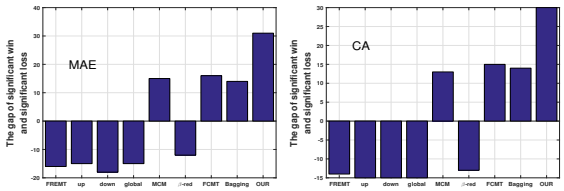


Fig. 1: Significance Test Results in terms of MAE and CA

7:3. We conduct the U_{11}-U_{12} division 10 times to obtain the mean validation MAE for each pair of parameters. And the parameters with the minimum mean validation MAE are chosen as the optimal parameter for training. The parameter p is chosen from 1 to $\min\{\lfloor |A|/2 \rfloor, 10\}$ with a step 1 and β is chosen from 0.1 to 0.5 with a step 0.1.

D. Experimental Results and Analysis

Table V and VI records the mean and standard deviation of test MAE and test CA. In each row, the method with the maximum value is in bold font; black dots mark behind the values of a method which FFREMT is significantly better than under the pairwise right-tailed Student's test with 95 percent significance level. The last row records the times of win-lose-tie by comparing FFREMT with the others, and the numbers in brackets record the times of significant win. It can be observed that FFREMT performs better on most data sets, and makes significant improvements in Wine-red, Boston housing, German, Student Score, Stalog, Waveform and Arcene.

To further analyze the results, we compare these methods by their confidence intervals [23]. Let $[L_{(a,d,m)}, U_{(a,d,m)}]$ be the 95 percent confidence interval of method a on data set d in the sense of performance measure m:

$$\begin{aligned} L_{(a,d,m)} &= \mu_{(a,d,m)} - 1.96\frac{\sigma_{(a,d,m)}}{\sqrt{n}}, \\ U_{(a,d,m)} &= \mu_{(a,d,m)} + 1.96\frac{\sigma_{(a,d,m)}}{\sqrt{n}}, \end{aligned} \tag{49}$$

where $\mu_{(a,d,m)}$ is the mean value and $\sigma_{(a,d,m)}$ is the standard deviation across n times runs. For two methods a and a', if $L_{a,d,m} > U_{a',d,m}$, then we say that a significantly wins a', otherwise a significantly loses to a'. Each bar in Fig. 1 represents the number gap between the significant win and lose times of a given method compared with the others on the first seven data sets. From Fig. 1, we can see that the performance of FFREMT is significantly better than the other six reduction-based ensemble methods and Bagging.

We show the influence of parameters on each data set in Fig. 2. The x-axis and the y-axis represent the value of β and p, respectively. The z-axis represents mean validation MAE. We can observe that the parameter p plays an important role in tuning the performances of the FFREMT algorithm. For Breast cancer, Wine, Student score and Stalog, a larger p value performs better. This signifies that for data sets with more monotonic consistency attributes, the parameter p should be larger. For Boston housing, Australian, German and Arcene, the influence of parameter β is more than that of parameter p.

We also show the comparison of mean reduction time and mean tree-building time on each data set in Fig. 3. It is easy to observe that the reduction methods based on significance measures are more time consuming than the discernibility matrix based methods. And in discernibility matrix based methods, the time is dominated by tree-building, which is meet the basic requirements of ensemble learning. These demonstrate that the proposed reduction method and ensemble method is effective and serviceable.

V. Conclusion and Future Work

In this paper, we mainly focus on developing a fusion method based on attribute reduction to solve ordinal classification problem. To achieve this, we firstly define a discernibility matrix with fuzzy dominance rough set and introduce the algorithms for finding reducts based on it. Each reduct forms a feature subspace with original information, and ordinal decision trees are built in these feature subspaces. Finally, we fuse these trees by voting. Numerical experimental results demonstrate that the proposed fusion method is effective and feasible. The major contribution of this paper is that we have theoretically defined a discernibility matrix using fuzzy dominance rough set and have proposed an effective fusion method for ordinal classification problem. We have verified that the reducts based on discernibility matrix are well suited to ensemble learning, and the combination of them is competent to further improve the generalization performance of the fuzzy monotonic decision tree. In addition, we have offered efficient procedures for implementing the laws of absorption and distribution, and for finding a reduct based on the discernibility matrix. In the future, we will work on eliminating the random consistency from the attribute dependence measure to describe the attribute more objectively and correctly. We will also consider attribute reduction and ensemble methods for mixture decision tables which contain both ordinal and nominal attributes.

Acknowledgment

This work was supported by National Key R&D Program of China (No. 2018YFB1004300), National Natural Science Foundation of China(Nos. 61672332, 61872226, 61802238, 61976120), Program for the Outstanding Innovative Teams of Higher Learning Institutions of Shanxi, Program for the San Jin Young Scholars of Shanxi, the Overseas Returnee Research Program of Shanxi Province (No. 2017023), Natural Science Foundation of Shanxi Province (Grant No. 201701D121052).

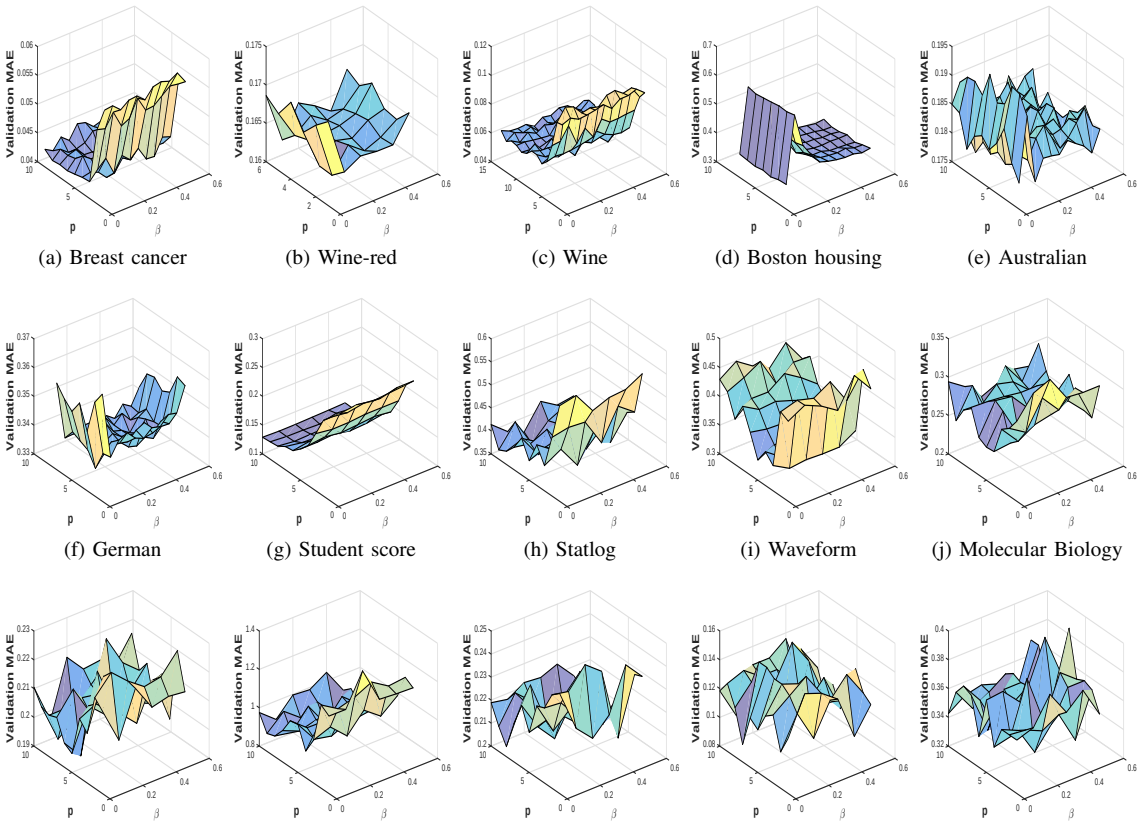
References

[1] L. A. Zadeh, "Fuzzy sets," *Information & Control*, vol. 8, no. 3, pp. 338–353, 1965.
[2] D. Dubois and H. Prade, "Rough fuzzy sets and fuzzy rough sets," *International Journal of General Systems*, vol. 17, no. 2-3, pp. 191–209, 1990.
[3] Z. Pawlak, "Rough sets," *International Journal of Computer and Information Sciences*, vol. 11, no. 5, pp. 341–356, 1982.
[4] A. M. Radzikowska and E. E. Kerre, "A comparative study of fuzzy rough sets," *Fuzzy Sets and Systems*, vol. 126, no. 2, pp. 137–155, 2002.
[5] D. S. Yeung, D. Chen, E. C. C. Tsang, J. W. T. Lee, and W. Xizhao, "On the generalization of fuzzy rough sets," *IEEE Transactions on Fuzzy Systems*, vol. 13, no. 3, pp. 343–361, 2005.
[6] R. Jensen and Q. Shen, "Fuzzy rough attribute reduction with application to web categorization," *Fuzzy Sets and Systems*, vol. 141, no. 3, pp. 469–485, 2004.

TABLE V: Comparison on MAE and CA

DATA ID	**MAE ± std**								
	[7]	[32]			[8]		[22]	[27]	OUR
	FREMT	up-sig	down-sig	global-sig	MCM	β-red	FCMT	Bagging	FFREMT
1	0.0581 ±0.0138 •	0.0579 ±0.0137 •	0.0567 ±0.0173 •	0.0594 ±0.0116 •	0.0474 ±0.0106 •	0.0754 ±0.1013	0.0442 ±0.0173	0.0434 ±0.0137	**0.0411** ±0.0096
2	0.2125 ±0.0134 •	0.2113 ±0.0197 •	0.2182 ±0.0137 •	0.2156 ±0.0153 •	0.2136 ±0.0101 •	0.3412 ±0.0810 •	0.2042 ±0.0168 •	0.1952 ±0.0163 •	**0.1617** ±0.0107
3	0.0796 ±0.0485 •	0.0803 ±0.0429 •	0.0795 ±0.0401 •	0.0917 ±0.0447 •	**0.0455** ±0.0299	0.0841 ±0.0415 •	0.0470 ±0.0227	0.0811 ±0.0422 •	0.0477 ±0.0298
4	0.4077 ±0.0435 •	0.3980 ±0.0498	0.4188 ±0.0390 •	0.3923 ±0.0467	0.3823 ±0.0492	0.3883 ±0.0354	**0.3812** ±0.0370	0.3958 ±0.0431	**0.3812** ±0.0374
5	0.1969 ±0.0309 •	0.1923 ±0.0179 •	0.1903 ±0.0188	0.1903 ±0.0239	0.1869 ±0.0214	0.1931 ±0.0242 •	0.1811 ±0.0320	**0.1633** ±0.0297	0.1777 ±0.0271
6	0.3855 ±0.0273 •	0.3693 ±0.0186 •	0.3789 ±0.0620 •	0.3808 ±0.0303 •	0.3616 ±0.0238 •	0.3751 ±0.0352 •	0.3713 ±0.0259 •	0.3696 ±0.0539 •	**0.3432** ±0.0231
7	0.2557 ±0.0319 •	0.2433 ±0.0326 •	0.2489 ±0.0288 •	0.2481 ±0.0262 •	0.1226 ±0.0192 •	0.2089 ±0.0345 •	0.1204 ±0.0220	0.1634 ±0.0239 •	**0.1097** ±0.0187
win-lose-tie	7(7)-0-0	7(6)-0-0	7(6)-0-0	7(5)-0-0	6(4)-1-0	7(5)-0-0	6(2)-0-1	6(4)-1-0	
DATA ID	**CA ± std**								
	[7]	[32]			[8]		[22]	[27]	OUR
	FREMT	up-sig	down-sig	global-sig	MCM	β-red	FCMT	Bagging	FFREMT
1	0.9419 ±0.0138 •	0.9421 ±0.0137 •	0.9433 ±0.0172 •	0.9405 ±0.0116 •	0.9526 ±0.0106 •	0.9246 ±0.1013	0.9558 ±0.0173	0.9565 ±0.0137	**0.9589** ±0.0097
2	0.7875 ±0.0134 •	0.7887 ±0.0197 •	0.7818 ±0.0137 •	0.7844 ±0.0153 •	0.7864 ±0.0101 •	0.6588 ±0.0810 •	0.7958 ±0.0168 •	0.8048 ±0.0163 •	**0.8383** ±0.0107
3	0.9250 ±0.0459 •	0.9227 ±0.0391 •	0.9235 ±0.0375 •	0.9159 ±0.0359 •	**0.9545** ±0.0298	0.9174 ±0.0408 •	0.9530 ±0.0227	0.9204 ±0.0418 •	0.9523 ±0.0297
4	0.6577 ±0.0343 •	0.6588 ±0.0349 •	0.6448 ±0.0305 •	0.6580 ±0.0306 •	0.6696 ±0.0405	0.6656 ±0.0325	0.6675 ±0.0297	0.6642 ±0.0317	**0.6794** ±0.0323
5	0.8031 ±0.0309 •	0.8077 ±0.0179 •	0.8097 ±0.0188	0.8097 ±0.0239	0.8131 ±0.0214	0.8069 ±0.0242 •	0.8189 ±0.0320	**0.8367** ±0.0297	0.8223 ±0.0271
6	0.6145 ±0.0273 •	0.6307 ±0.0186 •	0.6211 ±0.0620 •	0.6192 ±0.0303 •	0.6384 ±0.0238 •	0.6249 ±0.0352 •	0.6287 ±0.0259 •	0.6304 ±0.0539 •	**0.6568** ±0.0231
7	0.7454 ±0.0318 •	0.7570 ±0.0324 •	0.7513 ±0.0287 •	0.7519 ±0.0262 •	0.8774 ±0.0192 •	0.7911 ±0.0345 •	0.8796 ±0.0220	0.8366 ±0.0239 •	**0.8903** ±0.0187
win-lose-tie	7(7)-0-0	7(7)-0-0	7(6)-0-0	7(6)-0-0	6(4)-1-0	7(5)-0-0	7(2)-0-0	6(4)-1-0	

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TFUZZ.2019.2953024, IEEE Transactions on Fuzzy Systems

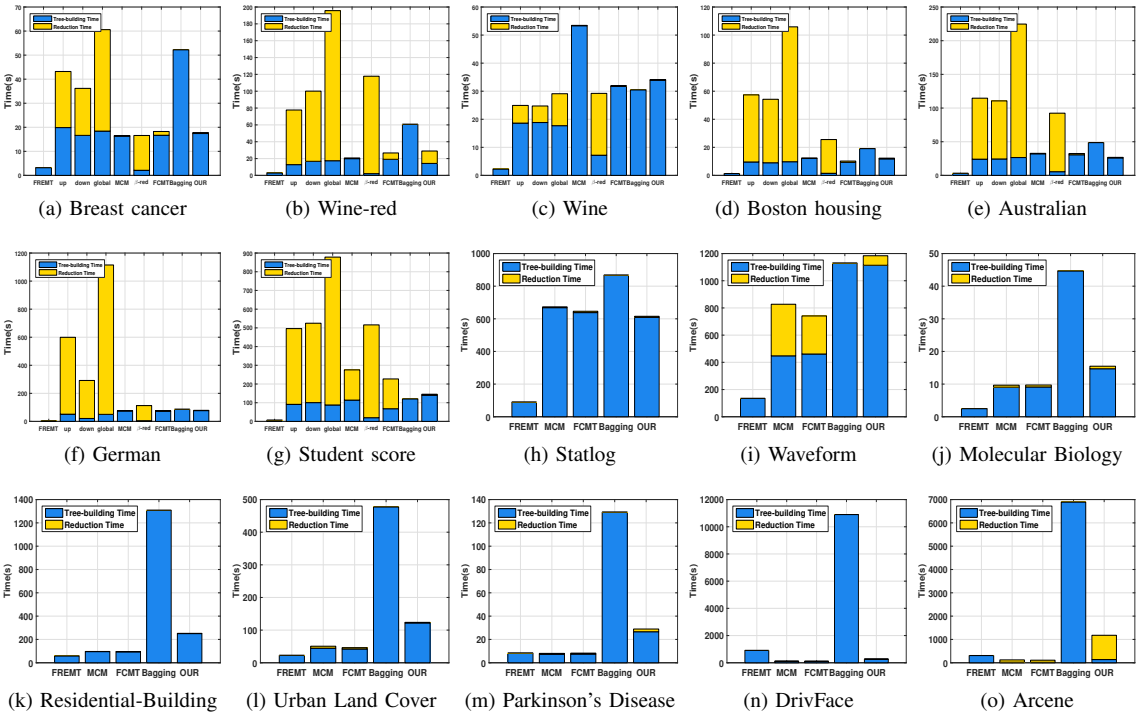


(a) Breast cancer (b) Wine-red (c) Wine (d) Boston housing (e) Australian

(f) German (g) Student score (h) Statlog (i) Waveform (j) Molecular Biology

(k) Residential-Building (l) Urban Land Cover (m) Parkinson's Disease (n) DrivFace (o) Arcene

Fig. 2: The Influence of Parameters on Validation MAE



(a) Breast cancer (b) Wine-red (c) Wine (d) Boston housing (e) Australian

(f) German (g) Student score (h) Statlog (i) Waveform (j) Molecular Biology

(k) Residential-Building (l) Urban Land Cover (m) Parkinson's Disease (n) DrivFace (o) Arcene

Fig. 3: Time Comparison

TABLE VI: Comparison on MAE and CA

DATA ID	**MAE ± std**				
	FREMT	MCM	FCMT	Bagging	FFREMT
8	0.5933 ±0.0396 •	0.3718 ±0.0218	0.3703 ±0.0188	0.3712 ±0.0132	**0.3644** ±0.0180
9	0.4341 ±0.0165 •	0.4632 ±0.0293 •	0.4585 ±0.0258 •	0.3901 ±0.0195 •	**0.3377** ±0.0360
10	0.4625 ±0.1100 •	0.3563 ±0.0646 •	0.3438 ±0.0442 •	0.3625 ±0.0754 •	**0.2406** ±0.0590
11	0.2426 ±0.0207 •	0.2139 ±0.0479	0.2078 ±0.0415	0.2035 ±0.0253	**0.1852** ±0.0396
12	1.5128 ±0.2321 •	0.8964 ±0.2757	0.9508 ±0.2691	1.0709 ±0.2385 •	**0.8236** ±0.2384
13	0.2722 ±0.0237 •	0.2394 ±0.0068 •	0.2299 ±0.0100 •	0.2106 ±0.0126	**0.2082** ±0.0125
14	0.1924 ±0.0479 •	0.1063 ±0.0118	0.1030 ±0.0081	0.1062 ±0.0161	**0.1030** ±0.0133
15	0.4194 ±0.0974 •	0.3419 ±0.0531	0.3452 ±0.0628	0.3226 ±0.0938	**0.3000** ±0.0900
win-lose-tie	8(8)-0-0	8(3)-0-0	8(3)-0-0	8(3)-0-0	
DATA ID	**CA ± std**				
	FREMT	MCM	FCMT	Bagging	FFREMT
8	0.7547 ±0.0143 •	0.8514 ±0.0069	0.8516 ±0.0069	0.8523 ±0.0047	**0.8526** ±0.0072
9	0.6656 ±0.0127 •	0.6633 ±0.0212 •	0.6686 ±0.0157 •	0.7079 ±0.0159 •	**0.7448** ±0.0257
10	0.5375 ±0.1100 •	0.6438 ±0.0646 •	0.6562 ±0.0442 •	0.6375 ±0.0754 •	**0.7594** ±0.0590
11	0.7617 ±0.0214 •	0.7991 ±0.0411	0.8052 ±0.0392	0.8000 ±0.0249	**0.8174** ±0.0408
12	0.5255 ±0.0714 •	0.7509 ±0.0752	0.7273 ±0.0647	0.6546 ±0.0781 •	**0.7564** ±0.0708
13	0.7278 ±0.0237 •	0.7607 ±0.0068 •	0.7701 ±0.0100 •	0.7894 ±0.0126	**0.7918** ±0.0125
14	0.8153 ±0.0454 •	0.8964 ±0.0093	0.8984 ±0.0068	0.8944 ±0.0162	**0.8993** ±0.0111
15	0.5807 ±0.0974 •	0.6581 ±0.0531	0.6548 ±0.0628	0.6774 ±0.0938	**0.7000** ±0.0900
win-lose-tie	8(8)-0-0	8(3)-0-0	8(3)-0-0	8(3)-0-0	

[7] Q. Hu, X. Che, L. Zhang, D. Zhang, M. Guo, and D. Yu, "Rank entropy-based decision trees for monotonic classification," *IEEE Transactions on Knowledge and Data Engineering*, vol. 24, no. 11, pp. 2052–2064, 2012.

[8] Y. Qian, H. Xu, J. Liang, B. Liu, and J. Wang, "Fusing monotonic decision trees," *IEEE Transactions on Knowledge and Data Engineering*, vol. 27, no. 10, pp. 2717–2728, 2015.

[9] Y. Qian, Y. Li, J. Liang, G. Lin, and C. Dang, "Fuzzy granular structure distance," *IEEE Transactions on Fuzzy Systems*, vol. 23, no. 6, pp. 2245–2259, 2015.

[10] Z. H. Zhou, "Abductive learning: towards bridging machine learning and logical reasoning," *Science China (Information Sciences)*, vol. 62, no. 7, pp. 076 101:1–076 101:3, 2019.

[11] D. Ślezak, "Approximate entropy reducts," *Fundamenta Informaticae*, vol. 53, no. 3-4, pp. 365–390, 2002.

[12] E. C. Tsang, D. Chen, D. S. Yeung, X.-Z. Wang, and J. W. Lee, "Attributes reduction using fuzzy rough sets," *IEEE Transactions on Fuzzy Systems*, vol. 16, no. 5, pp. 1130–1141, 2008.

[13] D. Chen, Q. Hu, and Y. Yang, "Parameterized attribute reduction with gaussian kernel based fuzzy rough sets," *Information Sciences*, vol. 181, no. 23, pp. 5169–5179, 2011.

[14] C. Z. Wang, Y. Qi, M. Shao, Q. Hu, D. Chen, Y. Qian, and Y. Lin, "A fitting model for feature selection with fuzzy rough sets," *IEEE Transactions on Fuzzy Systems*, vol. 25, no. 4, pp. 741–753, 2017.

[15] W. Ding, C. T. Lin, M. Prasad, Z. Cao, and J. D. Wang, "A layered-coevolution-based attribute-boosted reduction using adaptive quantum behavior PSO and its consistent segmentation for neonates brain tissue," *IEEE Transactions on Fuzzy Systems*, vol. 26, no. 3, pp. 1177–1191, 2018.

[16] W. Ding, C. T. Lin, and Z. Cao, "Deep neuro-cognitive co-evolution for fuzzy attribute reduction by quantum leaping PSO with nearest-neighbor memeplexes," *IEEE Transactions on Cybernetics*, vol. 49, no. 7, pp. 2744 – 2757, 2019.

[17] W. P. Ding, C. T. Lin, M. Prasad, S. B. Chen, and Z. J. Guan, "Attribute equilibrium dominance reduction accelerator (DCCAEDR) based on distributed coevolutionary cloud and its application in medical records," *IEEE Transactions on Systems Man & Cybernetics Systems*, vol. 46, no. 3, pp. 384–400, 2017.

[18] W. Chu and Z. Ghahramani, "Gaussian processes for ordinal regression," *Journal of Machine Learning Research*, vol. 6, no. Jul, pp. 1019–1041, 2005.

[19] T. B. Iwiński, "Ordinal information systems. I," *Bulletin of the Polish Academy of Sciences Mathematics*, vol. 36, no. 7, pp. 467–475, 1988.

[20] Y. Sai, Y. Y. Yao, and N. Zhong, "Data analysis and mining in ordered information tables," in *Proceedings of the 2001 IEEE International Conference on Data Mining*, 2001, pp. 497–504.

[21] S. Greco, B. Matarazzo, and R. Slowinski, "Rough approximation of a preference relation by dominance relations," *European Journal of Operational Research*, vol. 117, no. 1, pp. 63–83, 1999.

[22] H. Xu, W. Wang, and Y. Qian, "Fusing complete monotonic decision trees," *IEEE Transactions on Knowledge and Data Engineering*, vol. 29, no. 10, pp. 2223–2235, 2017.

[23] F. Li, Y. Qian, J. Wang, and J. Liang, "Multigranulation information fusion: A dempster-shafer evidence theory-based clustering ensemble method," *Information Sciences*, vol. 378, pp. 389–409, 2017.

[24] F. Li, Y. Qian, J. Wang, C. Dang, and B. Liu, "Cluster's quality evaluation and selective clustering ensemble," *ACM Transactions on Knowledge Discovery From Data*, vol. 12, no. 5, p. 60, 2018.

[25] F. Li, Y. Qian, J. Wang, C. Dang, and L. Jing, "Clustering ensemble based on sample's stability," *Artificial Intelligence*, vol. 273, pp. 37–55, 2019.

[26] D. D. Margineantu and T. G. Dietterich, "Pruning adaptive boosting," in *Proceedings of the 1997 International Conference on Machine Learning*, vol. 97, 1997, pp. 211–218.

[27] L. Breiman, "Bagging predictors," *Machine Learning*, vol. 24, no. 2, pp. 123–140, 1996.

[28] R. E. Schapire, Y. Freund, P. L. Bartlett, and W. S. Lee, "Boosting the margin: a new explanation for the effectiveness of voting methods," *Annals of Statistics*, vol. 26, no. 5, pp. 1651–1686, 1998.

[29] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[30] J. J. Rodriguez, L. I. Kuncheva, and C. J. Alonso, "Rotation forest: A new classifier ensemble method," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 28, no. 10, pp. 1619–1630, 2006.

[31] D. Dubois, H. F. Fargier, and P. Perny, "Qualitative decision theory with preference relations and comparative uncertainty: an axiomatic approach," *Artificial Intelligence*, vol. 148, pp. 219–260, 2003.

[32] Q. Hu, D. Yu, and M. Guo, "Fuzzy preference based rough sets," *Information Sciences*, vol. 180, no. 10, pp. 2003–2022, 2010.
[33] Y. Yao, "The two sides of the theory of rough sets," *Knowledge Based Systems*, vol. 80, pp. 67–77, 2015.
[34] A. Skowron and C. Rauszer, "The discernibility matrices and functions in information systems," *Fundamenta Informaticae*, pp. 331–362, 1992.
[35] H. S. Nguyen and A. Skowron, "Boolean reasoning for feature extraction problems," *Foundations of Intelligent Systems*, vol. 1325, pp. 117–126, 1997.
[36] Z. Pawlak and A. Skowron, "Rough sets and boolean reasoning," *Information Sciences*, vol. 177, no. 1, pp. 41–73, 2007.
[37] Q. H. Hu, M. Z. Guo, D. R. Yu, and J. F. Liu, "Information entropy for ordinal classification," *Science China Information Sciences*, vol. 53, no. 6, pp. 1188–1200, 2010.
[38] Q. Hu, D. Yu, Z. Xie, and X. Li, "Eros: Ensemble rough subspaces," *Pattern recognition*, vol. 40, no. 12, pp. 3728–3739, 2007.
[39] Y. Yao and Y. Zhao, "Discernibility matrix simplification for constructing attribute reducts," *Information Sciences*, vol. 179, no. 7, pp. 867–882, 2009.
[40] Y. Qian, J. Liang, W. Pedrycz, and C. Dang, "Positive approximation: An accelerator for attribute reduction in rough set theory," *Artificial Intelligence*, vol. 174, no. 9, pp. 597–618, 2010.
[41] D. Chen, S. Zhao, L. Zhang, Y. Yang, and X. Zhang, "Sample pair selection for attribute reduction with rough set," *IEEE Transactions on Knowledge and Data Engineering*, vol. 24, no. 11, pp. 2080–2093, 2012.
[42] G. Borowik, T. Luba, and D. Zydek, "Reduction of knowledge representation using logic minimization techniques," in *Proceedings of the 2011 International Conference on Systems Engineering*, 2011, pp. 482–485.

Jieting wang received a B.S. and M.S. degree at school of Mathematical Sciences from Shanxi University, China, in 2013 and 2015, respectively. She is a PhD candidate at Institute of Big Data Science and Industry, Shanxi University. Her research interest includes statistical machine learning and ensemble learning.

Yuhua Qian received the M.S. and Ph.D. degrees in computers with applications from Shanxi University, Taiyuan, China, in 2005 and 2011, respectively.

He is currently a Professor with the Key Laboratory of Computational Intelligence and Chinese Information Processing, Ministry of Education, Shanxi University. He is best known for multigranulation rough sets in learning from categorical data and granular computing. He is involved in research on pattern recognition, feature selection, rough set theory, granular computing, and artificial intelligence. He has authored over 80 articles on these topics in international journals. He served on the Editorial Board of the International Journal of Knowledge-Based Organizations and Artificial Intelligence Research. He has served as the Program Chair or Special Issue Chair of the Conference on Rough Sets and Knowledge Technology, the Joint Rough Set Symposium, and the Conference on Industrial Instrumentation and Control, and a PC Member of many machine learning, data mining, and granular computing conferences.

Feijiang Li received a B.S. degree at school of Computer Science and Technology from Northeast University, China, in 2012. He is a PhD candidate at Institute of Big Data Science and Industry, Shanxi University. His research interest includes machine learning and knowledge discovery.

Jiye Liang received the M.S. and Ph.D. degrees from Xi'an Jiaotong University, Xi'an, China, in 1990 and 2001, respectively. He is currently a Professor with the School of Computer and Information Technology, Shanxi University, Taiyuan, China, where he is also the Director of the Key Laboratory of Computational Intelligence and Chinese Information Processing of the Ministry of Education. He has authored more than 170 journal papers in his research fields. His current research interests include computational intelligence, granular computing, data mining, and knowledge discovery.

Weiping Ding (M'16-SM'19) received the Ph.D. degree in Computation Application, Nanjing University of Aeronautics and Astronautics (NUAA), Nanjing, China, in 2013. He was a Visiting Scholar at University of Lethbridge (UL), Alberta, Canada, in 2011. From 2014 to 2015, He is a Postdoctoral Researcher at the Brain Research Center, National Chiao Tung University (NCTU), Hsinchu, Taiwan. In 2016, He was a Visiting Scholar at National University of Singapore (NUS), Singapore. From 2017 to 2018, he was a Visiting Professor at University of Technology Sydney (UTS), Ultimo, NSW, Australia. His main research directions involve granular computing data mining, machine learning and big data analytics. He has published over 50 papers in flagship journals and conference proceedings as the first author, including IEEE Transactions on Fuzzy Systems, IEEE Transactions on Neural Network and Learning System, IEEE Transactions on Cybernetics, and so on. To data, he has held 12 approved invention patents in total over 20 issued patents. Dr. Ding currently serves on the Editorial Advisory Board of Knowledge-based Systems. He serves as an Associate Editor of IEEE Transactions on Fuzzy Systems, Information Sciences, Swarm and Evolutionary Computation, Journal of Intelligent & Fuzzy Systems, and IEEE Access, and a leading guest editor for four international journals. He serves/served as a program committee member for seven international conferences and workshops. Dr. Ding is a member of Task Force on Adaptive and Evolving Fuzzy Systems of IEEE CIS FSTC, a member of Soft Computing TC of IEEE SMC, a member of Granular Computing (GrC) TC of IEEE SMC, and also a member of Data Mining and Big Data Analytics TC of IEEE CIS.