Active Antinoise Fuzzy Dominance Rough Feature Selection Using Adaptive K-Nearest Neighbors

Binbin Sang, Weihua Xu, Hongmei Chen, *Member, IEEE*, and Tianrui Li, *Senior Member, IEEE*

***Abstract*—Feature selection methods with antinoise performance are effective dimensionality reduction methods for classification tasks with noise. However, there are few studies on robust feature selection methods for monotonic classification tasks. The fuzzy dominance rough set (FDRS) model is a nontrivial knowledge acquisition tool, which is widely used in feature selection of monotonic classification tasks. Nonetheless, this model has been proved in practice to be generally poorly fault-tolerance, and only one noisy sample can cause huge interference in acquiring knowledge. In view of these two issues, this article first designs an adaptive K-nearest neighbors strategy to calculate the density of samples. The noisy samples are identified according to their densities, and then an active antinoise FDRS model is proposed. Then, in the active antinoise fuzzy dominance rough approximation space, the class-separability is evaluated by the approximation operators of the proposed model, and the feature-redundancy is evaluated by the fuzzy ranking conditional mutual information. On this basis, a feature evaluation index is designed comprehensively considering class-separability and feature-redundancy. Finally, a feature selection algorithm is designed to select the feature subset with the highest classification performance. The experimental results show that the proposed algorithm has better robustness and classification performance.**

***Index Terms*—Active antinoise, adaptive K-nearest neighbors (AKNN), feature selection, fuzzy dominance rough sets (FDRSs), ordered decision systems (ODSs).**

I. Introduction

MONOTONIC classification task (MCT) is a common classification problem, which has two characteristics: one is that the class labels are discrete and ordered, and the other is that there is a monotonic dependency between features and decisions [1], [2]. For example, a credit evaluation agency will determine an individual's credit rating based on criteria (features) such as an individual's income, educational experience, and asset status, and so on. In practical applications, the MCT widely exists, such as medical diagnosis [3], financial risk prediction [4], enterprise rating [5], and so on. In rough set theory [6], the MCT can be represented by a data table, called an ordered decision system (ODS).

Fuzzy rough set (FRS) theory forms a mathematical model that can simulate human reasoning by integrating two complementary uncertainty reasoning ways: rough approximation and fuzzy granulation [7]. This theory has become a very popular knowledge acquisition tool in data mining, which has been successfully applied in outlier detection [8], [9], attribute reduction [10], [11], [12], classification [13], [14], clustering [15], and cognitive networks [16], [17], etc. However, after investigation, it is found that the FRS model is very sensitive to noise [18]. In the real world, the collected data is usually polluted by noise, which makes the fuzzy rough set model unable to exert its advantages in dealing with uncertainty problems [19]. Hence, the robustness of this theory has become a research hotspot in data mining. Over the past decade, a series of robust fuzzy rough set models have been proposed, such as variable precision $(\theta, \sigma)-$fuzzy rough sets [20], data-distribution-aware fuzzy rough sets [13], different classes' ratio fuzzy rough sets [21], relative distance-based fuzzy rough sets [14], fuzzy rough sets with representative samples [22], probability granular distance-based fuzzy rough sets [23], and so on. Moreover, Alcantud et al. [24] revealed a close connection between N-soft sets and rough structures of various types, then proposed an N-soft set approach to rough set. The research on robust fuzzy rough sets has achieved remarkable achievements, they only consider the fuzzy equivalence relation or fuzzy similarity relation between samples in nominal classification tasks. However, the above works ignore the preference relation between samples and the monotonously ordered relation between features and decisions in monotonic classification tasks. Obviously, these robust models are not suitable for knowledge acquisition of the ODS. Therefore, developing a robust FRS model for ODS is one of the research goals of this article.

The dominance-based rough set approach (DRSA) [25] provides a theoretical framework for uncertainty analysis and knowledge acquisition in the ODS. Subsequently, some extended models of DRSA were successively proposed, including stochastic dominance-based rough sets model [26], variable consistency dominance-based rough sets approach (VC-DRSA) [27], rough sets model based on multigranulation fuzzy

Manuscript received 3 October 2022; revised 26 March 2023; accepted 27 April 2023. Date of publication 2 May 2023; date of current version 1 November 2023. This work was supported in part by the Chongqing Normal University Foundation Project under Grant 22XLB019, and in part by the National Natural Science Foundation of China under Grant 61976245 and Grant 61976182. *(Corresponding author: Weihua Xu.)*

Binbin Sang is with the College of Computer and Information Science, Chongqing Normal University, Chongqing 401331, China (e-mail: sangbinbin@cqnu.edu.cn).

Weihua Xu is with the College of Artificial Intelligence, Southwest University, Chongqing 400715, China (e-mail: chxuwh@gmail.com).

Hongmei Chen and Tianrui Li are with the School of Computing and Artificial Intelligence, Institute of Artificial Intelligence and National Engineering Laboratory of Integrated Transportation Big Data Application Technology, Southwest Jiaotong University, Chengdu 611756, China (e-mail: hmchen@swjtu.edu.cn; trli@swjtu.edu.cn).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TFUZZ.2023.3272316.

Digital Object Identifier 10.1109/TFUZZ.2023.3272316

preference relation [28], dominance-based soft rough sets [29], dominance-based rough sets of multiscale intuitionistic fuzzy decision tables [30], soft dominance based rough sets [31], and composite dominance-based rough sets [32], etc. In order to more accurately describe the uncertainty from fuzzy and numerical ODS, Hu et al. [33] proposed a fuzzy preference-based rough set model, also known as the fuzzy dominance rough set (FDRS) model. This model inherits the advantages of FRS model and is an important knowledge acquisition model for ordinal classification. On the basis of FDRS model, the dominance-based rough fuzzy sets [34], fuzzy dominance neighborhood rough sets [35], and quantitative dominance neighborhood rough sets [36] are successively proposed. Although scholars have expanded some new models based on the FDRS model according to different requirements, its robustness has not been improved. In particular, mislabeled samples can greatly mislead the approximate computation results of the FDRS model [37], ultimately leading to wrong decision-making information for users. Therefore, in order to improve the defects of FDRS model in dealing ODS with noisy samples, it is necessary to explore a robust FDRS model in this article.

Feature selection (also known as attribute reduction) is one of the important applications of the FRS theory, which aims to mine useful and important features from high-dimensional and redundant data for subsequent learning tasks [38], [39], [40]. Feature selection methods for the ODS have also received much attention. Hu et al. [33] proposed a forward greedy feature selection method using FDRS-based dependency function as evaluation metric. Based on the classical feature selection algorithms ReliefF and Simba, Hu et al. [41] developed two feature selection algorithms O-ReliefF and O-Simba suitable for the ODS. Further, Hu et al. [1] proposed minimum-redundancy and maximum-relevance feature selection algorithm based on rank mutual information. Pan et al. [42] constructed a feature selection algorithm that maximizes the monotonic dependency function. Qian et al. [43] designed a fusing monotonic decision trees method based on maximal probability and developed a feature selection method with rank-preservation. Du and Hu [44] presented a feature selection method based on evidence theory for the ODS. Recently, Luo et al. [45] introduced a new fuzzy rank discrimination measure to characterize the uncertainty of monotonic classification, and designed a corresponding feature selection algorithm. However, the above feature selection methods for the ODS ignore three key factors that help to improve classification performance, which are the robustness of the evaluation metrics, the separability of classes, and the redundancy of features. This article will comprehensively consider these three factors to construct a new feature selection method for the ODS. In what follows, the importance of these three key factors for feature selection are analyzed, respectively.

Undoubtedly, an evaluation metric with good robustness is necessary for feature selection algorithms, which can accurately identify useful features for classification in noisy environments. Hu et al. [46] proposed a robust attribute reduction method based on soft FRSs. Zhu et al. [47] constructed a robust unsupervised spectral feature selection method by preserving local and global structures. Based on earth movers distance, Qu et al. [48] introduced a robust attribute reduction method using a strategy that minimizes the inconsistency between the discernibility of the reduct and the entire original attribute set. In practical applications, Dong et al. [49] proposed a key energy-consumption feature selection of thermal power systems based on robust attribute reduction with rough sets. It follows that robustness is very important for feature selection algorithms. However, the above robust feature selection algorithms do not care about ordinal classification, so they cannot effectively complete the feature selection tasks for ordered data. Thus, in this article, feature selection algorithms for ordinal classification should consider improving their robustness.

From the perspective of classification, the separability of classes in a certain feature subset space is a key indicator for evaluating the importance of this feature subset to classification. For multidimensional time series data in medicine, Fang et al. [50] developed an improved feature selection method that integrates the Kozachenko-Leonenko information entropy estimation method and the feature selection method based on class-separability (CS). Zhou et al. [51] designed an online feature selection considering separability between classes for high-dimensional class-imbalanced data. Chen et al. [52] proposed a spectral feature selection approach using kernel fuzzy rough approximation to describe the separability of classes. Hu et al. [12] proposed a robust attribute reduction method considering separability in fuzzy decision systems. Obviously, the separability of classes is closely related to the distribution of samples. Naturally, the effect of noisy samples on CS is also great. However, these feature selection methods considering CS do not design the metrics of CS from the perspective of antinoise performance. Consequently, it is necessary to design a robust metric to characterize the CS for ordinal classification. On the other hand, the redundancy between features is one of the important factors considered in feature selection algorithms [53], [54], [55], [56], which describes the repeated classification information provided by features [1]. Therefore, the feature selection method proposed for the ODS in this article not only considers the CS but also needs to explore the redundancy between features.

In view of the above analysis and investigation, this article develops a robust feature selection method considering CS and feature-redundancy for ordinal classification. The main contributions of this article can be summarized as follows.

1) An adaptive *K*-nearest neighbors (AKNN) strategy is proposed to calculate the density of samples. The noisy samples are identified according to their densities, and then a robust fuzzy dominance rough set model with active antinoise performance is proposed (i.e., AAnFDRS model).
2) The CS is evaluated by the approximation operators of the AAnFDRS model, and the feature-redundancy is evaluated by the fuzzy ranking conditional mutual information (FRCMI). On this basis, a feature evaluation index that comprehensively considers CS and feature-redundancy is constructed to evaluate the importance of features.
3) A novel feature selection algorithm (i.e., AAnFSCF) is designed to select the feature subset with the highest classification performance, and its time complexity is analyzed.

4) The classification performance, robustness, statistical tests, and parameter sensitivity analysis of the proposed algorithm are tested in the experiments and the results show that our algorithm has not only has better classification performance but also good robustness.

The rest of this article is organized as follows. Section II briefly reviews the basic concepts of ODS and FDRS. In Section III, an active antinoise FDRS using AKNN is proposed. Section IV proposes the feature evaluation index, and designs the feature selection algorithm. The experimental results and analysis are shown in Section V. Finally, Section VI concludes the article.

II. Preliminaries

This section introduces some related knowledge about ODS [57], [58] and FDRS [33].

A. Ordered Decision Systems (ODSs)

An ODS is a four-tuple $S^{\preceq}=\langle U, C\cup D, V\rangle$, where $U=\{x_i\}_{i=1}^{n}$ is a nonempty finite set of samples, $C=\{c_k\}_{k=1}^{m}$ is a nonempty finite set of conditional attributes, D is a nominal decision attribute, and V is the value domain of the entire information system. The V is composed of two parts, one is the value domain $\bigcup_{k=1}^{m} V_{c_k}$ ($V_{c_k}=\{v(x_i,c_k)|\forall x_i\in U\}$) under the condition attribute set C, and the other is the value domain $V_D=\{d_t\}_{t=1}^{T}$ under the decision attribute D.

In addition, the division of U under D is denoted as $U/D=\{Cl_t\}_{t=1}^{T}$, where $Cl_t=\{x_i\in U|v(x_i,D)=d_t, t\in\{1,2,\ldots,T\}\}$ is the decision class. Assuming that the decision values obey the preference order relationship, i.e., $d_1<\cdots<d_t<\cdots<d_T$, then the preference order relationship between decision classes as $Cl_1\prec\cdots\prec Cl_t\prec\cdots\prec Cl_T$. In FDRS, the upward union $Cl_t^{\succeq}=\bigcup Cl_{t'}(t\le t')$ and downward union $Cl_t^{\preceq}=\bigcup Cl_{t'}(t\ge t')$, where $t,t'\in\{1,2,\ldots,T\}$ are the two target sets to be approximated. In this article, we only consider the approximation operators for the upward union of the decision classes, because the the approximation operators for the downward union of the decision classes can also be similarly explored.

B. Fuzzy Dominance Rough Sets (FDRSs)

In an ODS, the fuzzy dominance relation between x_i and x_j under c_k is calculated by

$$R_{c_k}^{\preceq}(x_i,x_j)=1/\left(1+exp\left(-k(v(x_j,c_k)-v(x_i,c_k))\right)\right) \quad (1)$$

where k is a positive integer.

For any feature subset $P\subseteq C$, $R_P^{\preceq}(x_i,x_j)=\min_{c_k\in P} R_{c_k}^{\preceq}(x_i,x_j)$ indicates the extent to which x_j is better than x_i under P, and it can be abbreviated as $R_{(i,j)}^{\preceq P}$. The knowledge granules about x_i induced by $R_P^{\preceq}$ can be divided into fuzzy dominating set and fuzzy dominated set, which are denoted as follows:

$$R_P^{+}(x_i)=\left(R_{(i,1)}^{\preceq P},R_{(i,2)}^{\preceq P},\ldots,R_{(i,n)}^{\preceq P}\right), \quad (2)$$

$$R_P^{-}(x_i)=\left(R_{(1,i)}^{\preceq P},R_{(2,i)}^{\preceq P},\ldots,R_{(n,i)}^{\preceq P}\right). \quad (3)$$

Definition 1: For any $P\subseteq C$ and $t\in\{1,2,\ldots,T\}$, the fuzzy lower and upper approximations of the $Cl_t^{\preceq}$ under P are defined, respectively, as follows:

$$\underline{R_P^{\preceq}}(Cl_t^{\preceq})(x_i)=\inf_{x_j\in U}\max\left(1-R_P^{+}(x_i)(x_j),Cl_t^{\preceq}(x_j)\right), \quad (4)$$

$$\overline{R_P^{\preceq}}(Cl_t^{\preceq})(x_i)=\sup_{x_j\in U}\min\left(R_P^{-}(x_i)(x_j),Cl_t^{\preceq}(x_j)\right). \quad (5)$$

In [33], it was stipulated that $\underline{R_P^{\preceq}}(Cl_1^{\preceq})(x_i)=1$ and $\overline{R_P^{\preceq}}(Cl_1^{\preceq})(x_i)=1$. Moreover, (4) and (5) are simplified, respectively, as follows:

$$\underline{R_P^{\preceq}}(Cl_t^{\preceq})(x_i)=\inf_{x_j\notin Cl_t^{\preceq}}\left(1-R_P^{+}(x_i)(x_j)\right), \quad (6)$$

$$\overline{R_P^{\preceq}}(Cl_t^{\preceq})(x_i)=\sup_{x_j\in Cl_t^{\preceq}}\left(R_P^{-}(x_i)(x_j)\right). \quad (7)$$

From (6) and (7), it can be clearly observed that $\forall x_i\in U$, the membership of x_i to fuzzy set $\underline{R_P^{\preceq}}(Cl_t^{\preceq})$ is determined by the best sample that does not belong to class $Cl_t^{\preceq}$, and the membership of x_i to fuzzy set $\overline{R_P^{\preceq}}(Cl_t^{\preceq})$ is determined by the worst sample that belongs to class $Cl_t^{\preceq}$. Through the above analysis, we find that the approximation operators based on FDRS model are not robust to noise samples with wrong labels. The fundamental reason is that the degree to which a sample belongs to the approximation sets in Definition 1 depends on the best and worst samples in the sample set. This sensitive statistical rules have no fault-tolerant performance, and the calculation of approximation by this way directly lead to the instability of the approximation results to noise interference. This defect makes FDRS-based approximation operators very susceptible to noise samples with wrong labels in practical applications. Therefore, it is necessary to improve the robustness of the FDRS model.

III. Active Antinoise FDRSs Using AKNN

This section proposes an active antinoise fuzzy dominance rough set (AAnFDRS) model for ODS. This model uses an AKNN strategy to calculate the density of samples, and achieves the purpose of active antinoise by filtering out samples with low density. The relevant definitions are introduced as follows.

A. Density Calculation Method Using AKNN Strategy

In classification tasks, the density of samples is a key indicator for judging whether a sample is an abnormal sample (or called an outlier). The samples with higher density values are usually more likely to be correctly classified, and vice versa. In this subsection, we design a K-nearest neighbors strategy to calculate the density of samples. The traditional K-nearest neighbors strategy contains two key parameters, which are the number of nearest neighbors samples (K) and the distance (radius, r) between them. These two parameters are generally set subjectively, and they do not have good adaptability and generalization. Therefore, we actively learn these two parameters by considering the data distribution, and design the AKNN strategy to make it have better

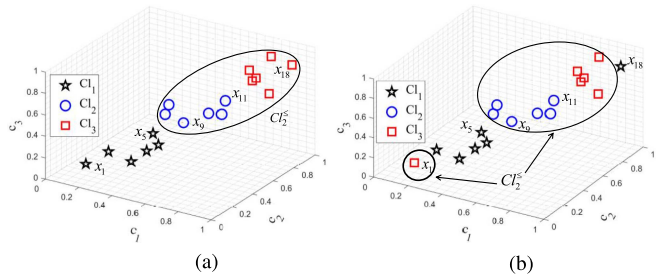


Fig. 1. Distribution of the ODS in no-noise and noisy environments. (a) The ODS without noise samples. (b) The ODS with two mislabeled samples.

adaptability and generalization. First, the adaptive parameters $\boldsymbol{r}$ and $\boldsymbol{K}$ are defined as follows.

Definition 2: For any $P \subseteq C$, the adaptive parameters $\boldsymbol{r}_P^{Cl_t^{\preceq}}$ of the upward union $Cl_t^{\preceq}$ under P is defined as follows:

$$\boldsymbol{r}_P^{Cl_t^{\preceq}} = \frac{1}{|Cl_t^{\preceq}|^2} \sum_{i=1}^{|Cl_t^{\preceq}|} \sum_{j=1}^{|Cl_t^{\preceq}|} \Delta_P^{Cl_t^{\preceq}}(x_i, x_j) \tag{8}$$

where $x_i, x_j \in Cl_t^{\preceq}$, the $\Delta_P^{Cl_t^{\preceq}}(x_i, x_j)$ represents the distance between x_i and x_j under P, which is calculated using the Euclidean distance.

Then, the parameter $\boldsymbol{K}_P^{Cl_t^{\preceq}}$ is learned based on the adaptive parameter $\boldsymbol{r}_P^{Cl_t^{\preceq}}$, which is defined as follows.

Definition 3: For any $P \subseteq C$, the adaptive parameters $\boldsymbol{K}_P^{Cl_t^{\preceq}}$ of the upward union $Cl_t^{\preceq}$ under P is defined as follows:

$$\boldsymbol{K}_P^{Cl_t^{\preceq}} = \frac{1}{|Cl_t^{\preceq}|} \sum_{i=1}^{|Cl_t^{\preceq}|} N_P^{Cl_t^{\preceq}}\left(x_i, \boldsymbol{r}_P^{Cl_t^{\preceq}}\right) \tag{9}$$

where $x_i \in Cl_t^{\preceq}$, the $N_P^{Cl_t^{\preceq}}(x_i, \boldsymbol{r}_P^{Cl_t^{\preceq}})$ represents the number of samples in the neighborhood of x_i with radius $\boldsymbol{r}_P^{Cl_t^{\preceq}}$.

Finally, the density of each sample in $Cl_t^{\preceq}$ is calculated based on the learned parameters $\boldsymbol{r}_P^{Cl_t^{\preceq}}$ and $\boldsymbol{K}_P^{Cl_t^{\preceq}}$, which is defined as follows.

Definition 4: For any $P \subseteq C$, the density of $x_i \in Cl_t^{\preceq}$ under P is defined as follows:

$$\boldsymbol{\rho}_P^{Cl_t^{\preceq}}(x_i) = \frac{N_P^{Cl_t^{\preceq}}(x_i, \boldsymbol{r}_P^{Cl_t^{\preceq}})}{\boldsymbol{K}_P^{Cl_t^{\preceq}}} \tag{10}$$

where $x_i \in Cl_t^{\preceq}$, the $N_P^{Cl_t^{\preceq}}(x_i, \boldsymbol{r}_P^{Cl_t^{\preceq}})$ represents the number of samples in the neighborhood of x_i with radius $\boldsymbol{r}_P^{Cl_t^{\preceq}}$.

Example 1: Fig. 1 shows the feature space distribution of an ODS in no-noise [as shown in Fig. 1(a)] and noise [as shown in Fig. 1(b), where x_1 and x_{18} are two noise samples] environments, where pentagrams, circles, and squares stand for samples coming from classes 1, 2, and 3, respectively. In the ODS, $U = \{x_i\}_{i=1}^{18}$, $C = \{c_1, c_2, c_2\}$, $Cl_1^{\preceq} = U$, $Cl_2^{\preceq} = \{x_7, x_8, \ldots, x_{18}\}$, and $Cl_3^{\preceq} = \{x_{13}, x_{12}, \ldots, x_{18}\}$.

In Fig. 1(a), according to Definitions 2–4, the densities of x_1 and x_{18} are calculated as $\boldsymbol{\rho}_C^{(Cl_2^{\preceq})^c}(x_1) = 0.67$ and $\boldsymbol{\rho}_C^{Cl_2^{\preceq}}(x_{18}) = 0.88$, respectively. In Fig. 1(b), the x_1 is misclassified into $Cl_2^{\preceq}$ and x_{18} is misclassified into $(Cl_2^{\preceq})^c$, and their densities are calculated as $\boldsymbol{\rho}_C^{Cl_2^{\preceq}}(x_1) = 0.14$ and $\boldsymbol{\rho}_C^{(Cl_2^{\preceq})^c}(x_{18}) = 0.23$, respectively. The above results show that the densities of the samples are greatly reduced due to the samples being mislabeled. At the same time, it also shows that the density calculation method proposed in this article can effectively identify abnormal samples.

B. Approximations in AAnFDRS Model

When constructing approximation operators with antinoise performance, the focus is on designing approximate calculation rules that can filter noise samples. This rule is used to filter out the noise samples existing in the classes boundary, and then screens out the appropriate correct samples (i.e., the best samples outside the classes and the worst samples inside the classes) to calculate the approximations. However, a very key issue is how to control the number of ignored samples in the process of filtering samples. If there are few samples filtered out, there may still be noisy samples in the classes boundary. Conversely, if the samples are filtered out excessively, it may cause overfitting noise. Therefore, how to find the critical value of filtering out the number of samples in the process of designing approximate calculation rules is the key to solving this issue. To address this issue, we design density critical (DC) values for the lower and upper approximations, respectively, which are defined as follows.

Definition 5: For lower approximation, $\forall P \subseteq C$, $x_i \in Cl_t^{\preceq}$, the DC value of samples outside the upward union $Cl_t^{\preceq}$ is defined as follows:

$$\underline{\boldsymbol{\rho}_{\boldsymbol{DC}}^{(\boldsymbol{Cl_t^{\preceq}})^c}}(x_i) = \inf_{x_j \in (Cl_t^{\preceq})^c} \max\left\{1 - \boldsymbol{\rho}_P^{(Cl_t^{\preceq})^c}(x_j) R_P^+(x_i)(x_j), \boldsymbol{\rho}_P^{(Cl_t^{\preceq})^c}(x_j)\right\}. \tag{11}$$

Definition 6: For upper approximation, $\forall P \subseteq C$, $x_i \in Cl_t^{\preceq}$, the DC value of samples inside the upward union $Cl_t^{\preceq}$ is defined as follows:

$$\overline{\boldsymbol{\rho}_{\boldsymbol{DC}}^{\boldsymbol{Cl_t^{\preceq}}}}(x_i) = \inf_{x_j \in Cl_t^{\preceq}} \max\left\{1 - \boldsymbol{\rho}_P^{Cl_t^{\preceq}}(x_j) R_P^-(x_i)(x_j), \boldsymbol{\rho}_P^{Cl_t^{\preceq}}(x_j)\right\}. \tag{12}$$

Then, the definitions of approximations in AAnFDRS are introduced as follows.

Definition 7: For any $P \subseteq C$, the fuzzy lower approximation of the upward union $Cl_t^{\preceq}$ under P is defined as follows:

$$\underline{\mathcal{R}_P^{\preceq}}(Cl_t^{\preceq})(x_i) = \inf_{x_j \in Lo} \{1 - R_P^+(x_i)(x_j)\}, \tag{13}$$

where $Lo = \{x_j | \boldsymbol{\rho}_P^{(Cl_t^{\preceq})^c}(x_j) \geq \underline{\boldsymbol{\rho}_{\boldsymbol{DC}}^{(\boldsymbol{Cl_t^{\preceq}})^c}}(x_i), x_j \in (Cl_t^{\preceq})^c\}$.

Definition 8: For any $P \subseteq C$, the fuzzy upper approximation of the upward union $Cl_t^{\preceq}$ under P is defined as follows:

$$\overline{\mathcal{R}_P^{\preceq}}(Cl_t^{\preceq})(x_i) = \sup_{x_j \in Up} \{R_P^-(x_i)(x_j)\} \tag{14}$$

where $Up = \{x_j | \boldsymbol{\rho}_P^{Cl_t^{\preceq}}(x_j) \geq \overline{\boldsymbol{\rho}_{\boldsymbol{DC}}^{\boldsymbol{Cl_t^{\preceq}}}}(x_i), x_j \in Cl_t^{\preceq}\}$.

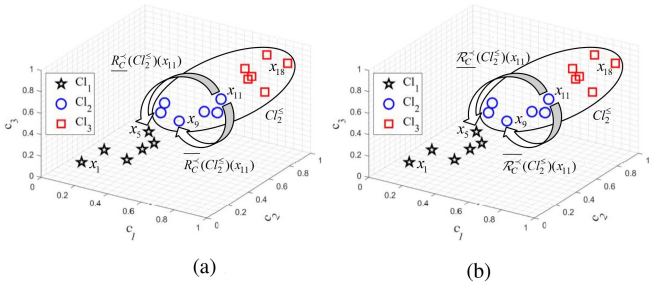


Fig. 2. Approximate calculation rules of FDRS and AAnFDRS in no-noise environment. (a) FDRS. (b) AAnFDRS.

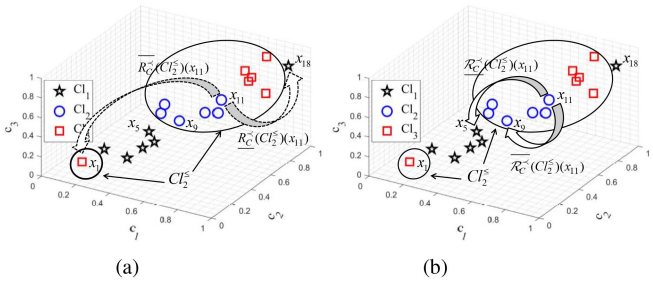


Fig. 3. Approximate calculation rules of FDRS and AAnFDRS in noise environment. (a) FDRS. (b) AAnFDRS.

Furthermore, we specify that $\underline{\mathcal{R}_P^{\prec}}(Cl_1^{\preceq})(x_i) = 1$ and $\overline{\mathcal{R}_P^{\prec}}(Cl_1^{\preceq})(x_i) = 1$.

C. Example Explanation

This section demonstrates the antinoise mechanism of the approximation operators in the AAnFDRS model with an example.

Example 2: Continuing from Example 1. Fig. 2 shows the approximate calculation rules of FDRS and AAnFDRS in the no-noise environment, respectively.

It can be seen from Fig. 2(a) that the degree to which x_{11} belongs to the lower approximation and the upper approximation of $Cl_2^{\preceq}$ is determined by the best sample x_5 outside $Cl_2^{\preceq}$ and the worst sample x_9 inside $Cl_2^{\preceq}$, respectively. Their calculation results as $\underline{R_C^{\prec}}(Cl_2^{\preceq})(x_{11}) = 0.94$ and $\overline{R_C^{\prec}}(Cl_2^{\preceq})(x_{11}) = 0.79$, where $k = 10$ in (1). Fig. 2(b) also presents the same approximate calculation rule. Their calculation results as $\underline{\mathcal{R}_C^{\prec}}(Cl_2^{\preceq})(x_{11}) = 0.94$ and $\overline{\mathcal{R}_C^{\prec}}(Cl_2^{\preceq})(x_{11}) = 0.79$. The above results show that the approximate calculation results of the FDRS model and the AAnFDRS model are consistent in the no-noise environment. Through the above analysis, it can be found that the proposed AAnFDRS model inherits the calculation calculation rules of the FDRS model, and its approximate calculation rules are reasonable.

Fig. 3 shows the approximate calculation rules of FDRS and AAnFDRS in the noise environment, respectively.

From Fig. 3(a), we can find that the degree to which x_{11} belongs to the lower approximation and the upper approximation of $Cl_2^{\preceq}$ is determined by the noise samples x_{18} and x_1, respectively. Their calculation results as $\underline{R_C^{\prec}}(Cl_2^{\preceq})(x_{11}) = 0.12$ and $\overline{R_C^{\prec}}(Cl_2^{\preceq})(x_{11}) = 0.99$. However, in Fig. 3(b), the approximate calculation rules for x_{11} avoid the influence of the noise samples x_1 and x_{18}. Their calculation results as $\underline{\mathcal{R}_C^{\prec}}(Cl_2^{\preceq})(x_{11}) = 0.94$ and $\overline{\mathcal{R}_C^{\prec}}(Cl_2^{\preceq})(x_{11}) = 0.79$. Based on the above calculation results, it is not difficult to find that the approximate calculation results based on FDRS model in noisy and no-noise environments are very different. Especially the calculation results in the noisy environment are obviously inconsistent with our intuitive reasoning. Because of $x_{11} \in Cl_2^{\preceq}$, intuitively $\underline{R_C^{\prec}}(Cl_2^{\preceq})(x_{11})$ should get higher values, but the rules for calculating approximations according to the FDRS model cause it to get a very small value. However, the approximate calculation results based on the AAnFDRS model are reasonable and consistent with our intuition reasoning, which are consistent with the calculation results in the no-noise environment. In Fig. 3(b), according to Definitions 5 and 6, the DC values are calculated as $\underline{\boldsymbol{\rho}_{DC}^{(\boldsymbol{Cl_2^{\preceq}})^c}}(x_{11}) = 0.80$ and $\overline{\boldsymbol{\rho}_{DC}^{\boldsymbol{Cl_2^{\preceq}}}}(x_{11}) = 0.86$, respectively. The densities of these two noise samples are $\boldsymbol{\rho}_C^{(Cl_2^{\preceq})^c}(x_{18}) = 0.23$ and $\boldsymbol{\rho}_C^{Cl_2^{\preceq}}(x_1) = 0.14$. Because of $\boldsymbol{\rho}_C^{(Cl_2^{\preceq})^c}(x_{18}) < \underline{\boldsymbol{\rho}_{DC}^{(\boldsymbol{Cl_2^{\preceq}})^c}}(x_{11})$ and $\boldsymbol{\rho}_C^{Cl_2^{\preceq}}(x_1) < \overline{\boldsymbol{\rho}_{DC}^{\boldsymbol{Cl_2^{\preceq}}}}(x_{11})$, the noise samples x_1 and x_{18} are ignored when computing the lower and upper approximations, respectively. After screening, samples x_5 and x_9 are respectively used as the most appropriate samples to calculate the approximations. This result shows that the AAnFDRS-based approximate calculation rules can effectively filter out the noise samples, and then select the appropriate samples to obtain the correct approximate calculation values.

Through the above analysis, it can be concluded that the AAnFDRS model has good robustness. Based on the distribution of the data, this model actively filters out noise samples by calculating the density of the samples, and selects proper samples from the process to effectively calculate the approximations, thereby achieving the purpose of antinoise.

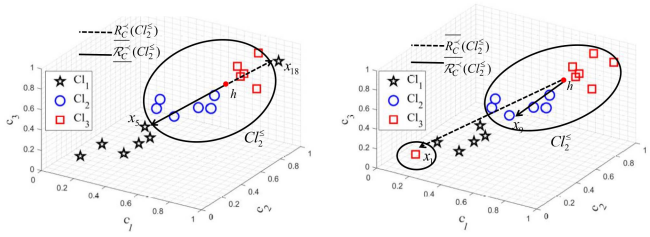
IV. AAnFDRS-Based Feature Selection Approach Exploring CS and Feature-Redundancy for ODS

This section proposes an AAnFDRS-based feature selection approach for ODS, which comprehensively explores CS and feature-redundancy.

A. Problem Statement

The separability of classes is characterized by the degree of between-class dispersion (BCD) and the degree of intraclass aggregation (ICA). The degree of BCD is usually measured by the distance between different classes, and the degree of ICA is usually measured by the distance between samples within the same class. As we all know, in classification tasks, the larger the between-class distance, the better; the smaller the intraclass distance, the better.

The between-class distance refers to the distance between the center point h of the selected class and the sample closest to the h point from different classes; the intraclass distance refers to the distance between the h point and the sample that is farthest



(a) (b)

Fig. 4. Degrees of between-class dispersion and intraclass aggregation for ODS in noisy environments. (a) The degrees of between-class dispersion based on FDRS and AAnFDRS, where h is the center point of $Cl_2^{\prec}$ and x_{18} is the noise sample. (b) The degree of intraclass aggregation based on FDRS and AAnFDRS, where h is the center point of $Cl_2^{\prec}$ and x_1 is the noise sample.

from the h point in the same class. For ordinal classification tasks, the between-class distance refers to the distance between the center point h of the selected class and the best sample from the different class; the intraclass distance refers to the distance between the h point and the worst sample in same class. This rule of searching for the best sample in different classes and the worst sample in same classes is very similar to the approximate calculation rules based on FDRS model. Therefore, for an ODS, it is reasonable to use the FDRS-based lower approximation and upper approximation to characterize the degree of BCD and the degree of ICA, respectively. However, as mentioned in Section II-B, this search strategy of statistical minima and maxima is sensitive and not robust to noisy samples. Next, this defect is revealed through Fig. 4.

From Fig. 4, we find that the BCD degree of class $Cl_2^{\prec}$ is determined by the noise sample x_{18} in the fuzzy dominance rough approximation space, i.e., $\underline{R_C^{\prec}}(Cl_2^{\prec})$ describes the dispersion degree of $Cl_2^{\prec}$ [as shown in Fig. 4(a)]; the ICA degree is determined by the noise sample x_1, i.e., $\overline{R_C^{\prec}}(Cl_2^{\prec})$ describes the aggregation degree of $Cl_2^{\prec}$ [as shown in Fig. 4(b)]. Obviously, this measure strategy is disturbed by noise samples in the fuzzy dominance rough approximation space, resulting in the output of erroneous discriminative information. Therefore, an antinoise search rule is needed to avoid the influence of noise samples in the process of calculating the BCD degree and ICA degree.

As mentioned in Section III, the proposed approximate calculation rules based on AAnFDRS model have the function of actively filtering out noisy samples and searching for the correct out-of-class best samples and the in-class worst samples. By using the approximate operators of the AAnFDRS model to describe the BCD degree and ICA degree, the interference of noise samples can be avoided, and the correct discrimination information can be output. As shown in Fig. 4, based on AAnFDRS model, the BCD degree of the class $Cl_2^{\prec}$ is determined by the sample x_5, i.e., $\underline{\mathcal{R}_C^{\prec}}(Cl_2^{\prec})$ describes the dispersion degree of $Cl_2^{\prec}$ [as shown in Fig. 4(a)]; the ICA degree is determined by the sample x_9, i.e., $\overline{\mathcal{R}_C^{\prec}}(Cl_2^{\prec})$ describes the aggregation degree of $Cl_2^{\prec}$ [as shown in Fig. 4(b)]. Clearly, this measure strategy is consistent with our intuitive reasoning. Therefore, it is reasonable to measure the BCD degree and the ICA degree based on the proposed AAnFDRS model.

B. Class-Separability

This subsection defines the BCD degree, ICA degree, and the CS based on the AAnFDRS model.

Definition 9: For any $P \subseteq C$ and $t \in \{2, 3, \ldots, T\}$, the BCD degree of $Cl_t^{\prec}$ under P is defined as follows:

$$\mathcal{BCD}_P(Cl_t^{\prec}) = \frac{\sum_{i=1}^{|Cl_t^{\prec}|} \underline{\mathcal{R}_P^{\prec}}(Cl_t^{\prec})(x_i)}{|Cl_t^{\prec}|}. \tag{15}$$

Definition 10: For any $P \subseteq C$ and $t \in \{2, 3, \ldots, T\}$, the ICA degree of $Cl_t^{\prec}$ under P is defined as follows:

$$\mathcal{ICA}_P(Cl_t^{\prec}) = \frac{\sum_{i=1}^{|Cl_t^{\prec}|} \overline{\mathcal{R}_P^{\prec}}(Cl_t^{\prec})(x_i)}{|Cl_t^{\prec}|}. \tag{16}$$

Since it is difficult to find the center points of the classes in the ODS, we use the average approximation results from all samples in the classes to characterize the BCD and ICA in Definitions 9 and 10, respectively. For an ODS, the larger the BCD, the better; the smaller the ICA, the better. The CS of single and global in the ODS are defined, respectively, as follows.

Definition 11: For any $P \subseteq C$ and $t \in \{2, 3, \ldots, T\}$, the CS of $Cl_t^{\prec}$ under P is defined as follows:

$$\mathcal{CS}_P(Cl_t^{\prec}) = \frac{\mathcal{BCD}_P(Cl_t^{\prec})}{\mathcal{ICA}_P(Cl_t^{\prec})}. \tag{17}$$

For an ODS, the larger the CS, the better.

Definition 12: For any $P \subseteq C$ and $t \in \{2, 3, \ldots, T\}$, the global class-separability (GCS) under P is defined as follows:

$$\mathcal{GCS}_P(D) = \frac{\sum_{t=2}^{T} \mathcal{CS}_P(Cl_t^{\prec})}{T-1}. \tag{18}$$

For an ODS, the larger the GCS, the better.

Definition 13: For any $P \subseteq C$ and $c_k \in C - P$, the GCS-based feature significance of c_k to P is defined as follows:

$$\mathcal{SIG}(c_k, P, D) = \mathcal{GCS}_{P \cup \{c_k\}}(D) - \mathcal{GCS}_P(D). \tag{19}$$

The larger $\mathcal{SIG}(c_k, P, D)$ is, the more important c_k is for ordinal classification.

C. Feature-Redundancy

In this section, FRCMI is defined to measure the redundancy between features in the ODS.

Definition 14 ([1]): For any $P \subseteq C \cup D$, the fuzzy ranking entropy (FRE) of P is defined as follows:

$$\mathcal{RE}^{\prec}(P) = -\frac{1}{|U|} \sum_{i=1}^{n} \log_2 \frac{|R_P^+(x_i)|}{|U|} \tag{20}$$

which reflects the information amount of the feature subset P in the ODS.

Definition 15 ([1]): For any $P, Q \subseteq C \cup D$, the fuzzy ranking mutual information of P and Q is defined as follows:

$$\mathcal{RMI}^{\prec}(P; Q) = -\frac{1}{|U|} \sum_{i=1}^{n} \log_2 \frac{|R_P^+(x_i)| \cdot |R_Q^+(x_i)|}{|U| \cdot |R_{P \cup Q}^+(x_i)|} \tag{21}$$

which reflects the amount of overlapping information provided by feature subsets P and Q in the ODS.

Definition 16: For any $P, Q \in C$, when Q is known, then the FRCMI of the P and D is defined as follows:

$$\mathcal{RCMI}^{\prec}(P; D|Q) = -\frac{1}{|U|}\sum_{i=1}^{n}\log_2\frac{|R^+_{P\cup Q}(x_i)|\cdot|R^+_{Q\cup D}(x_i)|}{|R^+_{P\cup Q\cup D}(x_i)|\cdot|R^+_{Q}(x_i)|} \tag{22}$$

which is used to measure the amount of information provided by the P for ordinal classification when Q is known.

Let P $(P \subseteq C)$ denotes the selected feature subset and c_k $(c_k \in C - P)$ denotes the current candidate feature.

Definition 17: The redundancy between the P and c_k with respect to D is defined as follows:

$$\mathcal{RED}^{\prec}(P, c_k; D) = \mathcal{RMI}^{\prec}(c_k; D) - \mathcal{RCMI}^{\prec}(c_k; D|P) \tag{23}$$

which reflects the amount of overlapping ordinal classification information provided by the selected feature set P and current candidate feature c_k for decision D. That is, when P is known, c_k provides redundant information for ordinal classification. According to Definition 15, the $\mathcal{RMI}^{\prec}(c_k; D)$ denotes the amount of information provided by the current candidate feature c_k for decision D. According to Definition 16, the $\mathcal{RCMI}^{\prec}(c_k; D|P)$ denotes the amount of information provided by the current candidate feature c_k for decision D under the selected feature subset P. Obviously, under the selected feature subset P, the amount of redundant information provided by the c_k for D can be calculated by $\mathcal{RMI}^{\prec}(c_k; D) - \mathcal{RCMI}^{\prec}(c_k; D|P)$. Therefore, it is reasonable to calculate the redundant information of a current candidate feature by Definition 17. This redundancy is a negative factor for feature evaluation.

D. Feature Evaluation Index and Feature Selection Algorithm

In this section, a feature evaluation index is proposed, which comprehensively considers CS and feature-redundancy based on the AAnFDRS model. Then, the corresponding feature selection algorithm is constructed.

Definition 18: The feature evaluation index is defined as follows:

$$\mathcal{J}(c_k) = \mathcal{SIG}(c_k, P, D) - \beta \cdot \mathcal{RED}^{\prec}(P, c_k; D) \tag{24}$$

where the parameter β is used to adjust the redundancy between features.

Next, an example is used to illustrate Definition 18.

Example 3: Continuing from Example 2.

The feature evaluation index is calculated in the noisy environment, and its data distribution is shown in Fig. 3. Assuming c_3 as the selected feature, then calculate the evaluation index values $\mathcal{J}(c_1)$ and $\mathcal{J}(c_2)$, where β is set to 0.9. According to Definitions 9–13, the significance of features c_1 and c_2 is calculated as follows:

$$\mathcal{SIG}(c_1, c_3, D) = 1.33, \mathcal{SIG}(c_2, c_3, D) = 1.23.$$

According to Definitions 15–17, the redundancy of features c_1 and c_2 is calculated as follows:

$$\mathcal{RCMI}^{\prec}(c_1; D|c_3) = 0.37, \mathcal{RCMI}^{\prec}(c_2; D|c_3) = 0.39.$$

According to Definition 18, the evaluation index values of c_1 and c_2 is calculated as $\mathcal{J}(c_1) = 0.99$ and $\mathcal{J}(c_2) = 0.87$.

Subsequently, an AAnFDRS-based feature selection algorithm considering class-separability and feature-redundancy (AAnFSCF) is given in Algorithm 1, and the feature selection strategy is described as shown in Fig. 5.

The time complexity of this algorithm is analyzed as follows. Steps 2–19 are to calculate the GCS between each feature and decision, where Steps 3–18 are to calculate the CS between each feature and single class and its time complexity is $O(|U|^2 + 2\sum_{t=2}^{T}|Cl_t^{\prec}|)$. The time complexity of Steps 2–19 is $O(|C|(|U|^2 + 2\sum_{t=2}^{T}|Cl_t^{\prec}|))$. Steps 20–22 are to select the feature with the best GCS and add it to the P, then delete it from C. Steps 25–35 are to continuously select the feature with the best feature evaluation index and add it to the P until an feature sequence is obtained. Concretely, Steps 26–30 calculate the GCS-based feature significance and redundancy between each candidate features and selected feature subset, its time complexity is $O((|C|-1)(|U|^2 + 2\sum_{t=2}^{T}|Cl_t^{\prec}|))$. The time complexity of Steps 25–35 is $O(|C|^2(|U|^2 + 2\sum_{t=2}^{T}|Cl_t^{\prec}|))$. Steps 36–38 are to use a specified classifier to cross-validate the classification accuracy of each sequential feature subsets, it is calculated at most $|C| - 1$ times. Step 39–40 are to select the feature subset sequence P_{opt} with highest classification accuracy and output the feature subset. In summary, the time complexity of this algorithm is $O(|C|^2|U|^2)$.

V. Experiments and Analysis

In this section, the robustness and classification performance of the algorithm AAnFSCF are evaluated by a series of experiments, respectively.

A. Experimental Data Preprocessing

Table I shows that 12 datasets from UCI are used for experiments. These datasets are preprocessed in two steps. The first step is to normalize the values under the conditional attribute set in each original dataset to the interval $[0, 1]$ via the min–max normalization method. The second step is to monotonize each dataset after normalization, and the specific operations are as follows. First, calculate the average value of samples under the conditional attributes. Then, samples with a larger average value are assigned a larger class label, and samples with a smaller average value are assigned a smaller class label. Considering that the number of class labels is far from the number of samples, in the process of assigning class labels based on the average value, we adopted a method of assigning values in batches according to the number of class labels.

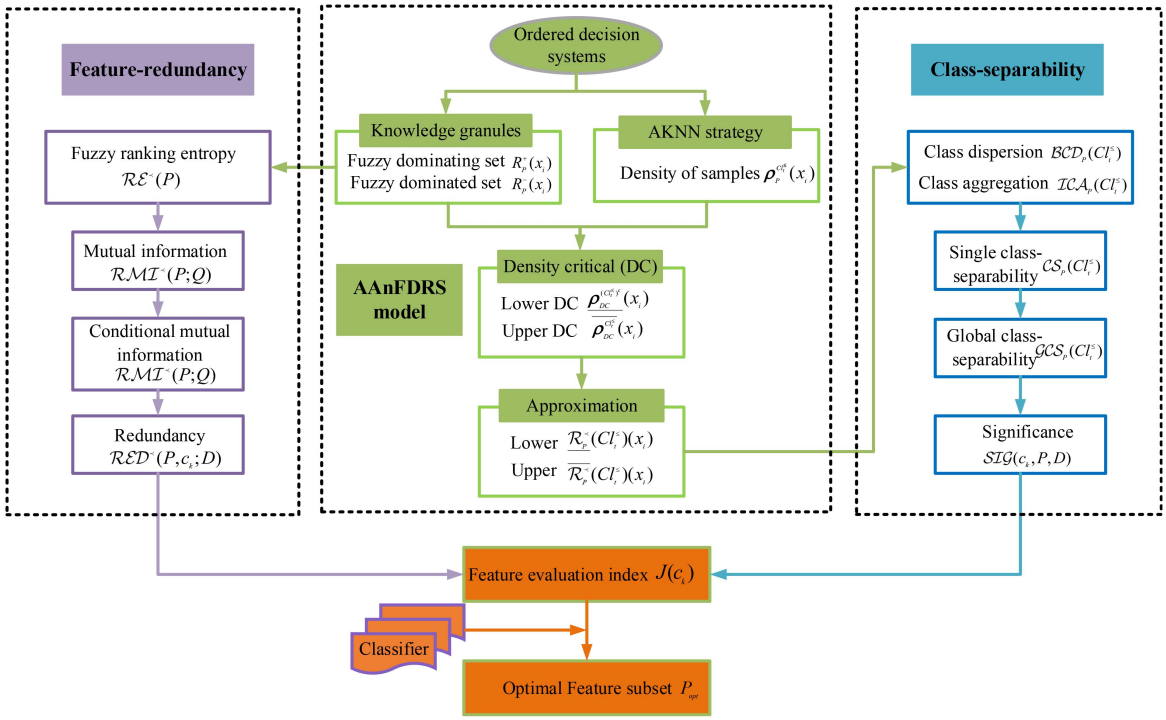


Fig. 5. Framework of AAnFSCF algorithm.

B. Classification Performance Evaluations of Algorithm AAnFSCF

The classification performance of the proposed algorithm AAnFSCF are evaluated by comparing with the existing feature selection methods applicable for the ODS.

1) Comparison Algorithms: The proposed algorithms are compared with seven feature selection (also known as attribute reduction) algorithms, including algorithms DRSQR, HARCC, AR-FDRS, AR-VCDRSA, mRMR-RMI, O-ReliefF, and O-Simba. Algorithm DRSQR is a DRSA-based QuickReduct method [58]. Algorithm HARCC is a heuristic attribute reduction method based on DRSA model [58]. Algorithm AR-FDRS is an attribute reduction method of forward greedy search based on FDRS model [33]. Algorithm AR-VCDRSA is an attribute reduction method of forward greedy search based on VCDRSA model [59]. In this experiment, for the algorithm AR-VCDRSA, we select the optimal feature subset of the consistency level l in the range of $[0.6, 0.9]$ with step size 0.1. Algorithm mRMR-RMI is an attribute reduction method with mRMR strategy using RMI as an uncertainty measure [1]. Algorithms O-Simba and O-ReliefF are extended by the classic algorithms ReliefF and Simba, respectively, for feature selection of ordered data [41]. The algorithms DRSQR, HARCC, AR-FDRS, and AR-VCDRSA each output an feature subset. The algorithms mRMR-RMI, O-ReliefF, and O-Simba each output an feature sequence, and their process of selecting the optimal feature subset is consistent with Steps 26–31 in Algorithm 1. Moreover, for the proposed algorithm AAnFSCF, we calculate the optimal feature subset of parameter β in the range [0, 1] with step size 0.1.

2) Experimental Preparation: First, assume that each original dataset is a clean dataset (i.e., no noise samples). Then, each data is randomly added 40% noise samples. Specifically, for each dataset, 40% samples with the maximum or minimum value of the class label are randomly selected, and then exchange the sample labels of different class labels. These relabeled samples are treated as noise samples.

The classification performance of algorithms are evaluated by classifiers Naive Bayes (NB), Support Vector Machine (SVM), and Classification And Regression Tree (CART). The default parameters setting of these classifiers are adopted. The ten-fold cross-validation is adopted to implement experiments. The original dataset is randomly divided into ten subsets, of which nine subsets are used as training data, and the remaining subset is used as test data. The experiments are repeated ten times, and the mean and standard deviation of the classification accuracy are calculated as the final result.

3) Experimental Results and Analysis: The classification performance of the eight feature reduction algorithms and the original data under different classifiers are recorded in Tables II, III, and IV. The classification performances are compared in terms of the mean classification accuracy (the left side of the "||" symbol) and the number of selected features (the right side of the "||" symbol), respectively. The parameters corresponding to classification accuracy are shown in parentheses after them. The last rows (Avg.) shows the average classification accuracy and the number of selected features of these algorithms on all datasets. The highest classification accuracy value corresponding to each data is highlighted in bold font.

The experimental results of Tables II–IV are analyzed as follows. Compared with the original data and seven contrasting

Algorithm 1: AAnFSCF Algorithm.

Input: A $S^{\preceq} = \langle U, C \cup D, V\rangle$ and parameter $\beta \in [0,1]$;
Output: The optimal feature subset P_{opt};

Initialize $P \leftarrow \emptyset$;
for *each* $c_k \in C$ **do**
  for t = *2 to* T **do**
    for *each* $x_{ti} \in Cl_t^{\preceq}$ **do**
      for *each* $x_{tj} \in (Cl_t^{\preceq})^c$ **do**
        Calculate the $\boldsymbol{\rho}^{(Cl_t^{\preceq})^c}(x_{tj})$ by Definitions 2-4;
        Calculate the $R_{c_k}^{+}(x_{ti})(x_{tj})$ by Eq. (1) $(k = 10)$;
      Calculate the $\underline{\boldsymbol{\rho}_{DC}^{(Cl_t^{\preceq})^c}}(x_{ti})$ by Definition 5;
      Calculate the $\underline{\mathcal{R}_{c_k}^{\preceq}}(Cl_t^{\preceq})(x_{ti})$ by Definition 7;
    for *each* $x_{ti} \in Cl_t^{\preceq}$ **do**
      for *each* $x_{tj} \in Cl_t^{\preceq}$ **do**
        Calculate the $\boldsymbol{\rho}^{Cl_t^{\preceq}}(x_{tj})$ by Definitions 2-4;
        Calculate the $R_{c_k}^{-}(x_{ti})(x_{tj})$ by Eq. (1);
      Calculate the $\overline{\boldsymbol{\rho}_{DC}^{Cl_t^{\preceq}}}(x_{ti})$ by Definition 6;
      Calculate the $\overline{\mathcal{R}_{c_k}^{\preceq}}(Cl_t^{\preceq})(x_{ti})$ by Definition 8;
    Calculate the $\mathcal{BCD}_{c_k}(Cl_t^{\preceq})$ by Definition 9;
    Calculate the $\mathcal{ICA}_{c_k}(Cl_t^{\preceq})$ by Definition 10;
    Calculate the $\mathcal{CS}_{c_k}(Cl_t^{\preceq})$ by Definition 11;
  Calculate the $\mathcal{GCS}_{c_k}(D)$ by Definition 12;
Find $c_k' = \underset{c_k}{arg}\ \underset{c_k \in C}{\max}\{\mathcal{GCS}_{c_k}(D)\}$;
$P \leftarrow P \cup \{c_k'\}$;
$C \leftarrow C - \{c_k'\}$;
$g \leftarrow 1$;
$l \leftarrow |C|$;
while $g < l$ **do**
  for *each* $c_k \in C$ **do**
    Calculate the $\mathcal{GCS}_{P \cup c_k}(D)$ by by Definition 12 // The detailed calculation process is consistent with Steps 2-19;
    Calculate the $\mathcal{SIG}(c_k, P, D)$ by Definition 13;
    Calculate the $\mathcal{RED}^{\preceq}(P, c_k; D)$ by Definition 17;
    Calculate the $\mathcal{J}(c_k)$ by Definition 18;
  Find $c_k' = \underset{c_k}{arg}\ \underset{c_k \in C}{\max}\{\mathcal{J}(c_p)\}$;
  $P \leftarrow P \cup \{c_k'\}$;
  $C \leftarrow C - \{c_k'\}$;
  $g \leftarrow g + 1$;
An feature sequence P is obtained, *i.e.*, $P = \{c_1', c_2', \cdots, c_l'\}$;
Let $P_1 = \{c_1'\}$, $P_2 = \{c_1', c_2'\}$, $\cdots$, $P_l = \{c_1', c_2', \cdots, c_l'\}$;
for $h = 1$ *to* l **do**
  Use a specified classifier to cross-validate the average classification accuracy $Acc(P_h)$ of the feature subset P_h;
$P_{opt} = \underset{P_h}{arg}\ \underset{P_h \subseteq P}{\max}\{Acc(P_h), h \in \{1, 2, \cdots, l\}\}$;
return P_{opt};

algorithms, our algorithm AAnFSCF wins 11 times, 12 times, and 9 times in Tables II–IV, respectively. However, other algorithms win up to two times. It shows that the classification performance of the algorithm AAnFSCF in most datasets is higher than that of the other algorithms and original feature set.

Moreover, it is not difficult to find that algorithms DRSQR and AR-FDRS have the worst classification performance. These

TABLE I
SUMMARY OF DATASETS

No.	Datasets	Abbreviation	Samples	Features	Classes
1	Wisconsin Prognostic Breast Cancer	WPBC	198	32	2
2	Sonar, mines versus rocks	Sonar	208	60	2
3	Ionosphere	Iono	351	34	2
4	Dermatology	Derm	358	34	6
5	Auto MPG	Auto	398	7	3
6	Housing	Hous	506	13	5
7	Wisconsin Diagnstic Breast Cancer	WDBC	569	30	2
8	Australian Credit	Aust	690	14	2
9	Credit Approval	Cred	690	14	2
10	German Credit	Cerm	1000	20	2
11	Wine Quality-red	Wred	1599	11	6
12	Wine Quality-white	Wite	4898	11	7

two algorithms are constructed based on the DRSA and FDRS models, respectively, and they have no fault tolerance for noise samples, which is the main reason for the poor classification performance of the two algorithms. Although the algorithm AR-VCDRSA uses a variable precision way (l is the fault-tolerance level) to construct a fault-tolerant mechanism, this algorithm is based on Boolean relation and cannot accurately describe the preference relationship in numerical features. The classification performance of algorithms mRMR-RMI and HARCC is moderate and close to the classification performance of the original data. The advantage of these two algorithms is that they comprehensively consider the relevance between features and decision, and the correlation between features. But these two algorithms still do not have antinoise mechanism. The algorithms O-Simba and O-ReliefF achieve better classification performance. The main reason is that these two algorithms use the information of class labels to select important features and avoid the influence of noise through multiple iterations. But their classification performance is still significantly worse than our algorithm AAnFSCF.

Overall, in terms of the size of the selected feature subset, the feature subset output by algorithm AAnFSCF is larger than that of algorithms DRSQR, AR-FDRS, and AR-VCDRSA, but smaller than that of algorithms HARCC, O-Simba, and O-ReliefF. The reason why algorithms DRSQR, AR-FDRS, and AR-VCDRSA obtain a smaller feature subset is that they unilaterally evaluate the dependence of features and decisions to determine the importance of features, and do not consider the classification information provided by class labels in the feature space and the correlation between features. Since the computational models of algorithms HARCC, O-Simba, and O-ReliefF are disturbed by noise samples, it is easy to lead to wrong feature evaluation results. Naturally, the feature subsets obtained by these four algorithms contain redundant or irrelevant features for classification. The feature subset obtained by our algorithm AAnFSCF is also much smaller than the original feature set. Compared with other algorithms, the algorithm AAnFSCF can accurately delete redundant features and better reduce the dimensionality of the feature space.

Subsequently, for feature selection algorithms that output feature sequences, we evaluate the impact of the number of

TABLE II
COMPARISON OF CLASSIFICATION PERFORMANCE OF DIFFERENT FEATURE SELECTION ALGORITHMS ON CLASSIFIER NB (%)

Datasets	Original data	DRSQR	HARCC	AR-FDRS	AR-VCDRSA(l)	mRMR-RMI	O-ReliefF	O-Simba	AAnFSCF (β)
WPBC	61.57±1.53‖32	60.30±0.49‖1	63.79±2.59‖13	55.51±1.10‖1	61.06±1.48(0.9)‖4	63.69±0.94‖2	61.16±2.12‖32	64.65±2.05‖6	**70.76±1.02(0.4)‖8**
Sonar	66.78±1.31‖60	63.94±0.00‖1	70.48±1.44‖9	64.52±0.59‖1	68.75±1.36(0.6)‖6	70.82±1.43‖18	68.08±0.82‖14	67.26±1.29‖49	**71.11±1.70(0.7)‖8**
Iono	60.46±0.63‖34	55.84±0.00‖1	59.00±0.80‖19	59.09±0.81‖2	61.48±0.38(0.7)‖10	61.77±0.86‖14	61.28±0.37‖11	61.79±0.56‖16	**63.79±0.84(0.5)‖4**
Derm	41.92±1.30‖34	20.45±0.27‖1	46.28±1.09‖30	19.46±0.49‖3	21.41±0.30(0.6)‖2	25.80±1.14‖33	49.05±1.48‖6	47.13±0.89‖5	**51.46±0.99(1.0)‖27**
Auto	72.07±0.28‖7	62.50±0.66‖3	73.01±0.70‖6	49.57±1.09‖1	72.14±0.49(0.6)‖7	72.81±0.36‖6	76.30±0.49‖5	71.79±0.54‖7	**78.57±0.50(0.5)‖1**
Hous	70.14±0.59‖13	38.40±1.14‖2	69.88±0.79‖12	56.54±0.94‖6	58.70±0.62(0.6)‖7	64.86±0.80‖12	71.46±0.84‖11	70.08±0.49‖13	**72.33±0.41(0.1)‖11**
WDBC	58.05±0.47‖30	55.69±0.53‖1	58.79±1.04‖22	57.63±0.37‖1	58.14±1.05(0.6)‖25	61.93±0.68‖11	59.19±0.64‖26	59.00±0.92‖29	**67.01±0.29(0.0)‖12**
Aust	54.74±0.67‖14	46.71±1.20‖1	54.67±0.55‖13	49.94±0.32‖4	55.13±0.44(0.6)‖5	56.41±0.50‖1	55.09±0.38‖10	55.07±0.24‖2	**58.78±1.14(0.8)‖13**
Cred	47.88±1.45‖14	**57.74±0.20‖1**	51.00±0.82‖13	42.13±0.10‖2	50.96±1.32(0.6)‖6	52.49±0.86‖2	**57.74±0.32‖1**	53.17±1.20‖2	53.38±2.07(0.7)‖10
Germ	55.56±1.10‖20	52.05±0.28‖1	55.78±0.47‖20	50.96±0.57‖2	59.44±0.38(0.9)‖1	59.46±0.45‖1	56.30±0.81‖19	58.39±0.61‖7	**66.54±0.83(0.0)‖17**
Wred	80.07±0.29‖11	20.18±0.37‖1	80.04±0.42‖11	20.15±0.66‖1	81.30±0.22(0.6)‖9	80.13±0.36‖10	80.99±0.21‖9	80.76±0.28‖10	**84.86±0.20(0.8)‖3**
Wite	73.24±0.19‖11	16.37±0.36‖1	73.45±0.16‖11	27.07±0.14‖3	65.78±0.20(0.6)‖3	40.63±0.29‖10	77.51±0.19‖9	74.24±0.28‖9	**78.49±0.17(0.3)‖4**
Avg.	59.39±0.80‖23.33	44.29±0.43‖1.25	59.87±0.85‖14.92	42.72±0.57‖2.25	55.30±0.72‖7.08	59.23±0.72‖10.00	61.86±0.70‖12.75	61.02±0.76‖12.92	**68.09±0.85‖9.83**

TABLE III
COMPARISON OF CLASSIFICATION PERFORMANCE OF DIFFERENT FEATURE SELECTION ALGORITHMS ON CLASSIFIER SVM (%)

Datasets	Original data	DRSQR	HARCC	AR-FDRS	AR-VCDRSA(l)	mRMR-RMI	O-ReliefF	O-Simba	AAnFSCF (β)
WPBC	59.49±1.52‖32	60.61±0.00‖1	60.56±0.16‖13	60.61±0.00‖1	60.61±0.00(0.6)‖1	60.86±1.10‖31	60.61±0.00‖9	60.61±0.00‖5	**73.08±0.98(0.1)‖21**
Sonar	71.83±0.88‖60	63.94±0.00‖1	64.09±0.46‖9	63.94±0.00‖1	63.94±0.00(0.6)‖6	72.40±1.31‖58	72.79±1.77‖56	72.69±0.87‖60	**73.99±1.41(0.4)‖56**
Iono	62.76±1.14‖34	55.84±0.00‖1	62.91±0.89‖19	55.19±0.76‖2	64.70±0.49(0.8)‖4	64.50±0.62‖9	63.90±0.95‖17	63.19±0.77‖30	**66.55±0.41(0.9)‖5**
Derm	63.46±1.31‖34	21.41±0.00‖1	64.37±1.51‖30	21.21±0.99‖3	22.87±0.95(0.7)‖4	32.68±1.33‖29	64.51±0.96‖31	64.08±1.73‖33	**65.69±1.71(1.0)‖22**
Auto	59.82±0.63‖7	52.04±0.88‖3	59.23±1.33‖6	44.34±1.00‖1	59.26±1.12(0.6)‖7	59.21±0.99‖6	59.59±1.26‖6	59.82±0.87‖7	**63.47±0.60(0.9)‖6**
Hous	77.87±0.54‖13	38.34±0.54‖2	78.34±0.58‖12	51.42±0.81‖6	50.40±0.96(0.6)‖7	61.07±0.79‖12	78.28±0.78‖13	78.22±0.44‖13	**78.52±0.37(0.1)‖11**
WDBC	59.00±0.44‖30	56.41±0.00‖1	59.00±0.60‖22	56.41±0.00‖1	59.44±0.65(0.6)‖25	59.65±0.48‖26	59.68±0.58‖27	59.56±0.17‖25	**63.29±0.31(0.1)‖13**
Aust	49.10±1.67‖14	47.23±1.16‖1	48.88±0.80‖13	49.19± 0.74‖4	52.74±2.15(0.8)‖3	54.39±0.80‖3	51.16±0.85‖10	52.10±0.25‖5	**60.46±0.87(0.9)‖13**
Cred	**58.41±0.00‖14**	58.12±0.00‖1	**58.41±0.00‖13**	58.12±0.00‖2	58.12±0.00(0.6)‖6	58.12±0.00‖1	**58.41±0.00‖14**	**58.41±0.00‖14**	**58.41±0.00(0.1)‖10**
Germ	51.88±0.90‖20	51.90±0.00‖1	51.52±0.89‖20	51.90±0.00‖2	52.89±0.63(0.6)‖3	53.79±0.77‖2	52.94±0.90‖11	52.75±0.65‖16	**60.07±0.34(0.0)‖16**
Wred	76.58±0.40‖11	19.41±0.15‖1	76.78±0.35‖11	19.31±0.40‖1	76.73±0.35(0.6)‖9	77.16±0.40‖10	76.72±0.28‖9	76.58±0.53‖11	**78.84±0.26(0.8)‖4**
Wite	83.73±0.20‖11	15.48±0.23‖1	83.74±0.12‖11	25.65±0.18‖3	64.06±0.13(0.6)‖3	39.58±0.39‖10	83.96±0.07‖9	83.62±0.20‖9	**84.64±0.12(0.8)‖8**
Avg.	64.49±0.80‖23.33	45.06±0.25‖1.25	63.98±0.64‖14.92	46.44±0.41‖2.25	57.15±0.62‖6.50	57.78±0.75‖16.42	65.21±0.70‖17.67	65.14±0.54‖19.00	**68.92±0.62‖15.42**

TABLE IV
COMPARISON OF CLASSIFICATION PERFORMANCE OF DIFFERENT FEATURE SELECTION ALGORITHMS ON CLASSIFIER CART (%)

Datasets	Original data	DRSQR	HARCC	AR-FDRS	AR-VCDRSA(l)	mRMR-RMI	O-ReliefF	O-Simba	AAnFSCF (β)
WPBC	61.01±3.21‖32	48.74±2.44‖1	61.57±1.96‖13	55.30±1.94‖1	64.90±2.21(0.8)‖4	69.29±2.13‖2	63.74±2.58‖16	61.31±1.49‖13	**73.08±2.68(0.1)‖16**
Sonar	64.23±3.13‖60	52.02±1.55‖1	66.44±2.30‖9	53.12±1.57‖1	66.54±1.99(0.9)‖6	66.87±3.57‖17	64.47±2.11‖47	**66.88±1.23‖24**	66.83±1.81(0.5)‖34
Iono	59.43±2.74‖34	55.84±0.00‖1	59.80±2.28‖19	52.36±1.35‖2	62.71±1.87(0.8)‖4	**64.33±1.69‖12**	61.85±2.01‖17	62.08±1.78‖27	62.56±1.50(0.0)‖28
Derm	65.21±2.91‖34	22.48±0.22‖1	65.55±1.30‖30	25.52±0.87‖3	22.48±0.35(0.6)‖2	30.76±1.55‖13	65.41±1.20‖26	66.62±1.49‖23	**67.52±1.25(1.0)‖20**
Auto	96.05±0.30‖7	62.98±1.85‖3	95.94±0.59‖6	51.35±0.21‖1	96.10±0.40(0.6)‖7	96.40±0.28‖6	**96.63±0.38‖5**	95.99±0.30‖7	94.59±0.70(0.7)‖4
Hous	81.48±1.10‖13	37.98±1.77‖2	80.83±0.97‖12	70.99±1.92‖6	60.28±1.18(0.7)‖5	75.36±1.07‖12	81.52±0.88‖11	80.02±1.04‖13	**81.92±1.01(0.0)‖12**
WDBC	56.49±1.39‖30	51.78±1.30‖1	54.82±2.32‖22	57.50±0.37‖1	56.22±2.55(0.6)‖25	56.63±2.11‖7	57.82±1.70‖20	57.15±1.76‖23	**64.34±1.16(0.5)‖8**
Aust	50.04±1.76‖14	45.77±1.87‖1	49.36±1.70‖13	49.81±1.29‖4	51.42±1.86(0.6)‖6	52.45±1.12‖1	51.62±1.57‖8	52.51±1.91‖3	**59.26±1.17(0.1)‖13**
Cred	66.22±1.59‖14	56.00±1.32‖1	66.96±1.67‖13	57.25±1.22‖2	65.62±1.64(0.6)‖6	65.72±1.35‖13	66.78±1.67‖14	66.80±1.97‖14	**67.25±1.47(1.0)‖11**
Germ	51.35±1.81‖20	51.76±0.71‖1	51.27±1.17‖20	52.74±0.78‖2	53.33±1.09(0.9)‖1	53.39±1.39‖3	55.70±0.68‖3	51.51±1.64‖19	**57.95±1.49(0.6)‖19**
Wred	89.98±0.55‖11	20.81±0.70‖1	89.66±0.58‖11	21.91±0.70‖1	89.55±0.54(0.6)‖9	89.28±0.50‖10	90.61±0.45‖9	90.46±0.44‖10	**90.83±0.46(0.4)‖8**
Wite	89.30±0.29‖11	17.78±0.19‖1	89.42±0.28‖11	37.85±0.36‖3	68.30±0.28(0.6)‖3	53.54±0.45‖10	89.52±0.29‖10	89.57±0.38‖9	**89.67±0.29(0.7)‖8**
Avg.	69.23±1.73‖23.33	43.66±1.16‖1.25	69.30±1.43‖14.92	48.81±1.05‖2.25	63.12±1.33‖6.50	64.50±1.43‖8.83	70.47±1.29‖15.50	70.08±1.29‖15.42	**72.98±1.25‖15.08**

selected features on classification performance. The relationship between the number of selected features and the classification accuracies of different algorithms with classifier SVM are shown in Fig. 6. Because algorithms DRSQR, HARCC, AR-FDRS, and AR-VCDRSA each output a feature subset, they do not need to learn the feature subset from the entire feature sequences. Therefore, the algorithms that output a feature subset do not need to perform this test, and naturally they do not appear in Fig. 6.

By observing Fig. 6, it is find that in addition to datasets Sonar, Derm, Hous, and Cred, our algorithm AAnFSCF quickly reaches the highest point of classification accuracy value on the remaining 8 datasets. Moreover, the classification accuracy values of the highest points of algorithm AAnFSCF are higher than that of other algorithms and the original data on most datasets. This characteristic is highlighted on datasets WPBC, Iono, Derm, Auto, WDBC, Aust, Germ, and Wred. This shows that the proposed algorithm AAnFSCF is effective, and it can accurately select the important features for ordinal classification.

C. Statistical Test

This section statistically tests the classification performance of eight algorithms on three different classifiers by using Friedman's test (F-test) and Nemenyi's posthoc test (N-test).

For F-test, the significance level is set as 0.1, and the null hypothesis that "all algorithms have the the same performance". The critical value is 1.796 from the F-test critical value table.

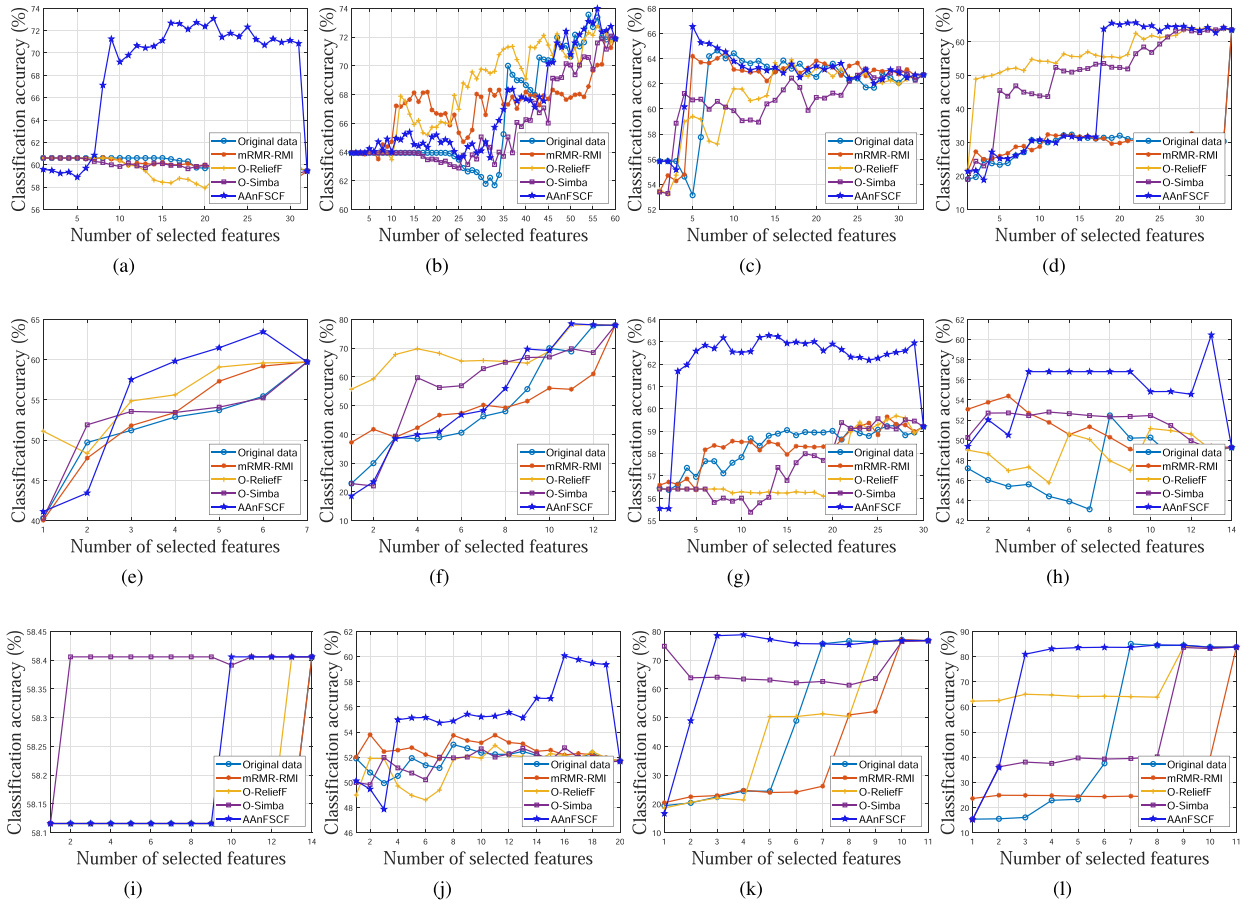


Fig. 6. Classification accuracy changes w.r.t. the number of selected features with classifier SVM. (a) WPBC. (b) Sonar. (c) Iono. (d) Derm. (e) Auto. (f) Hous. (g) WDBC. (h) Aust. (i) Cred. (j) Germ. (k) Wred. (l) Wite.

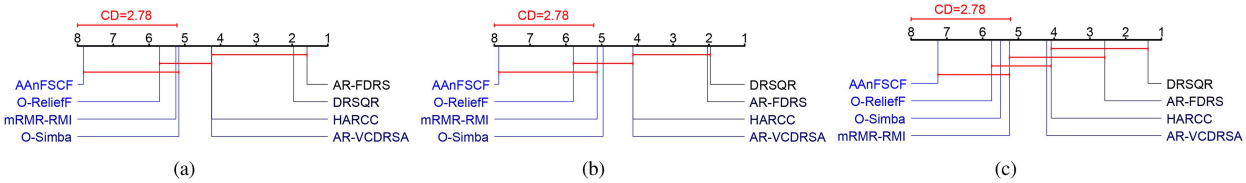


Fig. 7. Comparisons of the classification accuracy with the N-test on three classifiers. (a) NB. (b) SVM. (c) CART.

The values of τ_F on classifiers NB, SVM, and CART are 23.659, 18.700, and 15.228, respectively. Obviously, the values of τ_F on different classifiers are all larger than the critical value. Thus, the null hypothesis is rejected in the F-test.

Subsequently, the significant differences between any two algorithms are shown by using the N-test. The critical difference $CD = 2.780$ when the significance level is 0.1. The test results are shown in Fig. 7, the average ranking of each algorithm is marked on the number line. If the algorithms are connected by horizontal lines, there is no significant difference between the corresponding algorithms.

From Fig. 7, it is not difficult to find that our algorithm AAnFSCF significantly different from algorithms AR-FDRS, DRSQR, AR-VCDRSA, and HARCC on the three classifiers. Algorithms O-ReliefF and mRMR-RMI are not significantly different from the proposed algorithm AAnFSCF on the three classifiers, but algorithms AAnFSCF and O-Simba have a significant difference on classifier SVM. Furthermore, our algorithm AAnFSCF ranks first on different classifiers and far outperforms the second-ranked algorithm. Therefore, this statistical test also shows that the proposed algorithm AAnFSCF has good classification performance.

D. Parameter Sensitivity Analysis of Algorithm AAnFSCF

This section explores the sensitivity of the parameter β in algorithm AAnFSCF to classification accuracy tested by classifier NB. The experimental results are shown in Fig. 8.

By observing Fig. 8, it is easy to find that for most datasets, the first several features of the feature sequence output by the

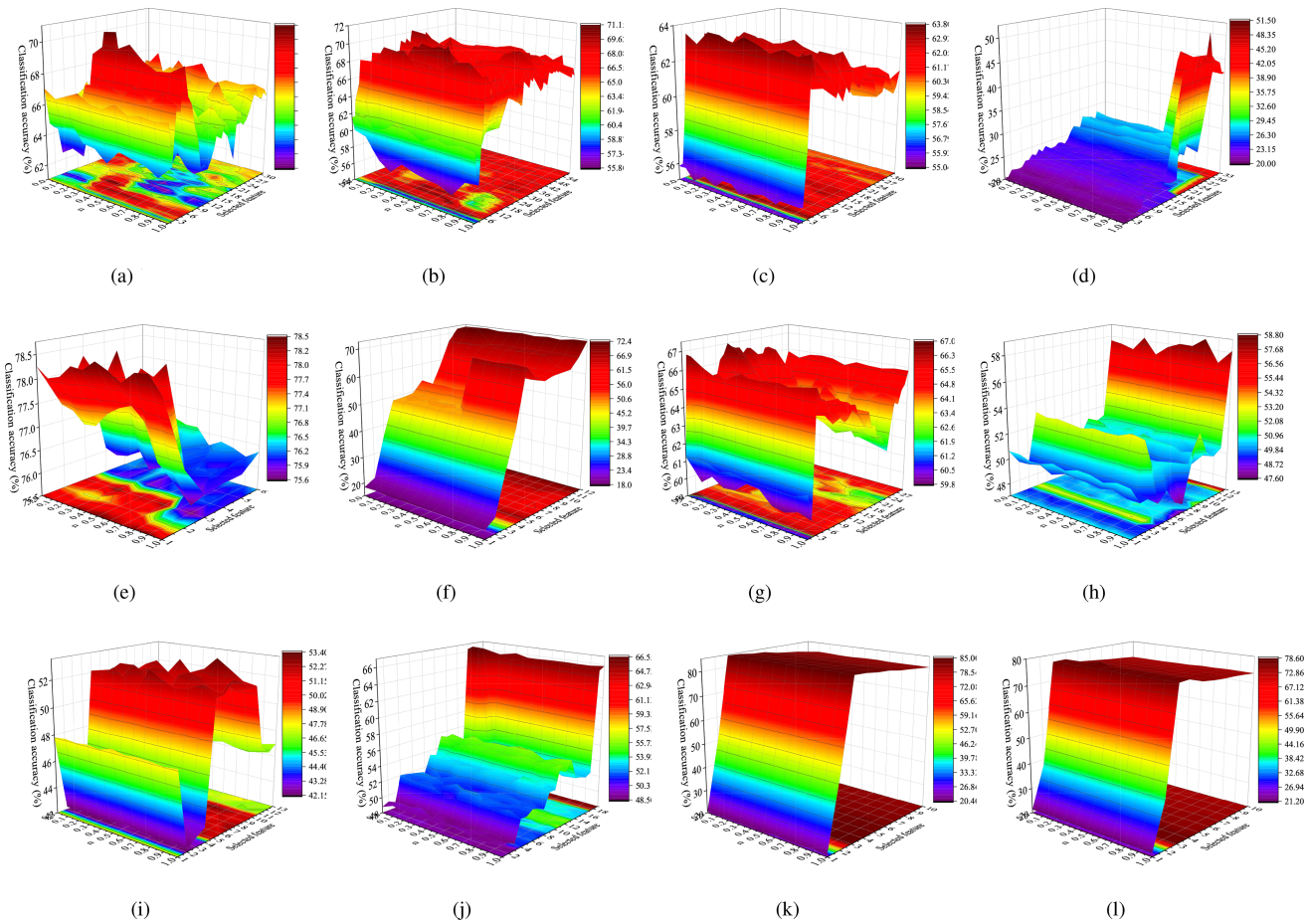


Fig. 8. Effect of parameter β and the selected features on the classification accuracy tested by classifier NB. (a) WPBC. (b) Sonar. (c) Iono. (d) Derm. (e) Auto. (f) Hous. (g) WDBC. (h) Aust. (i) Cred. (j) Germ. (k) Wred. (l) Wite.

algorithm AAnFSCF can reach the highest point of classification accuracy. This shows that features that provide important ordinal classification information can be extracted from the data by the algorithm AAnFSCF and placed at the front of the feature sequence. Moreover, for most datasets, the classification accuracy values corresponding to different parameters fluctuate significantly when the size of the feature subsets is the same. This shows that the feature sequences corresponding to different parameters are also different. This indicate that a better feature sequence for classification can be obtained by adjusting the parameter β.

E. Robustness Evaluations of Algorithm AAnFSCF

This section evaluates the robustness of feature selection algorithms in terms of feature sequences (subsets) and feature scores.

1) Evaluate the Robustness of Feature Sequences (subsets): Each data in Table I is added with 0%, 10%, 20%, 30%, 40%, and 50% noise samples respectively, so that each data generates six kinds of data with different noise levels. The same algorithm may obtain different feature sequences (subsets) on data containing different noise levels, and the similarity between them is used as a criterion to evaluate the robustness of the algorithm.

For the algorithm that outputs feature sequences, assuming that feature sequences $E_{l_i} = \{e_1^{l_i}, e_2^{l_i}, \ldots, e_m^{l_i}\}$ and $F_{l_j} = \{f_1^{l_j}, f_2^{l_j}, \ldots, f_m^{l_j}\}$ are respectively obtained at noise levels l_i and l_j, the similarity of these two feature sequences is calculated as follows:

$$S_{se}(E_{l_i}, F_{l_j}) = 1 - 6\sum_{k=1}^{m} \frac{(e_k^{l_i} - f_k^{l_j})^2}{m(m^2-1)}. \tag{25}$$

For the algorithm that outputs feature subsets, assuming that feature subsets $P_{l_i} = \{p_1^{l_i}, p_2^{l_i}, \ldots, p_{|P|}^{l_i}\}$ and $Q_{l_j} = \{q_1^{l_j}, q_2^{l_j}, \ldots, q_{|Q|}^{l_j}\}$ are obtained at noise levels l_i and l_j, respectively. The similarity of these two feature subsets is calculated as follows:

$$S_{su}(P_{l_i}, Q_{l_j}) = \frac{|P_{l_i} \cap Q_{l_j}|}{|P_{l_i}| + |Q_{l_j}| - |P_{l_i} \cap Q_{l_j}|}. \tag{26}$$

The same algorithm can obtain 6 feature sequences (subsets) under different noise levels, and the similarity of any two feature sequences (subsets) are calculated to form the similarity matrix

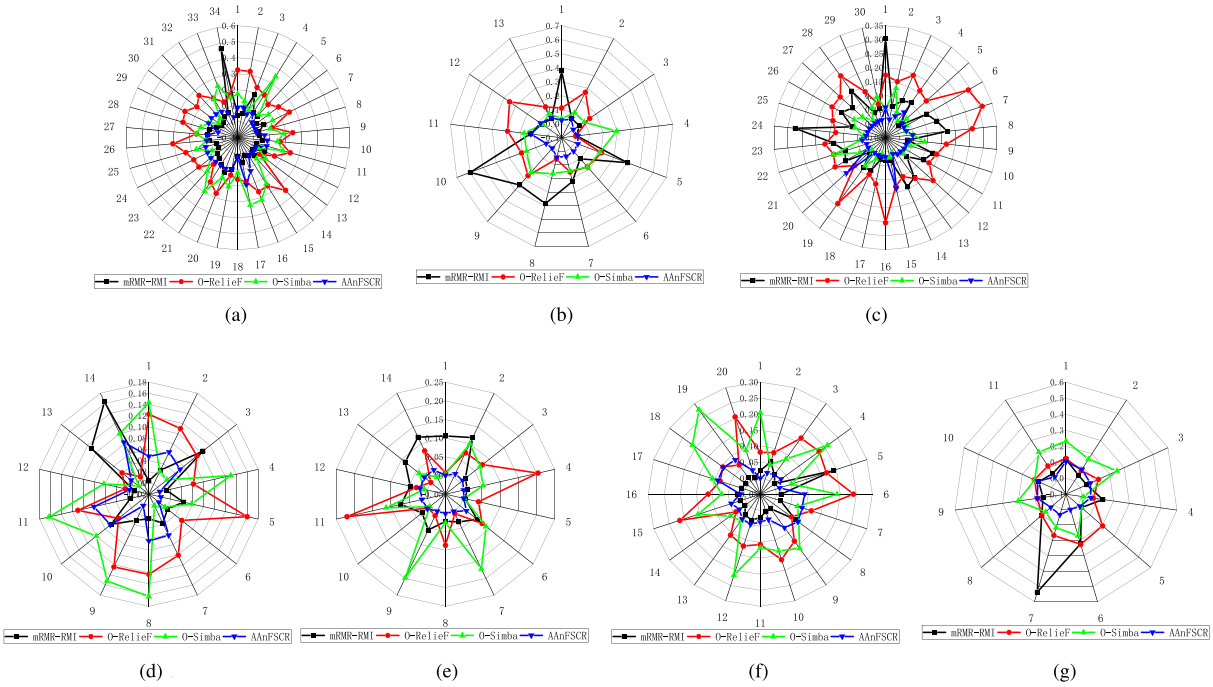


Fig. 9. Robustness evaluation of feature scores for different algorithms under different noise levels. (a) Derm. (b) Hous. (c) WDBC. (d) Aust. (e) Cred. (f) Germ. (g) Wred.

TABLE V
SIMILARITY OF FEATURE SEQUENCES (SUBSETS) UNDER DIFFERENT NOISE LEVELS

Datasets	DRSQR	HARCC	AR-FDRS	AR-VC DRSA	mRMR-RMI	O-ReliefF	O-Simba	AAnFSCF
WPBC	0.24	0.40	0.28	**0.70**	0.13	0.43	0.47	0.37
Sonar	0.30	0.30	0.39	**0.84**	0.20	0.34	0.55	0.46
Iono	0.57	0.51	**0.81**	0.32	0.24	0.25	0.23	0.60
Derm	0.35	0.65	0.52	0.76	0.01	0.67	0.41	**0.92**
Auto	0.42	**0.70**	0.49	0.15	0.28	0.36	0.53	0.47
Hous	0.40	0.61	0.42	0.19	0.36	0.35	0.48	**0.80**
WDBC	0.28	0.59	0.22	0.44	0.07	0.59	0.40	**0.95**
Aust	0.56	0.62	0.65	0.57	0.11	0.74	0.48	**0.92**
Cred	0.33	0.88	0.36	0.14	0.45	0.65	0.62	**0.92**
Germ	0.76	0.75	0.63	0.52	0.31	0.52	0.51	**0.87**
Wred	**0.80**	0.73	0.37	0.11	0.26	0.47	0.46	**0.80**
Wite	**0.77**	0.73	0.70	0.26	0.36	0.47	0.67	0.73
Avg.	0.48	0.62	0.49	0.42	0.23	0.49	0.48	**0.73**

$\mathbb{S} = [S_{ij}]_{6\times 6}$, where S_{ij} refers to the similarity between the generated feature sequences (subsets) at noise levels i and j. The average value of the similarity matrix $\mathbb{S}$ is the final evaluation index, and the larger it is, the more robust the algorithm is. The experimental results are shown in Table V, where the number in bold face indicates the highest similarity corresponding to the current row.

From Table V, we find that our algorithm AAnFSCF achieves relatively high similarity values, which is also proved from the perspective of the average. This proves that the feature sequence generated by our algorithm AAnFSCF has good robustness in noisy environment.

2) Evaluate the Robustness of Feature Scores: For algorithms that generate feature sequences, we compare their robustness of feature scores (also known as feature importance values) under different noise levels. Specifically, for a certain algorithm, we get the score of each feature under different noise levels, and calculate the standard deviation of these feature scores as the evaluation index. The smaller the evaluation index, the better. For example, the scores obtained for feature c under different noise levels can form a vector $c = (v_c^{l_{0\%}}, v_c^{l_{10\%}}, \ldots, v_c^{l_{50\%}})$, and then the standard deviation $std(c)$ of the values in the vector c is calculated as the experimental result. The experimental results of seven datasets selected are shown in Fig. 9 in the form of radar graph, where a point in the subgraph corresponds to a $std(c)$ value. Obviously, an algorithm is more robust if it covers a smaller area. By observing Fig. 9, we can roughly find that the coverage area corresponding to the AAnFSCF algorithm is the smallest, where the subgraphs Fig. 9(b), (c), (e), and (g) are the most significant. This manifests that our algorithm AAnFSCF has better robustness in noisy environments.

VI. CONCLUSION

This article develops a novel feature selection method for ordinal classification from the perspectives of robustness, CS, and feature-redundancy. First, a robust FDRS model is proposed, which calculates the density of samples by using an AKNN strategy and designs an approximate calculation mechanism to actively antinoise. Then, the CS and feature-redundancy are defined in this robust fuzzy dominance rough approximation space, where the CS is characterized by approximation sets and the feature-redundancy describes the redundancy between features by FRCMI. Finally, a robust feature evaluation index that comprehensively considers CS and feature-redundancy is proposed, and the corresponding feature selection algorithm AAnFSCF is also designed. Extensive experiments are performed in noisy

environments, and the results demonstrate that our algorithm AAnFSCF not only has better classification performance but also good robustness. However, the robust FDRS model and feature selection method proposed in this article are only suitable for data mining from single granularity knowledge space. For more complex or multimodal ordinal classification tasks, it is necessary to develop a robust FDRS model and feature selection method from a multigranularity perspective, thereby improving their generalization ability in practical applications. Therefore, in future work, we will explore robust FDRS model from a multigranularity perspective and investigate multigranularity uncertainty measures for feature selection of ordinal classification.

REFERENCES

[1] Q. H. Hu et al., "Feature selection for monotonic classification," *IEEE Trans. Fuzzy Syst.*, vol. 20, no. 1, pp. 69–81, Feb. 2012.

[2] J. T. Wang, Y. H. Qian, F. J. Li, J. Y. Liang, and W. P. Ding, "Fusing fuzzy monotonic decision trees," *IEEE Trans. Fuzzy Syst.*, vol. 28, no. 5, pp. 887–900, May 2020.

[3] J.-R. Cano, P. A. Gutirrez, B. Krawczyk, M. Wozniak, and S. Garca, "Monotonic classification: An overview on algorithms, performance measures and data sets," *Neurocomputing*, vol. 341, pp. 168–182, 2019.

[4] M.-J. Kim and I. Han, "The discovery of experts' decision rules from qualitative bankruptcy data using genetic algorithms," *Expert Syst. Appl.*, vol. 25, no. 4, pp. 637–646, 2003.

[5] J.-R. Cano, N. R. Aljohani, R. A. Abbasi, J. S. Alowidbi, and S. García, "Prototype selection to improve monotonic nearest neighbor," *Eng. Appl. Artif. Intell.*, vol. 60, pp. 128–135, 2017.

[6] Z. Pawlak, "Rough sets," *Int. J. Comput. Inf. Sci.*, vol. 11, no. 5, pp. 341–356, 1982.

[7] D. Dubois and H. Parde, "Rough fuzzy sets and fuzzy rough sets," *Int. J. Gen. Syst.*, vol. 17, no. 2-3, pp. 191–209, 1990.

[8] Z. Yuan, H. M. Chen, T. R. Li, B. B. Sang, and S. Wang, "Outlier detection based on fuzzy rough granules in mixed attribute data," *IEEE Trans. Cybern.*, vol. 52, no. 8, pp. 8399–8412, Aug. 2022.

[9] Z. Yuan, H. M. Chen, T. R. Li, J. Liu, and S. Wang, "Fuzzy information entropy-based adaptive approach for hybrid feature outlier detection," *Fuzzy Sets Syst.*, vol. 421, pp. 1–28, 2021.

[10] P. Maji and P. Garai, "IT2 fuzzy-rough sets and max relevance-max significance criterion for attribute selection," *IEEE Trans. Cybern.*, vol. 45, no. 8, pp. 1657–1668, Aug. 2015.

[11] P. Maji and P. Garai, "Fuzzy–rough simultaneous attribute selection and feature extraction algorithm," *IEEE Trans. Cybern.*, vol. 43, no. 4, pp. 1166–1177, Aug. 2013.

[12] M. Hu, E. C. Tsang, Y. T. Guo, and W. H. Xu, "Fast and robust attribute reduction based on the separability in fuzzy decision systems," *IEEE Trans. Cybern.*, vol. 52, no. 6, pp. 5559–5572, Jun. 2022.

[13] S. An, Q. H. Hu, W. Pedrycz, P. F. Zhu, and E. C. Tsang, "Data-distribution-aware fuzzy rough set model and its application to robust classification," *IEEE Trans. Cybern.*, vol. 46, no. 12, pp. 3073–3085, Dec. 2016.

[14] S. An, E. H. Zhao, C. Z. Wang, G. Guo, S. Y. Zhao, and P. Y. Li, "Relative fuzzy rough approximations for feature selection and classification," *IEEE Trans. Cybern.*, vol. 53, no. 4, pp. 2200–2210, Apr. 2023.

[15] P. Maji and P. Garai, "Rough hypercuboid based generalized and robust IT2 fuzzy c-means algorithm," *IEEE Trans. Cybern.*, vol. 51, no. 7, pp. 3641–3652, Jul. 2021.

[16] L. Concepción, G. Nápoles, I. Grau, and W. Pedrycz, "Fuzzy-rough cognitive networks: Theoretical analysis and simpler models," *IEEE Trans. Cybern.*, vol. 52, no. 5, pp. 2994–3005, May 2022.

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

[18] Q. H. Hu, L. Zhang, A. Shuang, D. Zhang, and D. R. Yu, "On robust fuzzy rough set models," *IEEE Trans. Fuzzy Syst.*, vol. 20, no. 4, pp. 636–651, Aug. 2012.

[19] X. D. Wu and X. Q. Zhu, "Mining with noise knowledge: Error-aware data mining," *IEEE Trans. Syst. Man, Cybern. Part A-Syst. Hum.*, vol. 38, pp. 917–932, Jul. 2008.

[20] Y. Q. Yao, J. S. Mi, and Z. J. Li, "A novel variable precision (θ, σ)-fuzzy rough set model based on fuzzy granules," *Fuzzy Sets Syst.*, vol. 236, pp. 58–72, 2014.

[21] Y. W. Li, S. X. Wu, Y. J. Lin, and J. H. Liu, "Different classes ratio fuzzy rough set based robust feature selection," *Knowl.-Based Syst.*, vol. 120, pp. 74–86, 2017.

[22] Y. W. Li, S. S. Wei, X. Liu, and Z. M. Zhang, "A novel robust fuzzy rough set model for feature selection," *Complexity*, vol. 2021, 2021, Art. no. 6685396. doi: 10.1155/2021/6685396.

[23] S. An, Q. H. Hu, and C. Z. Wang, "Probability granular distance-based fuzzy rough set model," *Appl. Soft Comput.*, vol. 102, 2021, Art. no. 107064. doi: 10.1016/j.asoc.2020.107064.

[24] J. C. R. Alcantud, F. Feng, and R. R. Yager, "An N-soft set approach to rough sets," *IEEE Trans. Fuzzy Syst.*, vol. 28, no. 11, pp. 2996–3007, Nov. 2020.

[25] S. Greco, B. Matarazzo, and R. Slowinski, "Rough approximation of a preference relation by dominance relations," *Eur. J. Oper. Res.*, vol. 117, no. 1, pp. 63–83, 1999.

[26] W. Kotlowski, K. Dembczyński, S. Greco, and R. Slowiński, "Stochastic dominance-based rough set model for ordinal classification," *Inf. Sci.*, vol. 178, no. 21, pp. 4019–4037, 2008.

[27] M. Szelag, S. Greco, and R. Slowiński, "Variable consistency dominance-based rough set approach to preference learning in multicriteria ranking," *Inf. Sci.*, vol. 277, pp. 525–552, 2014.

[28] W. Pan, K. She, and P. Y. Wei, "Multi-granulation fuzzy preference relation rough set for ordinal decision system," *Fuzzy Sets Syst.*, vol. 312, pp. 87–108, 2017.

[29] T. Shaheen, B. Mian, M. Shabir, and F. Feng, "A novel approach to decision analysis using dominance-based soft rough sets," *Int. J. Fuzzy Syst.*, vol. 21, pp. 954–962, 2019.

[30] B. Huang, H. X. Li, G. F. Feng, and X. Z. Zhou, "Dominance-based rough sets in multi-scale intuitionistic fuzzy decision tables," *Appl. Math. Comput.*, vol. 348, pp. 487–512, 2019.

[31] A. Ali, M. I. Ali, and N. Rehman, "Soft dominance based rough sets with applications in information systems," *Int. J. Approx. Reasoning*, vol. 113, pp. 171–195, 2019.

[32] Q. Huang, T. Li, Y. Huang, X. Yang, and H. Fujita, "Dynamic dominance rough set approach for processing composite ordered data," *Knowl.-Based Syst.*, vol. 187, 2020, Art. no. 104829. doi: 10.1016/j.knosys.2019.06.037.

[33] Q. H. Hu, D. R. Yu, and M. Z. Guo, "Fuzzy preference based rough sets," *Inf. Sci.*, vol. 180, no. 10, pp. 2003–2022, 2010.

[34] W. S. Du and B. Q. Hu, "Dominance-based rough fuzzy set approach and its application to rule induction," *Eur. J. Oper. Res.*, vol. 261, no. 2, pp. 690–703, 2017.

[35] B. B. Sang, H. M. Chen, L. Yang, T. R. Li, and W. H. Xu, "Incremental feature selection using a conditional entropy based on fuzzy dominance neighborhood rough sets," *IEEE Trans. Fuzzy Syst.*, vol. 30, no. 6, pp. 1683–1697, Jun. 2022.

[36] S. Y. Yang, H. Y. Zhang, B. D. Baets, M. Jah, and G. Shi, "Quantitative dominance-based neighborhood rough sets via fuzzy preference relations," *IEEE Trans. Fuzzy Syst.*, vol. 29, no. 3, pp. 515–529, Mar. 2021.

[37] Q. H. Hu, D. R. Yu, W. Pedrycz, and D. G. Chen, "Kernelized fuzzy rough sets and their applications," *IEEE Trans. Knowl. Data Eng.*, vol. 23, no. 11, pp. 1649–1667, Nov. 2011.

[38] C. Z. Wang, Y. Huang, M. W. Shao, Q. H. Hu, and D. G. Chen, "Feature selection based on neighborhood self-information," *IEEE Trans. Cybern.*, vol. 50, no. 9, pp. 4031–4042, Sep. 2020.

[39] W. P. Ding, C. T. Lin, and Z. H. Cao, "Shared nearest-neighbor quantum game-based attribute reduction with hierarchical coevolutionary spark and its application in consistent segmentation of neonatal cerebral cortical surfaces," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 30, no. 7, pp. 2013–2027, Jul. 2019.

[40] Y. Y. Yang, D. G. Chen, H. Wang, and X. Z. Wang, "Incremental perspective for feature selection based on fuzzy rough sets," *IEEE Trans. Fuzzy Syst.*, vol. 26, no. 3, pp. 1257–1273, Jun. 2018.

[41] Q. H. Hu, W. W. Pan, Y. P. Song, and D. R. Yu, "Large-margin feature selection for monotonic classification," *Knowl.-Based Syst.*, vol. 31, pp. 8–18, 2012.

[42] W. W. Pan, Q. H. Hu, Y. P. Song, and D. R. Yu, "Feature selection for monotonic classification via maximizing monotonic dependency," *Int. J. Comput. Intell. Syst.*, vol. 7, no. 3, pp. 543–555, 2014.

[43] Y. H. Qian, H. Xu, J. Y. Liang, B. Liu, and J. T. Wang, "Fusing monotonic decision trees," *IEEE Trans. Knowl. Data Eng.*, vol. 27, no. 10, pp. 2717–2728, Oct. 2015.

[44] W. S. Du and B. Q. Hu, "Attribute reduction in ordered decision tables via evidence theory," *Inf. Sci.*, vol. 364, pp. 91–110, 2016.

[45] C. Luo, H. Pi, T. R. Li, H. M. Chen, and Y. Y. Huang, "Novel fuzzy rank discrimination measures for monotonic ordinal feature selection," *Knowl.-Based Syst.*, vol. 240, 2022, Art. no. 108178. doi: 10.1016/j.knosys.2022.108178.

[46] Q. H. Hu, S. An, and D. R. Yu, "Soft fuzzy rough sets for robust feature evaluation and selection," *Inf. Sci.*, vol. 180, no. 22, pp. 4384–4400, 2010.

[47] X. F. Zhu, S. C. Zhang, R. Y. Hu, Y. H. Zhu, and j. K. Song, "Local and global structure preservation for robust unsupervised spectral feature selection," *IEEE Trans. Knowl. Data Eng.*, vol. 30, no. 3, pp. 517–529, Mar. 2018.

[48] Y. P. Qu, Z. Xu, C. J. Shang, X. L. Ge, A. S. Deng, and Q. Shen, "Inconsistency guided robust attribute reduction," *Inf. Sci.*, vol. 580, pp. 69–91, 2021.

[49] L. J. Dong, D. G. Chen, N. L. Wang, and Z. H. Lu, "Key energy-consumption feature selection of thermal power systems based on robust attribute reduction with rough sets," *Inf. Sci.*, vol. 532, pp. 61–71, 2020.

[50] L. Y. Fang et al., "Feature selection method based on mutual information and class separability for dimension reduction in multidimensional time series for clinical data," *Biomed. Signal Process. Control*, vol. 21, pp. 82–89, 2015.

[51] P. Zhou, X. G. Hu, P. P. Li, and X. D. Wu, "Online feature selection for high-dimensional class-imbalanced data," *Knowl.-Based Syst.*, vol. 136, pp. 187–199, 2017.

[52] J. K. Chen, Y. J. Lin, J. S. Mi, S. Z. Li, and W. P. Ding, "A spectral feature selection approach with kernelized fuzzy rough sets," *IEEE Trans. Fuzzy Syst.*, vol. 30, no. 8, pp. 2886–2901, Aug. 2022.

[53] S. Singha and P. P. Shenoy, "An adaptive heuristic for feature selection based on complementarity," *Mach. Learn.*, vol. 107, no. 12, pp. 2027–2071, 2018.

[54] W. B. Qian, X. D. Long, Y. L. Wang, and Y. H. Xie, "Multi-label feature selection based on label distribution and feature complementarity," *Appl. Soft Comput.*, vol. 90, 2020, Art. no. 106167.

[55] J. H. Wan, H. M. Chen, T. R. Li, Z. Yuan, J. Liu, and W. Huang, "Interactive and complementary feature selection via fuzzy multigranularity uncertainty measures," *IEEE Trans. Cybern.*, vol. 53, no. 2, pp. 1208–1221, Feb. 2023.

[56] Z. J. Chen et al., "Clustering-based feature subset selection with analysis on the redundancy ccomplementarity dimension," *Comput. Commun.*, vol. 168, pp. 65–74, 2021.

[57] S. Greco, B. Matarazzo, and R. Slowinski, "Rough sets theory for multicriteria decision analysis," *Eur. J. Oper. Res.*, vol. 129, no. 1, pp. 1–47, 2001.

[58] W. S. Du and B. Q. Hu, "A fast heuristic attribute reduction approach to ordered decision systems," *Eur. J. Oper. Res.*, vol. 264, no. 2, pp. 440–452, 2018.

[59] S. Greco, B. Matarazzo, S. Roman, and J. Stefanowski, "Variable consistency model of dominance-based rough sets approach," in *Proc. 2nd Int. Conf. Rough Sets Curr. Trends Comput.*, 2000, vol. 2005, pp. 170–181.