A Robust Pseudo Fuzzy Rough Feature Selection Using Linear Reconstruction Measure

Lin Qiu, Xingwei Wang, Yanpeng Qu, Kaimin Zhang, Fei Gao, Bo Yi, *Member, IEEE*, and Keqin Li, *Fellow, IEEE*

Abstract—**Fuzzy-rough sets (FRS) provide an outstanding theoretical tool for feature selection (FS). Whilst promising, the FRS model is sensitive to noisy information and ineffectively applicable to the data with large class density difference, with existing FRS-based FS methods only tackling one of these challenges. Therefore, to overcome both of these issues, this article presents a robust FS algorithm using linear reconstruction measure for the first time. First, a pseudo FRS model is proposed, where the distribution-aware linear reconstruction relation serving as the fuzzy similarity relation is constructed by considering the insight of meaningful information (i.e., distribution information of samples and density information of classes) to enhance the robustness and the pseudofuzzy rough approximations are further redefined based on k-Nearest Neighbor (kNN) granules determined by the linear reconstruction coefficients to empower the antinoise ability. Then, the pseudo FRS model is employed to guide the robust FS algorithm from the perspective of *redundant filter*, *strongly relevant priority*, and *discriminative selection* to determine the final feature subset. The experimental results on 31 datasets and practical applications (i.e., cancer diagnosis and face recognition) demonstrate that the reduct gained by the proposed approach generally outperforms those attained by alternative implementations of FRS-based FS and state-of-the-art FS techniques.**

Index Terms—**Antinoise ability, linear reconstruction measure (LRM), k-Nearest Neighbor (kNN), pseudofuzzy-rough sets (FRS), robust feature selection (FS).**

I. Introduction

FEATURE selection (FS) is presently one of the most prominent data preprocessing approaches. By eliminating irrelevant and redundant features, FS contributes to achieve a subset of the original features preserving their original meanings unaltered while improving performance of a specific task with regard to efficiency and data comprehension [1], [2]. The implementation of FS may follow three strategies: embedded, wrapper, and filter. The key factor of the filter strategy-based ones is to design an evaluation function in terms of specific criteria to evaluate the feature subset and select the features that satisfy the defined conditions. Fuzzy-rough sets (FRS) [3], [4] capture uncertainty inherent in data, information, or knowledge by integrating the concepts of vagueness (for fuzzy sets [5]) and indiscernibility (for rough sets [6]), which are utilized to guide the FS process with significant success.

In essence, FRS-based FS (FRS-FS) approaches are implemented by making the most of the principles associated with the dependency degrees, such as, a series of heuristic algorithms developed by Jensen, Shen, et al. [7], [8], [9], [10], [11], Hu, et al. [12], [13], [14], Yang et al. [15], and Wang et al. [16]. However, most existing FRS-FS algorithms neglect meaningful information embedded in the possible structure of inherent grouped features, which may lead to redundant information. Therefore, FRS-FS algorithms have undergone further refinement through integration with the consideration of feature grouping [17], [18], [19], [20]. For instance, correlation coefficient in FRFG [17], graph theory in GBFG [18] and FGS-RFRAS [19], and k-means in EL-TSFRFS [20] are employed to group redundant features, respectively, and distinct FS techniques are implemented subsequently to determine the discriminative feature subset by using FRS.

While promising, the aforementioned methods are sensitive to noise information and data distribution. In order to overcome these drawbacks, many researchers are dedicating their efforts to the exploration of robust FRS-FS algorithms. To improve the antinoise ability, some robust FRS models have been proposed by improving the calculation of fuzzy approximations in classical FRS, such as, β-precision FRS model [21], VQRS model [22], VPFRS model [23], FVPRS model [24], PFRS model [25], OWA-FRS model [26], SFRS model [27], k-means FRS model [28], k-trimmed FRS model [28], k-median FRS model [28], k-order FRS model [29], and RFRS model [28]. These robust models focus on mitigating the impact of the noisy information in data on the lower approximation, by neglecting some nearest samples to expand the lower approximation (i.e., β-precision FRS, PFRS, SFRS, and k-trimmed FRS) or substituting the minimum statistics with robust approximation operators (i.e., VQRS, VPFRS, FVPRS, OWA-FRS, k-means FRS, and k-median FRS). Moreover, some approaches are proposed to overcome the sensitiveness of noisy data by conducting with different fuzzy similarity relations, such as,

Manuscript received 6 February 2024; revised 21 May 2024; accepted 3 July 2024. Date of publication 9 July 2024; date of current version 8 October 2024. This work was supported in part by the National Key Research and Development Program of China under Grant 2022YFB4500800, and in part by the National Natural Science Foundation of China under Grant 62032013 and Grant 92267206. Recommended by Associate Editor R. Jensen. *(Corresponding author: Xingwei Wang.)*

Lin Qiu, Xingwei Wang, Kaimin Zhang, Fei Gao, and Bo Yi are with the College of Computer Science and Engineering, Northeastern University, Shenyang 110819, China (e-mail: qiulin960229@163.com; wangxw@mail.neu.edu.cn; 2201918@stu.neu.edu.cn; gaof9264@163.com; yibo@mail.neu.edu.cn).

Yanpeng Qu is with the College of Artificial Intelligence, Dalian Maritime University, Dalian 116026, China (e-mail: yanpengqu@dlmu.edu.cn).

Keqin Li is with the Department of Computer Science, State University of New York, New Paltz, NY 12561 USA (e-mail: lik@newpaltz.edu).

Digital Object Identifier 10.1109/TFUZZ.2024.3424809

kernel function-based metrics [30], [31], [32], and information theory-based metrics [33], [34], [35], [36].

Another kind of robust FRS-FS technique aims at addressing the sensitiveness of data distribution [37], [38], [39], [40], [41]. In [37], the relative distance-based FS is developed by integrating the absolute distance with the class density to relieve the impact of data distribution on the FRS model. Also, in [38], the k-Nearest Neighbor (kNN) granules information is used for designing the relative distance-based uncertainty measure through which, a robust FRS model proves to be highly effective for datasets exhibiting significant density variations. In addition, to improve the generalization of FRS model, the relative similarity is developed by considering the label distribution among data samples in the proposed model [39]. In [40], a fitting FRS model is developed by analyzing the data similarity distributions with respect to the class decisions. In [41], a local composite entropy to the uncertainty and decision distribution in fuzzy decision systems is developed to comprehensively account for the distribution characteristics of imbalanced data.

While showing potential, the abovementioned robust FRS-FS works concentrate exclusively on addressing either the weakness of antinoise ability or the sensitiveness of data distribution, yet both of these issues significantly impact the generalization of FRS models, subsequently degrading the overall performance of classification tasks. To combat the potential adverse impacts of noisy information and data distribution concurrently, a novel *L*inear *R*econstruction *M*easure-based *R*obust *P*seudo *F*uzzy *R*ough *F*eature *S*election (LRM-RPFRFS) supported with a robust pseudo FRS model is presented in this article, enhancing the performance of FS.

Particularly, the distribution-aware linear reconstruction relation serving as the fuzzy similarity relation is developed using linear reconstruction measure (LRM) [42] advocating for the evaluation of a feature subset and the pseudofuzzy rough approximations are redefined by determining the kNN granules in terms of the linear reconstruction coefficients. Furthermore, the robust LRM-k-PFRS model is constructed to improve the antinoise ability and overcome the sensitiveness of data distribution in this article. In the proposed approach, we integrated the LRM-k-PFRS model into the *RF-SRP-DS* selection strategy by fully considering three aspects: *redundant filter (RF)*, *strongly relevant priority (SRP),* and *discriminative selection (DS)* to determine the final feature subset.

The proposed LRM-RPFRFS approach is fully investigated through systematic experimental validation and evaluation. The experimental studies are implemented in reference to six FRS-FS methods (i.e., four feature grouping-based FRS-FS methods and two heuristic FRS-FS methods) and four state-of-the-art FS methods (i.e., two wrapped strategy-based FS methods and two filter strategy-based FS methods), including: FRFG [17], GBFG [18], EL-TSFRFS [20], FGS-RFRAS [19], FRMR [15], HARCM [16], GSA [43], GWO [44], ReliefF [45], PCC [46]. Furthermore, the determined reducts are evaluated by the following four different classifiers: J48 [47], Bagging [48], Jrip [49], and Part [50], respectively. The comparative results demonstrate that LRM-RPFRFS outperforms the rest, ensuring the effectiveness, generalization, and robustness, across 31 datasets, including: 18 benchmark datasets, 5 noisy datasets, 3 synthetic datasets with different data distributions, 3 biological datasets and 2 face datasets.

The contributions of this article are outlined from following two perspectives.

1) A robust pseudo FRS (i.e., LRM-k-PFRS) model is presented by utilizing LRM for the first time.
   a) Compared to existing fuzzy similarity relations (which mainly consider the similarity between two samples), the distribution-aware linear reconstruction relation serving as the fuzzy similarity relation in LRM-k-PFRS fully considers the sight of certain meaningful information, including distribution information of samples and density information of classes. Therefore, the LRM-k-PFRS model is effectively applicable for diverse data distributions.
   b) The definitions of pseudofuzzy rough approximations in LRM-k-PFRS do not only use the nearest neighbor from other decision classes, but also calculates the average value of dissimilarities between the sample and kNN granules. Compared to existing kNN granules-based FRS models (which employ the point-to-point measures), the linear reconstruction coefficients used to determine the kNN granules consider not only the distance information between two samples, but also the relationship between all samples. Thus, the kNN granules in LRM-k-PFRS can capture more meaningful information and furthermore, the LRM-k-PFRS model can effectively alleviate the sensitiveness of noisy samples.
2) The strategy of *RF-SRP-DS* is designed to guide the supervised filter-based FS algorithm (i.e., LRM-RPFRFS).
   a) Compared to the most existing heuristic FRS-FS approaches (which only focus on discrimination) and feature grouping-based FRS-FS approaches (which ignore the relevance), *RF-SRP-DS* initially filters out the redundant features, then ranks the rest based on the relevance between conditional features and decision feature (with highly correlated features prioritized). Subsequently, LRM-k-PFRS model is employed to determine a more discriminative feature subset. Therefore, *RF-SRP-DS* comprehensively considers three essential aspects: redundancy, relevance, and discrimination.
   b) Compared to existing feature grouping-based FRS-FS approaches (which use correlation coefficient and threshold setting, information theory & graph theory and k-means), *RF-SRP-DS* employs LRM and graph theory to form feature groups, where LRM thoroughly considers the distributional relationships between features to evaluate their redundancy, thereby grouping features that offer parallel discriminative capabilities.

The rest of this article is structured as follows. In Section II, the problems of classical FRS model are described. The pseudorobust model called LRM-k-PFRS and the novel FS approach called LRM-RPFRFS are presented in Sections III and IV, respectively. In Section V, the results of comparative experimental studies are presented and discussed. Finally, Section VI concludes this article.

II. Classical FRS Model and Its Problems

Let $DT = (\mathbb{U}, \mathbb{C} \cup \mathbb{D})$ denote a decision table, where $\mathbb{U} = \{x_t | t = 1, \ldots, n\}$ is a nonempty set of finite samples; $\mathbb{C} = \{f_i | i = 1, \ldots, m\}$ is a nonempty set of condition features; and $\mathbb{D} = \{y_t \in \{l_1, \ldots, l_r\} | t = 1, \ldots, n\}$ is the set of decision feature. Suppose that $\mathbb{D}$ partitions the samples in $\mathbb{U}$ into r crisp equivalence classes $\mathbb{U}/\mathbb{D} = \{\mathbb{D}_1, \mathbb{D}_2, \ldots, \mathbb{D}_r\}$, where $\forall x_k \in \mathbb{D}_j, y_k \in l_j$.

The definitions of the fuzzy lower and upper approximations of a decision class $\mathbb{D}_k$, $k = 1, \ldots, r$ with respect to $P \subseteq \mathbb{C}$ are defined as follows:

$$\underline{R_P}(\mathbb{D}_k)(x_i) = \min_{x_j \notin \mathbb{D}_k} \{1 - R_P(x_i, x_j)\} \tag{1}$$

$$\overline{R_P}(\mathbb{D}_k)(x_i) = \max_{x_j \in \mathbb{D}_k} R_P(x_i, x_j) \tag{2}$$

with $x_i \in \mathbb{U}$ and R_P the fuzzy similarity relation induced by the subset of features P

$$R_P(x_i, x_j) = \cap_{a \in P} R_a(x_i, x_j) \tag{3}$$

where $R_a(x_i, x_j)$ denotes the degree to which the objects x_i and x_j are regarded to be similar with respect to the feature a; and it is a fuzzy similarity relation as it satisfies

1) Reflexivity: $R_a(x_i, x_i) = 1, \forall x_i \in \mathbb{U}$.
2) Symmetry: $R_a(x_i, x_j) = R_a(x_j, x_i), \forall x_i, x_j \in \mathbb{U}$.

$\underline{R_P}(\mathbb{D}_k)(x_i)$ (i.e., the minimum one among the dissimilarities between x_i and all the samples from the domain $\mathbb{U} - \mathbb{D}_k$) and $\overline{R_P}(\mathbb{D}_k)(x_i)$ (i.e., the maximum one among the similarities between x_i and all the samples from $\mathbb{D}_k$) indicate that the membership degree of x_i certainly and possibly belonging to the class $\mathbb{D}_k$, respectively.

Furthermore, the fuzzy positive region of $\mathbb{D}$ with respect to $P \subseteq \mathbb{C}$ is defined as follows:

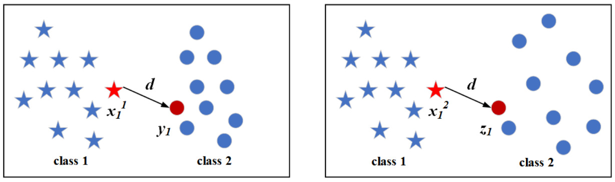
$$\mathrm{POS}_P(\mathbb{D})(x_i) = \bigcup_{k=1}^{r} \underline{R_P}(\mathbb{D}_k)(x_i). \tag{4}$$

Based on the definition of fuzzy positive region, the fuzzy dependency degree is defined as follows:

$$\gamma_P(\mathbb{D}) = \frac{\sum_{i=1}^{n} \mathrm{POS}_P(\mathbb{D})(x_i)}{|\mathbb{U}|}. \tag{5}$$

While promising, the definitions of (1) and (2) in the classical FRS [7] are sensitive to noise in data and do not reflect the uncertainty information of data precisely when the difference of class density is significantly obvious. The mentioned problems are illustrated as follows (taking the lower approximation as the example).

Sensitiveness to Noise Information: As shown in (1), the lower approximation in classical FRS is defined based on the dissimilarity between the sample and the nearest neighbor from other decision classes. However, in practical applications, data is often unavoidably accompanied by noise for a variety of reasons. Thus, if the nearest neighbor happens to be the noise sample, this will not only reduce the reliability and accuracy of the lower approximation, but also degrade the performance of the downstream classification task.



(a) (b)

Fig. 1. Two kinds of data distributions. (a) Data with distribution 1. (b) Data with distribution 2.

Sensitiveness to Data Distribution: To clearly illustrate the problem, a simple example is used as shown in Fig. 1. In Fig. 1, the data distributions D_1 and D_2 are generated based on the feature subsets P_1 and P_2, respectively, where the distribution for class 1 (c_1) are same and the distribution for class 2 (c_2) are different. The density difference between class 1 and class 2 is small for distribution D_1, while that of distribution D_2 is large. The lower approximations of x_1^1 and x_1^2 can be calculated according to (1)

$$\underline{R_{P_1}} c_1(x_1^1) = 1 - R_{P_1}(x_1^1, y_1)$$

$$\underline{R_{P_2}} c_1(x_1^2) = 1 - R_{P_2}(x_1^2, z_1).$$

Suppose that the following fuzzy similarity relation is employed:

$$R(x, y) = \exp\left(-\frac{||x - y||^2}{\sigma}\right).$$

or

$$R(x, y) = 1 - ||x - y||^2.$$

It can be seen that $R_{P_1}(x_1^1, y_1) = R_{P_2}(x_1^2, z_1)$, thus $\underline{R_{P_1}} c_1(x_1^1) = \underline{R_{P_2}} c_1(x_1^2)$. That is, x_1^1 and x_1^2 have the same uncertainty degree. But, it is obviously that the x_1^1 has less uncertainty (i.e., larger certainty) belonging to c_1 than x_1^2. Therefore, the lower approximation of classical FRS performs poorly for evaluating the certainty of data in this case, where the data has a large class density difference.

It is crucially important to overcome both of these problems, which significantly impact the generalization of FRS models, subsequently degrading the overall performance of the downstream classification task. However, most existing FRS-FS methods only aim at addressing one of the two problems. Therefore, the robust LRM-k-PFRS model is developed to improve the antinoise ability and overcome the sensitiveness of data distribution concurrently, by combining the distribution-aware linear reconstruction relation with kNN granules in this article.

III. Robust PseudoFRS Model

Considering the sensitiveness of the classical FRS model to noise information and data distribution, the distribution-aware linear reconstruction relation is developed with the support of

LRM [42] first and then the robust LRM-k-PFRS model is proposed by integrating the distribution-aware linear reconstruction relation with kNN granules.

A. Pseudo Fuzzy Similarity Relation

Most fuzzy similarity relations typically only consider the similarity between two samples, inadvertently ignoring the meaningful information about certain inherent relationships between samples, such as the distribution information and density information. Thus, the distribution-aware linear reconstruction relation is developed to serve as the fuzzy similarity relation in this work.

1) Distribution-Aware Linear Reconstruction Relation: The dataset in a supervised task can be represented as a decision table $DT = (\mathbb{U}, \mathbb{C} \cup \mathbb{D})$. Moreover, $X \in R^{n\times m} = [x_1, x_2, \ldots, x_n] = [f_1; f_2; \ldots; f_m]$ represents the data matrix, where n and m denote the number of samples and features, respectively, the ith row x_i represents the ith sample, and the jth column f_j represents the jth feature. $\bar{X}_t \in R^{n\times 1} = [x_1, x_2, \ldots, x_n] = [f_t]$ consists only of a single conditional feature $f_t \in \mathbb{C}$.

For each data sample x_i, LRM [42] intends to reconstruct it with all samples in $\mathbb{U}$ by minimizing the Euclidean distance between $\mathbf{w_i}X$ and x_i, where $\mathbf{w_i} \in R^n = [w_{i,1}, w_{i,2}, \ldots, w_{i,n}]$ represents the linear reconstruction coefficient vector and each element $w_{i,j} \in \mathbf{w_i}$ represents the similarity relation between x_i and x_j, that is, utilizing all samples to represent themselves. In this way, based on the LRM, the linear reconstruction coefficient matrix W_t^* on feature $f_t \in \mathbb{C}$ can be formulated as the linear reconstruction process with the L_2-norm regularization term, which is also known as ridge regression

$$\begin{aligned} W_t^* &= \min_W \left\{ \sum_{i=1}^{n} \|x_i - w_i \bar{X}_t\|_2^2 + \alpha \sqrt{\sum\nolimits_{i=1}^{n} \|w_i\|_2^2} \right\} \\ &= \min_W \{\|\bar{X}_t - W\bar{X}_t\|_F^2 + \alpha \|W\|_F^2\}. \end{aligned} \tag{6}$$

where $W \in R^{n\times n} = [w_1, \ldots, w_n]$ denotes the reconstruction coefficient matrix between the samples and themselves; W_t^* is the optimal solution of W; α is the regularization parameter that is used to balance the loss function and regularization term.

Each coefficient $w_{i,j}$ of W_t^* reflects the underlying effect of x_j on x_i, where the positive/zero/negative value of the coefficient indicates the positive/null/negative correlation between x_i and x_j. Especially, a larger absolute value of coefficient indicates the higher correlation between the two samples. To illustrate the meaning of the reconstruction coefficient matrix, a random optimal solution $W_t^* \in R^{5\times 5}$ is supposed as follows:

$$W_t^* = \begin{bmatrix} 0.6 & 0.06 & 0.05 & 0.09 & 0.2 \\ 0.09 & 0.7 & -0.1 & 0.25 & 0.07 \\ 0.08 & -0.02 & 0.6 & 0.17 & 0.13 \\ 0.02 & 0.15 & 0.12 & 0.8 & -0.06 \\ 0.13 & 0.05 & 0.1 & -0.12 & 0.9 \end{bmatrix}.$$

In this example, there are five samples $\{x_1, x_2, x_3, x_4, x_5\}$. The values of coefficients in the ith row of W_t^* imply the similarity correlations between the ith sample x_i and all samples. For example, the correlations between x_1 and x_1, x_2, x_3, x_4, x_5 are 0.6, 0.06, 0.05, 0.09, 0.2, respectively.

In order to obtain the distribution-aware linear reconstruction relation $R_{f_t}^{\text{LRM}}$, perform the following operations on W_t^*: For $\forall i, j \in (0, n]$, 1) $W_t^*[i][j] = |W_t^*[i][j]|$, which does not impact the evaluation of membership degree; 2) $W_t^*[i][j] = W_t^*[j][i] = \min(W_t^*[i][j], W_t^*[j][i])$, which is implemented to further improve the robustness. Thus, the linear reconstruction relation $R_{\{f_t\}}^{\text{LRM}}(x_i, x_j)$ is $W_t^*[i][j]$. For $\forall P \in \mathbb{C}$, the linear reconstruction relation induced by P is computed by

$$R_P^{\text{LRM}}(x_i, x_j) = \cap_{f_t \in P} R_{\{f_t\}}^{\text{LRM}}(x_i, x_j). \tag{7}$$

Obviously, R_P^{LRM} satisfies non-negativity and symmetry, but does not satisfy reflexivity. Therefore, the proposed model constructed using R_P^{LRM} is not a FRS model in the strict sense, we call it the pseudo FRS model. Since the membership degree of data sample to a specific class is primarily determined by its similarity correlation between itself and its neighbors independent of reflexivity, R_P^{LRM} can effectively evaluate the similarity between samples by incorporating the distributional information of samples, serving as the fuzzy similarity relation. Moreover, R_P^{LRM} is monotonic, as follows.

Property 1: Let $B_1 \subseteq B_2 \subseteq \mathbb{C}$, then, $R_{B_2}^{\text{LRM}} \subseteq R_{B_1}^{\text{LRM}}$.

Proof: According to the (7), for $\forall x, y \in \mathbb{U}$, $B_1 \subseteq B_2 \subseteq \mathbb{C}$, we have $R_{B_2}^{\text{LRM}}(x, y) = \cap_{a \in B_2} R_a^{\text{LRM}}(x, y) \leq \cap_{a \in B_1} R_a^{\text{LRM}}(x, y) = R_{B_1}^{\text{LRM}}(x, y)$. Thus, it can be concluded that $R_{B_2}^{\text{LRM}} \subseteq R_{B_1}^{\text{LRM}}$.

2) Theoretical Analysis: It is reported that in [42], the linear reconstruction coefficients can effectively indicate the similarity correlation between a given sample and the other samples. Let x_0 be a given sample which is represented as $x_0 = \sum_{i=1}^{n} w_{0,i} x_i$, where $\mathbf{w_0} = [w_{0,1}, w_{0,2}, \ldots, w_{0,n}]$ denotes the reconstruction coefficient vector generated by the LRM. The $w_{0,j}$ is computed by

$$w_{0,j} = \frac{1}{2}\left(d^2\left(\sum_{i=1, i\neq j}^{n} w_{0,i} x_i, x_0 \right) - d^2(x_0, x_j) \right). \tag{8}$$

where $d(x_0, x_j)$ represents the Euclidean distance between x_0 and x_j. The detailed description of the derivation process of (8) can be consulted in [42]. In (8), the $d^2(\sum_{i=1, i\neq j}^{n} w_{0,i} x_i, x_0)$ represents the Euclidean distance between x_0 and other representing samples $x_i, i \neq j$. The larger $d^2(\sum_{i=1, i\neq j}^{n} w_{0,i} x_i, x_0)$ indicates that x_0 is more prominent for forming the linear reconstruction subspace. The larger $w_{0,j}$ (i.e., the larger $d^2(\sum_{i=1, i\neq j}^{n} w_{0,i} x_i, x_0)$ as well as the smaller $d^2(x_0, x_j)$), the larger similarity correlation between x_0 and x_j is. From (8), it can be indicated that the linear reconstruction coefficients obtained by LRM can consider not only the information of Euclidean distance between two samples, but also the relationships of all data samples. In this way, inherent similarity information embedded in geometric distribution of data can be captured and reflected. Thus, the linear reconstruction coefficients coded by LRM are used as the fuzzy similarity relation to determine the kNN granules in LRM-k-PFRS, which can select kNN granules with more meaningful information.

Following the use of the example in Fig. 1, where the samples with the label c_2 in Fig. 1(a) and 1(b) are denoted as $Y = [y_1, \ldots, y_{10}]$ and $Z = [z_1, \ldots, z_{10}]$, respectively, the lower approximations of x_1^1 and x_1^2 can be calculated as follows. First, according to (6), the LRM of x_1^1 and x_1^2 can be formulated with Y and Z, as follows:

$$w^1 = \min_w \{\|x_1^1 - wY\|_2^2 + \alpha\|w\|_2^2\}$$
$$w^2 = \min_w \{\|x_1^2 - wZ\|_2^2 + \alpha\|w\|_2^2\}.$$

where $w^1 = [w_1^1, \ldots, w_{10}^1]$ and $w^2 = [w_1^2, \ldots, w_{10}^2]$. Thus, according to (8), the distribution-aware linear reconstruction relations can be represented approximatively as

$$R_{P_1}(x_1^1, y_1) = w_1^1 = \frac{1}{2}\left(d^2\left(\sum_{t=2}^{10} w_t^1 y_t, y_1\right) - d^2(x_1^1, y_1)\right)$$
$$R_{P_2}(x_1^2, z_1) = w_1^2 = \frac{1}{2}\left(d^2\left(\sum_{t=2}^{10} w_t^2 z_t, z_1\right) - d^2(x_1^2, z_1)\right).$$

Finally, the pseudofuzzy rough lower approximations of x_1^1 and x_1^2 can be obtained

$$\underline{R_{P_1}}c_1(x_1^1) = 1 - R_{P_1}(x_1^1, y_1)$$
$$\underline{R_{P_2}}c_1(x_1^2) = 1 - R_{P_2}(x_1^2, z_1).$$

For $R_{P_1}(x_1^1, y_1)$ and $R_{P_2}(x_1^2, z_1)$, it can be seen from Fig. 1 that, $d(x_1^1, y_1) = d(x_1^2, z_1) = d$, thus $d^2(x_1^1, y_1) = d^2(x_1^2, z_1)$. Moreover, since the density of c_2 with distribution D_1 is larger than that with distribution D_2, $d^2(\sum_{t=2}^{10} w_t^1 y_t, y_1) < d^2(\sum_{t=2}^{10} w_t^2 z_t, z_1)$. Therefore, $R_{P_1}(x_1^1, y_1) < R_{P_2}(x_1^2, z_1)$. A further conclusion can be drawn: $\underline{R_{P_1}}c_1(x_1^1) > \underline{R_{P_2}}c_1(x_1^2)$. Thus, the x_1^1 has larger certainty degree of belonging to c_1 than x_1^2.

Importantly, it requires that the fuzzy similarity relation in FRS should be appropriate for different data distributions. The results analyzed above show the LRM can capture important information, such as the distribution information of data samples and the density information of data classes. Thus, distribution-aware linear reconstruction relation has good generalization performance and can solve the problem of sensitiveness of data distribution mentioned in Section II.

B. Pseudo FRS Model

By using the pseudofuzzy similarity relation, a robust pseudo FRS model called LRM-k-PFRS is proposed.

Definition 1: Given a $DT = (\mathbb{U}, \mathbb{C} \cup \mathbb{D})$, $P \in \mathbb{C}$ and $\mathbb{D}_j \in \mathbb{U}/\mathbb{D}$. In LRM-$k$-PFRS, the pseudolower and upper approximations of $y \in \mathbb{U}$ with respect to P are defined, respectively, as follows:

$$\underline{R_{kP}}^{\mathrm{LRM}}(\mathbb{D}_j)(y) = \frac{1}{k}\sum_{i=1}^{k}(1 - R_P^{\mathrm{LRM}}(y, x_i)) \tag{9}$$

where $\{x_1, x_2, \ldots, x_k\}$ is the set of kNN (i.e., has the k largest linear reconstruction coefficients) of $y \in \mathbb{U}$ in subset $\mathbb{U} - \mathbb{D}_j$

$$\overline{R_{kP}}^{\mathrm{LRM}}(\mathbb{D}_j)(y) = \frac{1}{k}\sum_{i=1}^{k} R_P^{\mathrm{LRM}}(y, x_i) \tag{10}$$

where $\{x_1, x_2, \ldots, x_k\}$ is the set of kNN (i.e., has the k largest linear reconstruction coefficients) of $y \in \mathbb{U}$ in subset $\mathbb{D}_j$.

As the calculation of the fuzzy rough approximations of a sample only cover the dissimilarity information of the nearest sample in classical FRS, the precision of the membership degree to the positive domain may be degraded. This will further destroy the accuracy of the lower approximation when the dataset contains noisy information. To alleviate the problem of sensitiveness of noise in data mentioned in Section II, the strategy of kNN granules is employed when calculating the pseudofuzzy rough approximations. Moreover, linear reconstruction coefficients used to determine the kNN granules consider not only the distance information between two samples, but also the relationship between all samples, which can capture more meaningful information.

Definition 2: Given a $DT = (\mathbb{U}, \mathbb{C} \cup \mathbb{D})$, $P \in \mathbb{C}$ and $\mathbb{D}_j \in \mathbb{U}/\mathbb{D}$. In LRM-$k$-PFRS, the pseudofuzzy dependency degree of $\mathbb{D}$ with respect to P is defined, as follows:

$$\gamma_{kP}^{\mathrm{LRM}}(\mathbb{D}) = \frac{\sum_{y \in \mathbb{U}} \max_{j=1}^{L} \underline{R_{kP}}^{\mathrm{LRM}}(\mathbb{D}_j)(y)}{|\mathbb{U}|}. \tag{11}$$

Since the pseudofuzzy rough lower approximation can indicate that the membership degree of a data sample certainly belongs to a specific class, the pseudofuzzy dependency degree reflects the percentage of samples that can be exactly classified in terms of the feature subset $P \in \mathbb{C}$.

The $\underline{R_{kP}}^{\mathrm{LRM}}(\mathbb{D}_j)(y)$, $\overline{R_{kP}}^{\mathrm{LRM}}(\mathbb{D}_j)(y)$ and $\gamma_{kP}^{\mathrm{LRM}}(\mathbb{D})$ satisfies the following properties:

Property 2: Let $B_1 \subseteq B_2 \subseteq \mathbb{C}$, then, $\underline{R_{kB_1}}^{\mathrm{LRM}}(\mathbb{D}_j)(y) \leq \underline{R_{kB_2}}^{\mathrm{LRM}}(\mathbb{D}_j)(y)$.

Property 3: Let $B_1 \subseteq B_2 \subseteq \mathbb{C}$, then, $\overline{R_{kB_2}}^{\mathrm{LRM}}(\mathbb{D}_j)(y) \leq \overline{R_{kB_1}}^{\mathrm{LRM}}(\mathbb{D}_j)(y)$.

Property 4: Let $B_1 \subseteq B_2 \subseteq \mathbb{C}$, then, $\gamma_{kB_1}^{\mathrm{LRM}}(\mathbb{D}) \leq \gamma_{kB_2}^{\mathrm{LRM}}(\mathbb{D})$.

Proof: According to the Property 1, for a given k and $\forall y \in \mathbb{U}$, since $B_1 \subseteq B_2 \subseteq \mathbb{C}$ and $R_{B_2}^{\mathrm{LRM}} \subseteq R_{B_1}^{\mathrm{LRM}}$, we have $\frac{1}{k}\sum_{i=1}^{k} R_{B_2}^{\mathrm{LRM}}(y, x_i) \leq \frac{1}{k}\sum_{i=1}^{k} R_{B_1}^{\mathrm{LRM}}(y, x_i)$. Thus, $\overline{R_{kB_2}}^{\mathrm{LRM}}(\mathbb{D}_j)(y) \leq \overline{R_{kB_1}}^{\mathrm{LRM}}(\mathbb{D}_j)(y)$. Moreover, we have $\frac{1}{k}\sum_{i=1}^{k}(1 - R_{B_1}^{\mathrm{LRM}}(y, x_i)) \leq \frac{1}{k}\sum_{i=1}^{k}(1 - R_{B_2}^{\mathrm{LRM}}(y, x_i))$. Thus, $\underline{R_{kB_1}}^{\mathrm{LRM}}(\mathbb{D}_j)(y) \leq \underline{R_{kB_2}}^{\mathrm{LRM}}(\mathbb{D}_j)(y)$. Further, $\gamma_{kB_1}^{\mathrm{LRM}}(\mathbb{D}) \leq \gamma_{kB_2}^{\mathrm{LRM}}(\mathbb{D})$.

It can be indicated from the above analysis that, the pseudofuzzy dependency degree in LRM-k-PFRS is monotonic, which can be used to guide the FS process to select the discriminative features in this article.

IV. Feature Grouping and Selection

LRM-k-PFRS model is employed to guide a robust FS algorithm called LRM-RPFRFS from the perspective of *RF*, *SRP* and *DS* (i.e., the selection strategy of *RF-SRP-DS*) to determine

TABLE I
ORIGINAL DATASET

$x \in \mathbb{U}$	f_1	f_2	f_3	f_4	f_5	f_6	f_7	f_8	f_9	$\Longrightarrow$	$\mathbb{D}$
x_1	56	79.4	127.5	4	2.1	24	12.6	98	48		0
x_2	40.6	68.9	115.8	8	3.5	36	17.8	106	65		0
x_3	4.5	6.7	21	430	90	4200	1.8	8.9	4.7		0
x_4	7.6	10.7	45	740	160	6400	1.4	7.5	4.2		0
x_5	11.7	15	101	10	4.6	48	24.8	148	76		1
x_6	8.9	12.7	67	18	8.6	87	32.6	152	84		1
x_7	2.3	4.8	8.8	320	65	3500	0.23	1.2	2.1		1
x_8	6.8	8.7	39	580	120	5300	0.68	2.6	2.6		1

the final feature subset. After implementing the feature grouping as a pretreatment, the process of FS is implemented.

A. Feature Grouping

As an essential step for implementing LRM-RPFRFS, potentially redundant features are grouped. First, the correlation matrix $W_{\mathrm{CM}} \in R^{m\times m}$ is calculated, where we design to use all features in $\mathbb{C}$ to reconstruct themselves, i.e., reconstruct each conditional feature $f_t \in \mathbb{C}$ with all features, as follows:

$$\begin{aligned} W_{\mathrm{CM}} &= \min_{W}\left\{\sum_{t=1}^{m}\|f_i - Xw_i\|_2^2 + \gamma\sqrt{\sum\nolimits_{t=1}^{m}\|w_i\|_2^2}\right\} \\ &= \min_{W}\{\|X - XW\|_F^2 + \gamma\|W\|_F^2\} \end{aligned} \tag{12}$$

where $W \in R^{m\times m} = [w_1, \ldots, w_m]$ denotes the reconstruction coefficient matrix between each feature and themselves; W_{CM} is the optimal solution of W; γ is the regularization parameter. Similar to the W_t^*, W_{CM} is also subjected to the following treatments: For $\forall i, j \in (0, m]$, 1) $W_{\mathrm{CM}}[i][j] = |W_{\mathrm{CM}}[i][j]|$; 2) $W_{\mathrm{CM}}[i][j] = W_{\mathrm{CM}}[j][i] = \min(W_{\mathrm{CM}}[i][j], W_{\mathrm{CM}}[j][i])$. Then, the undirected correlation graph $G = \{\vartheta, \varepsilon, W_{\mathrm{CM}}\}$ is constructed, where $\vartheta = \{f_t | t = 1, \ldots, m\}$ is the vertex set with m features; ε is the edge set; the weight of the edge between pairwise features $f_i, f_j \in \mathbb{C}(i \neq j)$ is $W_{\mathrm{CM}}[i][j]$. The G is quite complex in analysing the redundancy between features, particularly when the number of features is large. Thus, a maximum spanning tree (MST) of G is generated by using the same mentality of the Kruskal's algorithm for creating the minimum spanning tree to streamline the number of edges, where the generated MST covers all vertices with the least edges. After that, a pruning strategy is implemented. Because a vertex may be connected to more than one node in MST, in this case, each vertex retains the only edge with the largest weight and the other edges are deleted in the pruning strategy. Finally, features that are connected each other are the members of the same group.

To illustrate this process, a simple dataset is used as in Table I, consisting of eight objects, each involving nine conditional features and one decision feature. In advance of feature grouping, the values of each feature are normalized using the min–max normalization method, mapping the original feature measurements made on different scales onto a notionally common scale, as in Table II. According to the $W_{\mathrm{CM}} \in R^{9\times 9}$, the G is constructed. After that, the process of feature grouping is shown in Fig. 2. First, the MST is created using Kruskal's algorithm. Then, the edges between pairwise features (f_3, f_5) and (f_3, f_8) are deleted after implementing the pruning strategy. Thus, all features are partitioned into three groups, which are $\{f_1, f_2, f_3\}$, $\{f_4, f_5, f_6\}$, $\{f_7, f_8, f_9\}$. It can be seen from the Table II, the features within the same group possess nearly redundant information, leading to parallel discriminative capabilities.

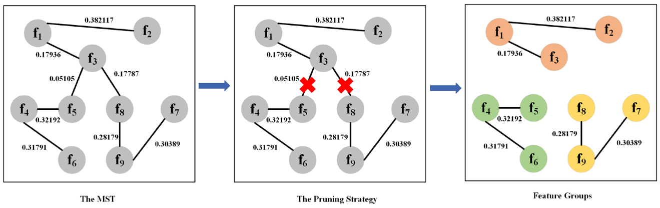


Fig. 2. Process of feature grouping.

Graphs provide an intuitive representation of relationships between features, enhancing the comprehension of how features interrelate. In addition, graph theory-based feature grouping techniques can effectively identify groups of features that are highly correlated or redundant. Compared to existing feature grouping-based FRS-FS approaches [18], [19], *RF-SRP-DS* employs LRM to calculate the relationships between features to construct the undirected correlation graph, which thoroughly considers the distributional relationships and captures more important information.

B. Feature Selection

Given the resulting groups of features, the selection strategy of *RF-SRP-DS* is employed to implement the robust FS, comprehensively considering three essential aspects: redundancy, relevance, and discrimination.

Redundant Filter: Exclusive lasso (EL) is utilized to select the representative features from each group, which formulates the process of linear reconstruction with the $L_{1,2}$-norm. The m features are partitioned into g groups in the process of feature grouping

$$\begin{aligned} \widehat{w} &= \min_{w}\{\|y - \mathbf{X}w\|_2^2 + \lambda\|w\|_1^2\}, \\ \|w\|_1^2 &= \sqrt{\sum\nolimits_{i=1}^{g}\left(\sum\nolimits_{j=1}^{N_i}|w_{ij}|\right)^2}. \end{aligned} \tag{13}$$

Here, $\widehat{w} = [\widehat{w}_1, \widehat{w}_2, \ldots, \widehat{w}_m]^T$ is the reconstruction coefficients of EL; y represents the class labels of all instances; λ is the regularization parameter and N_i is the quantity of the features in the ith group. The $L_{1,2}$-norm enforces intergroup nonsparsity via L_2-norm over members within each group and intragroup sparsity via L_1-norm over groups. In this manner, EL performs FS by guaranteeing the inclusion of at least one feature will be selected from each group. Consequently, any redundant features within these groups concerning the decision-making process will be eliminated, that is, the features with zero reconstruction coefficients are discarded within each group.

Strongly Relevant Priority: The representative features with the nonzero linear reconstruction coefficients are ranked in terms of the absolute value of their coefficients, where the more

TABLE II
NORMALIZED DATASET

$x \in \mathbb{U}$	f_1	f_2	f_3	f_4	f_5	f_6	f_7	f_8	f_9	$\Longrightarrow$	$\mathbb{D}$
x_1	1.0000	1.0000	1.0000	0.0000	0.0000	0.0000	0.3821	0.6419	0.5604		0
x_2	0.7132	0.8592	0.9014	0.0054	0.0088	0.0019	0.5428	0.6949	0.7680		0
x_3	0.0409	0.0254	0.1027	0.5788	0.5566	0.6549	0.0485	0.0510	0.0317		0
x_4	0.0986	0.0790	0.3049	1.0000	1.0000	1.0000	0.0361	0.0417	0.0256		0
x_5	0.1750	0.1367	0.7767	0.0081	0.0158	0.0037	0.7590	0.9734	0.9023		1
x_6	0.1229	0.1058	0.4903	0.0190	0.0411	0.0098	1.0000	1.0000	1.0000		1
x_7	0.0000	0.0000	0.0000	0.4293	0.3983	0.5451	0.0000	0.0000	0.0000		1
x_8	0.0837	0.0522	0.2544	0.7826	0.7466	0.8274	0.0139	0.0092	0.0061		1

relevant features associated with the decision feature are granted the higher rankings within each group.

Discriminative Selection: LRM-RPFRFS utilizes the interior rankings of the representative features within their own groups to implement the search for determining the final discriminative feature subset of the original dataset. The pseudofuzzy dependency degree defined in LRM-k-PFRS model [i.e., (11)] is employed to measure the discriminative capacity of the candidate subset for guiding the FS process. During each iteration, priority is given to the top-ranked representative features within each group when updating the selected features. The feature contributing the most significant growth in pseudofuzzy dependency degree is then added to the current subset and removed from its original group. If a group becomes empty as a result of the feature removal, no further operations are conducted on that specific group. The stopping criteria consists of two aspects: 1) the improvement of the pseudofuzzy dependency degree between two successive iterations is negligibly small; 2) the pseudofuzzy dependency degree of the selected feature subset reaches that of the entire set of conditional features.

Following the use of the example dataset in Table I, the optimal solution of (13) is

$$\widehat{w} = [0, -0.801, 0, 0, -0.105, 0, 0.536, 0.002, 0].$$

It can be seen from $\widehat{w}$ that features f_2, f_5, f_7, f_8 are selected due to the nonzero coefficients. The ranking results of representative features in each group are

$$\{f_2\}, \{f_5\} \text{ and } \{f_7, f_8\}.$$

After that, the top-ranked features in each group including f_2, f_5, f_7 are considered in the first iteration, whose pseudofuzzy dependency degree values are calculated first

$$\gamma k^{\text{LRM}}_{\{f_2\}}(\mathbb{D}) = 0.96250$$
$$\gamma k^{\text{LRM}}_{\{f_5\}}(\mathbb{D}) = 0.86074$$
$$\gamma k^{\text{LRM}}_{\{f_7\}}(\mathbb{D}) = 0.88377.$$

As the feature f_2 results in the greatest increase in pseudofuzzy dependency degree, it will be selected and added to the candidate feature subset, and then removed from its group. The first feature group becomes empty, therefore, this group will not be considered in any future iterations. In the next iteration, the values of pseudofuzzy dependency degree on $\{f_2, f_5\}$ and $\{f_2, f_7\}$ are considered

$$\gamma k^{\text{LRM}}_{\{f_2,f_5\}}(\mathbb{D}) = 0.99699$$
$$\gamma k^{\text{LRM}}_{\{f_2,f_7\}}(\mathbb{D}) = 0.96533.$$

Thus, the feature f_5 is selected and added to the candidate subset, and then removed from its group. In the next iteration, the respective value of pseudofuzzy dependency degree on $\{f_2, f_5, f_7\}$ is calculated

$$\gamma k^{\text{LRM}}_{\{f_2,f_5,f_7\}}(\mathbb{D}) = 0.99699.$$

Since the improvement of the pseudofuzzy dependency degree is zero, that is, the feature f_7 can not increase the value of pseudo fuzzy dependency degree, the LRM-RPFRFS algorithm terminates. The final reduct returned by LRM-RPFRFS is $\{f_2, f_5\}$.

TABLE III
TIME COMPLEXITY ANALYSIS

Step of Alg.1	Time Complexity
Line 3	$O(n \times m^2)$
Line 4-5	$O(m \times (m-1)/2 \times \log(m \times (m-1)/2))$
Line 6	$O(m)$
Line 8-10	$O(I \times g)$
Line 12–27	$O(s)$–$O(s^2)$

C. Algorithm

The LRM-RPFRFS algorithm is summarized in psuedocode as given in Algorithm 1, where the value of ε in Line 27 is fixed at 0.0005. The time complexity of each step is listed in Table III, where the maximum number of iterations is denoted by I during the optimization for solving the EL [i.e., (13)]; the number of features returned by EL is denoted as s ($g \leq s \leq m$). In general, the worst case of time complexity of LRM-RPFRFS is therefore, $O(n \times m^2 + m \times (m-1)/2 \times \log(m \times (m-1)/2) + m + I \times g + m^2)$.

V. EXPERIMENTAL EVALUATION

This section presents a systematic evaluation of LRM-RPFRFS experimentally. After an introduction to the experimental setup in Section V-A, the parameters training and the experimental results are presented and discussed in Sections V-B and V-C, respectively.

Algorithm 1: Linear Reconstruction Measure-Based Robust Pseudo Fuzzy Rough Feature Selection.

```
Input:
  $\mathbb{C}$, set of all conditional features;
  $\mathbb{D}$, set of decision feature;
  $k$, number of nearest neighbors in LRM-$k$-PFRS;
  $\lambda$, the regularization parameter in exclusive lasso.
Output: $R$, reduct.
$R \leftarrow \emptyset$;
//Feature Grouping.
$W_{CM} \in \boldsymbol{R}^{m\times m} \leftarrow$ Calculate the correlation matrix between features via (13);
$G = \{\vartheta, \varepsilon, W_{CM}\} \leftarrow$ Construct the undirected graph in terms of $W_{CM}$;
$MST = \{\vartheta, \bar{\varepsilon}, W_{CM}\} \leftarrow$ Create the maximum spanning tree of $G$ via Kruskal's algorithm;
$FG = \{FG_1, ..., FG_g\} \leftarrow$ Divide the features into $g$ groups via the pruning strategy;
//Redundant Filter.
$\widehat{w} \leftarrow$ Solve (12);
$FG^s = \{FG_1^s, \ldots, FG_G^s\} \leftarrow$ Select representative features in each group via EL with $\lambda$;
//Strongly Relevant Priority.
$FG^r = \{FG_1^r, \ldots, FG_k^r\} \leftarrow$ Rank representative features in each group in terms of $\widehat{w}$;
//Discriminative Selection.
$\gamma_{k_{prev}}^{LRM}(\mathbb{D}) \leftarrow 0$; $\gamma_{k_{best}}^{LRM}(\mathbb{D}) \leftarrow 0$;
repeat
    $F_{best} \leftarrow \emptyset$; $DEP_{best} \leftarrow 0$;
    $\gamma_{k_{prev}}^{LRM}(\mathbb{D}) \leftarrow \gamma_{k_{best}}^{LRM}(\mathbb{D})$;
    foreach $FG_i^r \in FG^r$ do
        $f \leftarrow$ highestRankedFeature($FG_i^r$);
        $T \leftarrow R \cup \{f\}$;
        if $\gamma_{k_{\{T\}}}^{LRM}(\mathbb{D}) > DEP_{best}$ then
            $F_{best} \leftarrow f; DEP_{best} \leftarrow \gamma_{k_{\{T\})}}^{LRM}(\mathbb{D})$;
            $v \leftarrow i$;
        end
    end
    $FG_v^r \leftarrow FG_v^r - \{F_{best}\}$;
    $R \leftarrow R \cup \{F_{best}\}$;
    $\gamma_{k_{best}}^{LRM}(\mathbb{D}) \leftarrow DEP_{bset}$;
until $(\gamma_{k_{best}}^{LRM}(\mathbb{D}) - \gamma_{k_{prev}}^{LRM}(\mathbb{D})) < \varepsilon$ or $\gamma_{k_{best}}^{LRM}(\mathbb{D}) == \gamma_{k_{\mathbb{C}}}^{LRM}(\mathbb{D})$;
return $R$
```

A. Experimental Setup

1) Datasets: A total of 31 datasets[1,2,3,4] are used for the following experimental evaluation. The basic information about these datasets are summarized in Table IV. Note that these datasets are normalized via the min–max normalization method in this article. Stratified 10 × 10-fold cross validation (10-FCV) is employed in the following experimentation.

TABLE IV
DATASETS USED FOR EVALUATION

Type	#	Dataset	Abbreviation	Instances	Features	Classes
Benchmark Datasets	1	Page-blocks	PAB	5473	11	5
	2	Leaf	LEA	340	16	30
	3	Diabetic	DIA	1151	18	2
	4	Vehicle	VEH	846	19	4
	5	Sobar	SOB	72	20	2
	6	Parkinson	PAR	195	23	2
	7	Wpbc	WPB	197	33	2
	8	Ionosphere	ION	230	35	2
	9	CELL	CEL	12000	35	2
	10	NASA	NAS	4687	36	2
	11	Statlog	STA	654	37	2
	12	CM1	CM1	344	38	2
	13	PC1	PC1	735	38	2
	14	MC2	MC2	125	40	2
	15	KC3	KC3	200	40	2
	16	SPECTFHeart	SPE	267	45	2
	17	Libras	LIB	360	91	15
	18	Bankrupt	BAK	6819	95	2
Noisy Datasets	19	Leaf_noise	LEA_N	340	16	30
	20	Diabetic_noise	DIA_N	1151	18	2
	21	Vehicle_noise	VEH_N	846	19	4
	22	Sobar_noise	SOB_N	72	20	2
	23	Parkinson_noise	PAR_N	195	23	2
Synthetic Datasets	24	Data_1STD	D1S	100	40	4
	25	Data_3STD	D3S	100	40	4
	26	Data_5STD	D5S	100	40	4
Biological Datasets	27	Colon	COL	62	2000	2
	28	Prostate_GE	PRO	102	5966	2
	29	ALLAML	ALL	72	7130	2
Face Datasets	30	Genk	GEN	500	577	2
	31	WarpPIE10P	WAR	210	2421	10

2) Comparison Approaches: In this article, we present a comparison on the reduced size and the classification accuracy of the selected features, between LRM-RPFRFS and six FRS-FS methods (i.e., four feature grouping-based FRS-FS methods and two heuristic FRS-FS methods) and four state-of-the-art FS methods (i.e., two wrapped strategy-based FS methods and two filter strategy-based FS methods).

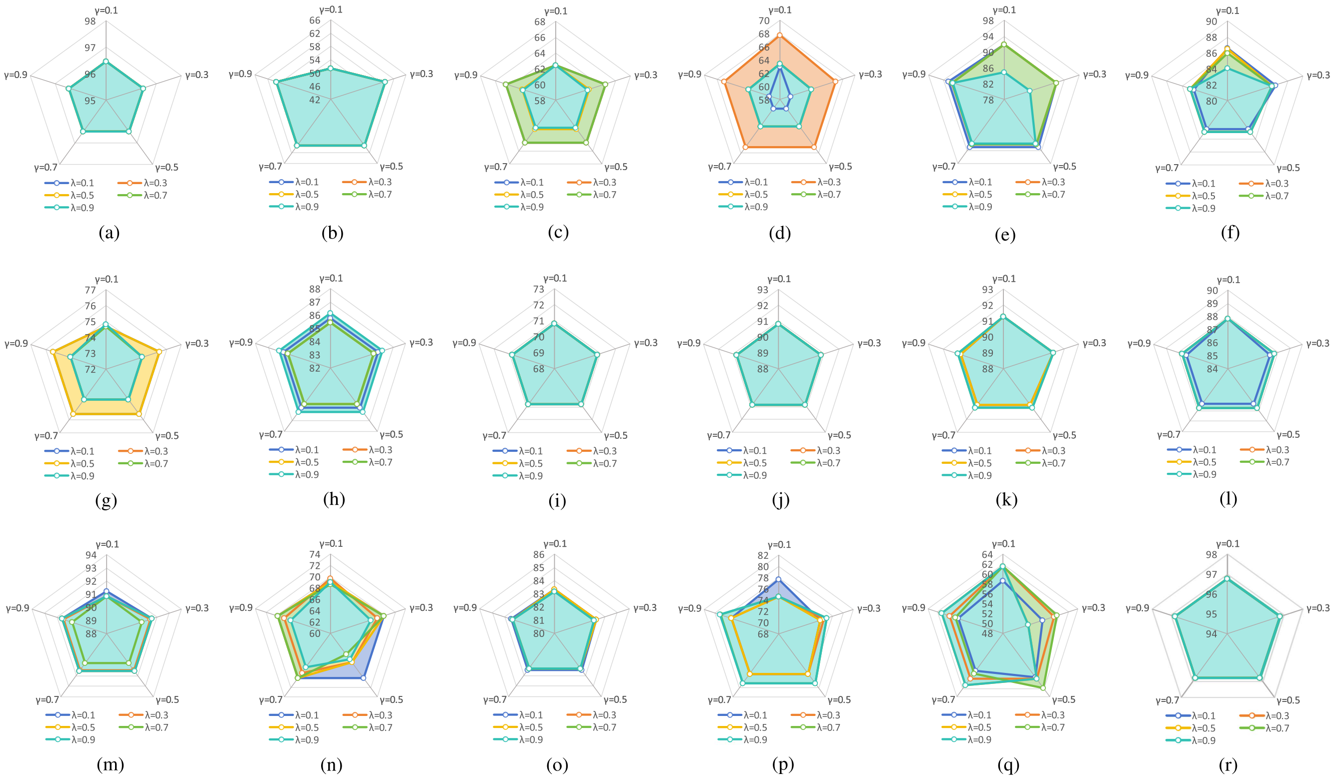
1) Feature grouping-based FRS-FS methods: FRFG [17], GBFG [18], EL-TSFRFS [20], FGS-RFRAS [19];
2) Heuristic FRS-FS methods: FRMR [15], HARCM [16].
3) Wrapped strategy-based FS methods: GSA [43], GWO [44].
4) Filter strategy-based FS methods: PCC [46], ReliefF [45].

3) Classifiers: The determined reducts are evaluated by the following four different classifiers: J48 [47], Bagging [48], Jrip [49], and Part [50], respectively.

B. Training Parameters

As shown in Sections III and IV, LRM-RPFRFS has four parameters: 1) The regularization parameter γ in (12) is used to form feature groups. 2) The regularization parameter λ in (13) is used to filter redundant features in each group. 3) The regularization parameter α in (6) is used to construct the distribution-aware linear reconstruction relation. 4) The number of nearest neighbors k in (9) is used to calculate the pseudofuzzy rough approximations. Ideally, for each dataset, we should search throughout the entire range space to determine the optimal values of the four parameters, which results in a significant amount of time consumption. Fortunately, we have discovered that these four parameters can be divided into two independent groups for training, i.e., γ & λ and α & k, since the FS process guided by LRM-k-PFRS (which involves two parameters α and k) relies

[1] [Online]. Available: https://archive.ics.uci.edu/ml/index.php
[2] [Online]. Available: https://jundongl.github.io/scikit-feature/datasets.html
[3] [Online]. Available: https://github.com/klainfo/NASADefectDataset
[4] [Online]. Available: https://www.kaggle.com/datasets



(a) (b) (c) (d) (e) (f)

(g) (h) (i) (j) (k) (l)

(m) (n) (o) (p) (q) (r)

Fig. 3. Classification accuracy (%) versus regularization parameter γ versus regularization parameter λ, by J48. (a) PAB. (b) LEA. (c) DIA. (d) VEH. (e) SOB. (f) PAR. (g) WPB. (h) ION. (i) CEL. (j) NAS. (k) STA. (l) CM1. (m) PC1. (n) MC2. (o) KC3. (p) SPE. (q) LIB. (r) BAK.

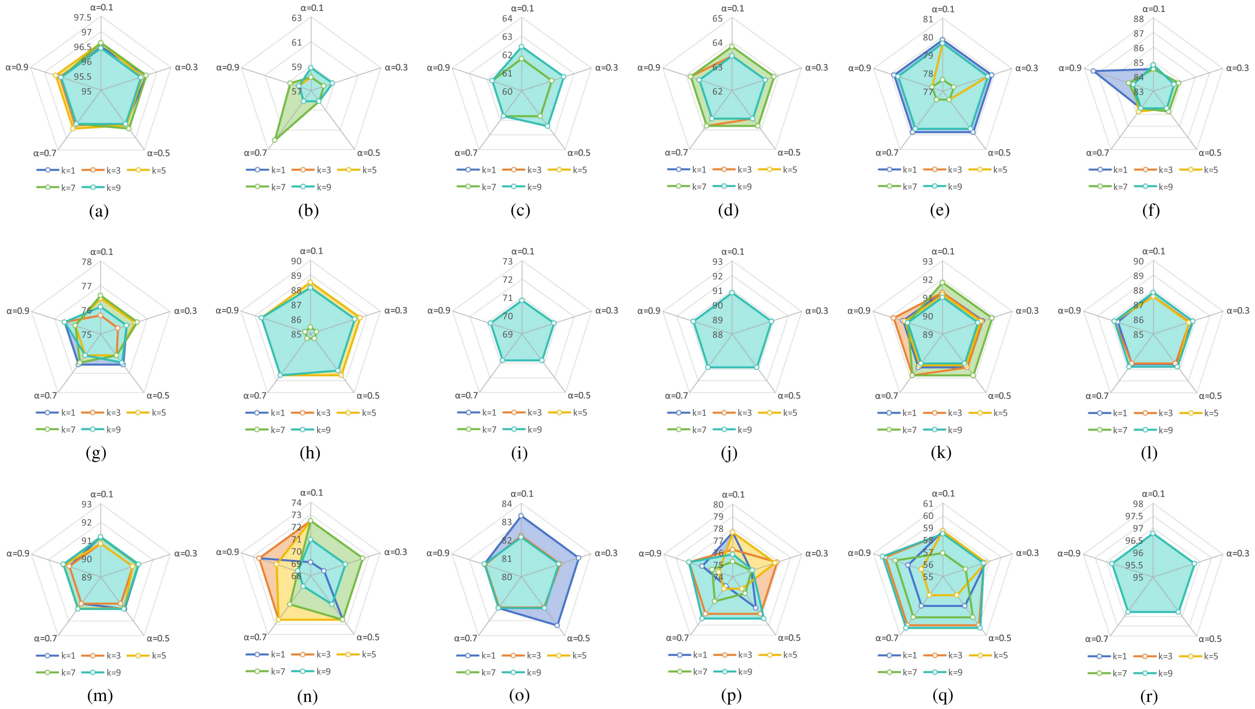
on the result obtained by filtering out the redundant ones from each resulted feature group (which involves two parameters γ and λ) as input.

The values of regularization parameters α, λ, and γ are, respectively, set to 0.1, 0.3, 0.5, 0.7, and 0.9. The values of k are, respectively, set to 1, 3, 5, 7, and 9. As indicated in a substantial amount of literatures, to approximately achieve the optimal performance of the proposed approach, we can fix a parameter at a specific value while searching the optimal value of another parameter across its entire range. Next, we will determine which parameter of each group to fix through experimental analysis. Due to space constraints, this article primarily presents the results obtained on the 18 benchmark datasets.

1) γ & λ: We validate the impact of γ and λ on the experimental results by evaluating the classification accuracies of the obtained feature subsets on the J48 classifier. In the following experimental results as shown in Fig. 3, we fix the values of α at 0.1 and k at 1. It can be seen from Fig. 3 that, there are five cases in the experimental results: 1) for four datasets, i.e., *PAB*, *CEL*, *NAS*, *BAK*, the values of both γ and λ have no impact on the experimental results; 2) for one dataset, i.e., *LEA*, the values of γ have an impact on the experimental results; 3) for five datasets, i.e., *ION*, *STA*, *CM1*, *PC1*, *KC3*, the values of λ have an impact on the experimental results; 4) for six datasets, i.e., *DIA*, *VEH*, *SOB*, *PAR*, *WPB*, *MC2*, the values of both γ and λ have an impact on the experimental results, where the effect of λ is more significant; 5) for two datasets, i.e., *SPE*, *LIB*, the values of both γ and λ have a significant impact on the experimental results.

In summary, we fix γ at a specific value and search for the optimal value of λ, since γ has a minor impact on the experimental results for most cases. For different values of λ, except for three datasets, i.e., *VEH*, *SPE*, *LIB* (accuracy losses are 4.7%, 0.36%, and 2.8%, respectively, which are acceptable), all other datasets can achieve their maximum values when γ is set to 0.1. In addition, for the majority of datasets, as γ increases, the number of feature groups either remains constant or increases. Due to the impact of the number of feature groups on the computational complexity of the subsequent FS process, we set γ to 0.1 in order to enhance the efficiency of the LRM-RPFRFS.

2) α & k: We validate the impact of α and k on the experimental results by evaluating the classification accuracies of the obtained feature subsets on the J48 classifier. In the following experimental results as shown in Fig. 4, we fix the values of γ and λ at 0.1. It can be seen from Fig. 4 that, there are five cases in the experimental results: 1) for three datasets, i.e., *CEL*, *NAS*, *BAK*, the values of both α and k have no impact on the experimental results; 2) for four datasets, i.e., *PAB*, *CM1*, *PC1*, *KC3*, the values of k have an impact on the experimental results; 3) for two datasets, i.e., *LEA*, *PAR*, the values of both α and k have an impact on the experimental results, where the effect of α is more significant; 4) for eight datasets, i.e., *DIA*, *VEH*, *SOB*, *ION*, *STA*, *MC2*, *SPE*, *LIB*, the values of both α and k have an impact on the experimental results, where the effect of k is more



(a) (b) (c) (d) (e) (f)

(g) (h) (i) (j) (k) (l)

(m) (n) (o) (p) (q) (r)

Fig. 4. Classification accuracy (%) versus number of nearest neighbors k versus regularization parameter α, by J48. (a) PAB. (b) LEA. (c) DIA. (d) VEH. (e) SOB. (f) PAR. (g) WPB. (h) ION. (i) CEL. (j) NAS. (k) STA. (l) CM1. (m) PC1. (n) MC2. (o) KC3. (p) SPE. (q) LIB. (r) BAK.

significant; 5) for one dataset, i.e., *WPB*, the values of both α and k have a significant impact on the experimental results.

In summary, k has a more significant impact on the experimental results than α for most cases. Therefore, we fix α at a specific value and search for the optimal value of k. For different values of α, except for four datasets, i.e., *LEA*, *PAR*, *SPE*, *LIB* (accuracy losses are 3.24%, 2.55%, 0.64%, and 1.53%, respectively, which are acceptable), all other datasets can achieve their maximum values when α is set to 0.1. Moreover, as long as the value of α is fixed at 0.1, the variation of k can cover all or more possible selected feature subsets for most datasets.

3) Parameters Setting: Based on the above analysis, the values of γ and α are set to 0.1. Next, we proceed to set the values of λ and k.

Taking the Bagging classifier as representative, the classification accuracies on 18 benchmark datasets are shown in Fig. 5. Looking into Fig. 5, the classification accuracies do not exhibit any specific pattern with the variation of λ and k. Moreover, for the different values of λ and k, the change of the parameters may lead to different gaps in classification accuracy: 1) two datasets, i.e., *LEA*, *WAR*, are significant gap (12.59% of *LEA* and 11.70% of *WAR*); 2) two datasets, i.e., *PC1*, *BAK*, are tiny gap (0.61% of *PC1* and 0.90% of *BAK*); 3) the remaining datasets are small gap (the maximum is 4.63% of *PAB* and the minimum is 1.00% of *CEL*). Thus, the values of λ and k are set to the values associated with the highest classification accuracies achieved by the specific classifier. In addition, given four different classifiers (i.e., J48, Bagging, Jrip, and Part), the values of k and λ corresponding to the highest classification accuracies are typically different. Consequently, distinct values of k and λ are set for each classifier on 31 datasets, as displayed in Tables VI–IX.

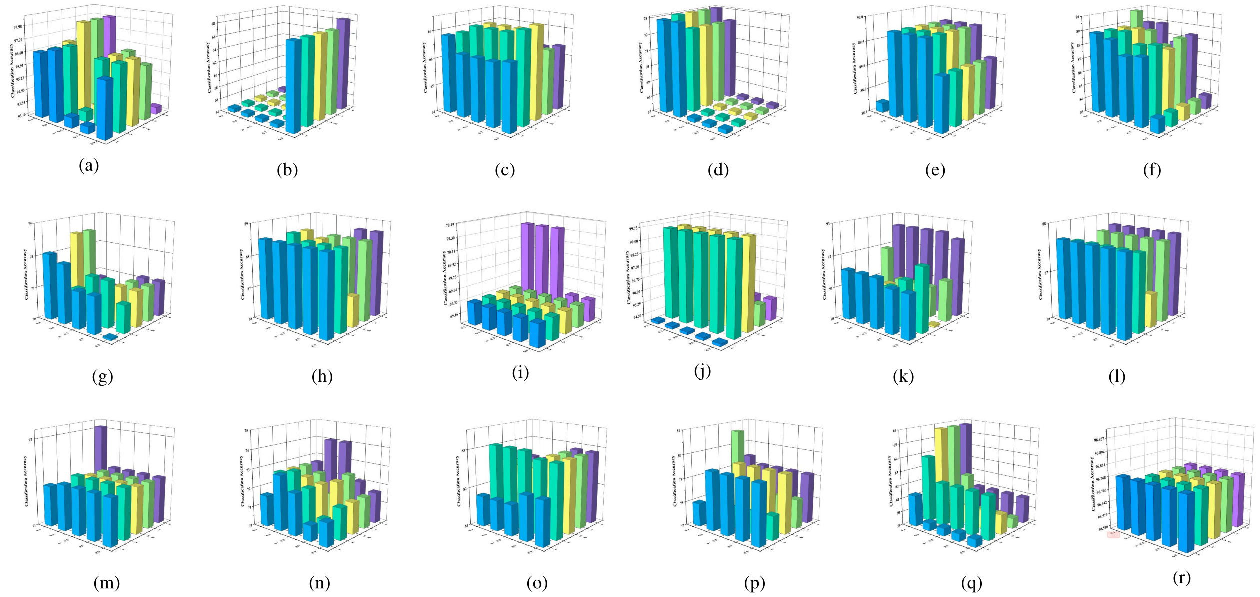
TABLE V
REDUCT SIZE

Dataset	**LRM-RPFRFS**	FRFG	GBFG	EL-TSFRFS	FGS-RFRAS	FRFM	HARCM	GSA	GWO	PCC	RELIEF
PAB	3	9	10	6	4	5	4	1	1	3	3
LEA	7	9	14	12	3	9	12	11	6	7	7
DIA	7	8	18	10	4	5	13	5	8	7	7
VEH	13	9	9	10	4	4	12	8	6	13	13
SOB	5	5	4	7	5	6	6	3	5	5	5
PAR	5	6	6	7	6	6	8	7	7	5	5
WPB	7	6	6	7	9	8	16	12	21	5	7
ION	8	8	8	8	8	13	15	10	15	8	8
CEL	4	30	32	0	3	4	31	7	12	4	4
NAS	5	8	9	9	6	7	1	7	10	5	5
STA	4	8	9	11	12	3	19	11	17	4	4
CM1	4	8	8	9	6	4	17	9	11	4	4
PC1	4	18	37	10	9	4	17	10	16	4	4
MC2	4	16	9	8	8	4	12	12	18	4	4
KC3	5	14	21	8	8	7	13	12	18	4	4
SPE	9	7	7	8	7	12	21	18	26	9	9
LIB	8	11	8	8	16	7	50	18	26	9	9
BAK	2	10	11	2	11	3	52	28	18	2	2
LEA_N	5	9	14	13	14	8	14	4	7	5	5
DIA_N	3	8	17	10	17	12	17	7	6	3	3
VEH_N	5	9	9	13	9	6	8	9	13	5	6
SOB_N	5	5	5	9	4	4	6	2	6	5	5
PAR_N	5	6	6	9	5	5	6	4	7	5	5
D1S	9	3	3	3	2	2	2	2	12	3	3
D3S	14	3	3	3	3	2	2	2	12	3	3
D5S	15	3	3	3	2	2	2	2	12	3	3
COL	22	5	6	5	5	5	6	943	687	22	20
PRO	4	5	5	5	4	4	4	2959	1903	4	4
ALL	6	4	5	4	3	3	3	3513	2969	6	6
GEN	8	7	8	7	6	6	5	278	237	8	8
WAR	5	6	7	6	6	5	6	1120	773	5	5

As in general, the choice of the values for λ and k affects the performance of LRM-RPFRFS slightly. Thus, if the requirement of classification accuracy for the current task is not very high, the parameters can be selected randomly. Otherwise, a careful offline selection of the appropriate parameters is necessary before LRM-RPFRFS is applied.

C. Results and Analysis

1) Size of Reduct: By summarizing the size of the reducts over each dataset, Table V presents a comparison on the reduced size of the selected features, between LRM-RPFRFS and 10 comparison approaches. The experimental results collectively show that, the reduct size obtained by the LRM-RPFRFS approach is smaller than or comparable to those achievable with



(a) (b) (c) (d) (e) (f)

(g) (h) (i) (j) (k) (l)

(m) (n) (o) (p) (q) (r)

Fig. 5. Classification accuracy (%) versus number of nearest neighbors k versus regularization parameter λ, by Bagging. (a) PAB. (b) LEA. (c) DIA. (d) VEH. (e) SOB. (f) PAR. (g) WPB. (h) ION. (i) CEL. (j) NAS. (k) STA. (l) CM1. (m) PC1. (n) MC2. (o) KC3. (p) SPE. (q) LIB. (r) BAK.

TABLE VI
J48 CLASSIFICATION ACCURACY (%) WITH REDUCTS RETURNED BY DIFFERENT FS METHODS

Dataset	**LRM-RPFRFS** (k,λ)	RAW-DATA	FRFG	GBFG	EL-TSFRFS	FGS-RFRAS	FRFM	HARCM	GSA	GWO	PCC	RELIEF
PAB	97.82 (5,0.3)	96.99	96.99	96.99	96.75	96.49	95.78	96.50	92.83	93.63	95.90	94.14
LEA	71.46 (1,0.9)	61.71*	71.46	71.46	63.29*	27.82*	60.59*	61.97*	62.65*	56.21*	58.85*	62.56*
DIA	65.83 (5,0.3)	64.40	64.67	64.40	64.24	62.29	65.69	63.55	64.78	63.13	63.96	64.90
VEH	70.49 (3,0.3)	72.28	66.24	71.71	65.89	64.60	58.10*	70.16	69.46	68.90	70.81	69.71
SOB	93.11 (9,0.1)	83.04*	84.89*	87.91	91.88	86.57*	78.57*	83.96*	80.34*	87.50	86.30*	88.07
PAR	87.51 (9,0.7)	84.74	87.22	88.12	85.12	86.99	86.42	87.27	87.87	81.03	86.01	90.11
WPB	76.03 (7,0.1)	73.86	74.24	74.75	74.18	75.72	75.72	73.67	75.61	73.67	74.94	76.48
ION	86.83 (1,0.1)	86.13	87.78	87.61	85.87	84.83	85.35	87.83	83.26	84.35	84.52	89.00
CEL	70.80 (9,0.1)	64.26	64.79	64.31	65.67	70.82	70.82	65.10	71.24	70.82	70.89	70.78
NAS	99.64 (1,0.3)	99.56	99.58	89.41*	99.58	88.54*	85.38*	83.89*	99.48	99.59	99.53	99.59
STA	91.96 (7,0.1)	92.49	93.15	91.18	93.35	91.23	89.62	92.66	92.36	93.15	89.47	88.69
CM1	87.80 (9,0.1)	83.12*	84.49	86.37	85.91	87.74	87.30	82.77*	86.52	86.11	87.01	87.42
PC1	91.93 (9,0.1)	90.54	90.83	90.49	91.05	91.20	91.04	90.87	91.62	91.31	91.36	91.58
MC2	74.52 (7,0.7)	65.84*	68.69	67.72	63.94*	66.69*	72.14	61.82*	68.88	68.06	72.28	66.90
KC3	84.40 (3,0.1)	80.00	79.70	79.45	78.75	77.95*	83.00	77.65*	83.20	79.55	80.50	79.55
SPE	77.68 (1,0.1)	75.78	76.68	76.88	73.80	76.09	77.54	75.76	74.75	74.95	76.49	74.99
LIB	62.03 (7,0.7)	69.36v	62.17	58.42	63.03	55.95*	62.53	67.83	66.75	66.06	40.83*	28.00*
BAK	96.77 (3,0.1)	95.98*	96.69	96.52*	96.77	96.60*	96.77	96.09*	96.38*	96.27*	96.73*	96.77
LEA_N	60.97 (1,0.9)	60.15	56.68	60.12	60.32	26.71*	60.53	60.44	52.68*	58.29	53.47*	53.18*
DIA_N	65.25 (1,0.3)	64.07	64.87	64.07	64.52	62.32	65.09	63.58	64.34	63.86	64.26	61.80
VEH_N	67.35 (5,0.3)	62.23	66.15	68.71	69.41	56.44*	61.33*	70.22	70.39	70.47	58.59*	63.83
SOB_N	93.11 (3,0.1)	83.04*	84.89*	80.64*	87.54	86.57*	83.38*	83.96*	82.98	82.46*	76.39*	88.07
PAR_N	88.14 (7,0.1)	85.24	87.08	88.54	85.40	88.22	85.72	88.21	87.16	84.07	86.43	88.92
D1S	98.80 (1,0.1)	98.40	96.20	98.80	97.90	97.30	98.80	98.80	83.70*	97.80	97.30	97.40
D3S	95.70 (1,0.5)	93.80	91.40	94.90	94.20	94.40	92.90	94.20	90.10	93.20	91.60	92.70
D5S	94.70 (3,0.1)	90.90	79.50*	79.50*	73.20*	63.00*	78.70*	87.40*	91.20	76.40*	73.10*	92.80
COL	86.95 (1,0.1)	81.95	91.71v	76.50	77.86*	75.81*	81.64	62.60*	85.45	79.12	82.45	80.93
PRO	93.39 (5,0.3)	79.75*	93.14	90.86	91.03	85.13*	82.73*	89.06	80.84	87.95	89.64	88.75
ALL	93.57 (5,0.1)	80.73*	92.96	81.43*	96.52	89.07	92.23	94 21	82.95*	86.96	88.05	92.43
GEN	59.90 (5,0.9)	54.76	52.16*	49.70*	51.78*	52.70*	53.66	51.06*	51.88*	51.56*	52.78*	55.00
WAR	79.86 (9,0.1)	78.95	74.62	70.86*	75.15	77.41	66.90*	81.14	81.19	79.38*	60.33*	71.19*
Summary	(v//*)	1/22/8	1/26/4	0/24/7	0/26/5	0/17/14	0/22/9	0/20/11	0/22/9	0/25/6	0/21/10	0/27/4
AVE_ACC	82.72	79.16	80.05	78.98	79.48	75.59	78.26	78.85	79.12	78.90	77.44	78.91
AVE_RANK	2.52	7.35	6.11	6.58	6.44	7.94	6.53	7.02	6.58	7.76	7.37	5.77
#BEST	19	2	2	2	3	0	2	1	2	1	0	5

alternative methods on most of the 31 datasets. Only for three datasets, i.e., *VEH*, *D3S*, *D5S*, the reduct size obtained by LRM-RPFRFS is the largest among all methods. Although in some cases, the reduct size of LRM-RPFRFS is not the smallest, the classification performance of the reduct generated by LRM-RPFRFS is superior.

2) *Potential of Reduct:* It is important to ascertain whether the small reducts generated by LRM-RPFRFS retains sufficient information of the original datasets to entail high discriminating ability. A systematic comparison has been made, regarding the classification accuracy based on the use of raw datasets and reduced datasets returned by LRM-RPFRFS and 10 comparison approaches. The experimental results obtained by four popular classifiers including: J48 [47], Bagging [48], Jrip [49], and Part [50], which are shown in Tables VI–IX, respectively, where the average classification accuracies are obtained using the averaged results of 10-FCV with the best results per dataset underlined. Within each of these tables, the average accuracies

TABLE VII
BAGGING CLASSIFICATION ACCURACY (%) WITH REDUCTS RETURNED BY DIFFERENT FS METHODS

Dataset	**LRM-RPFRFS** (k,λ)	RAW-DATA	FRFG	GBFG	EL-TSFRFS	FGS-RFRAS	FRFM	HARCM	GSA	GWO	PCC	RELIEF
PAB	98.02 (5,0.3)	97.24	97.24	97.24	96.04	96.59	96.07	96.56	92.78	92.35	96.16	94.60
LEA	68.38 (1,0.9)	69.32	63.38	69.12	69.88	31.41*	66.65	70.03	70.29	62.89	62.06	66.12
DIA	67.36 (5,0.9)	66.96	65.65	66.96	66.98	61.31*	66.32	66.82	69.52	66.96	69.35	69.21
VEH	73.07 (9,0.1)	73.08	69.77	70.85	68.51	65.92*	59.68*	72.42	71.89	71.48	71.22	72.08
SOB	89.75 (1,0.3)	85.75	87.79	88.70	87.82	88.30	81.46*	85.64	81.20*	89.36	88.50	89.07
PAR	89.91 (9,0.7)	87.80	88.03	88.31	87.47	88.71	87.73	87.52	88.68	84.55	86.75	89.71
WPB	78.08 (1,0.3)	77.74	78.24	79.23	77.38	74.79	76.99	78.08	75.32	78.50	76.47	75.69
ION	88.70 (3,0.3)	89.09	89.78	89.78	85.78	86.00	87.04	87.52	85.13	86.65	87.22	90.13
CEL	70.34 (9,0.1)	69.91	70.08	69.94	69.88	70.81	70.02	69.98	69.05	69.59	69.13	68.72
NAS	99.56 (1,0.3)	99.52	99.49	90.16*	99.49	85.99*	85.62*	83.99*	99.50	99.53	99.47	99.56
STA	92.63 (9,0.7)	94.10	93.81	92.26	94.45	92.65	91.09	94.28	92.92	94.24	90.76	90.26
CM1	87.77 (9,0.1)	86.95	87.07	87.39	87.31	87.51	87.22	87.04	87.42	87.31	87.10	87.65
PC1	92.03 (9,0.1)	91.82	91.93	91.39	91.54	91.12	91.91	91.96	91.95	91.53	91.10	91.51
MC2	74.05 (9,0.5)	69.71	71.83	71.24	69.08	70.02	70.52	67.24*	70.69	70.47	71.46	70.56
KC3	83.05 (3,0.1)	81.00	81.95	80.70	80.90	79.85	83.25	80.50	82.65	81.00	80.40	81.00
SPE	80.65 (7,0.1)	79.95	78.63	78.70	78.67	79.17	80.49	80.59	80.47	80.10	80.02	79.93
LIB	65.75 (5,0.1)	70.89	65.53	62.00	66.86	60.39	69.56	71.29	69.42	69.58	41.39*	29.58*
BAK	96.77 (1,0.1)	96.08	96.62	96.76	96.60	96.81	96.68	96.71	96.61	96.68	96.73	96.73
LEA_N	67.85 (1,0.9)	67.06	63.47	63.62	64.44	30.91*	66.76	67.65	59.21*	63.88	60.47*	56.91*
DIA_N	67.03 (1,0.3)	66.69	66.41	66.69	67.31	64.66	66.93	66.93	67.49	67.65	68.17	66.56
VEH_N	69.94 (5,0.3)	64.95	69.67	69.90	70.18	58.64*	64.20	72.40	72.63	72.24	59.53*	63.92*
SOB_N	88.34 (3,0.3)	85.75	87.79	82.71	87.79	88.30	79.55*	85.64	82.43	82.43	78.16*	88.07
PAR_N	90.26 (7,0.1)	87.36	88.44	88.62	87.93	88.54	87.41	87.67	88.03	87.11	86.90	89.14
D1S	100.00 (1,0.1)	100.00	100.00	100.00	100.00	100.00	100.00	100.00	83.60*	99.90	100.00	100.00
D3S	96.40 (5,0.1)	94.70	95.20	96.20	93.30	94.40	93.90	94.20	93.00	95.00	95.40	95.90
D5S	94.10 (7,0.3)	92.30	82.40*	82.40*	73.20*	71.60*	84.40*	89.70	92.10	82.60*	86.50*	92.70
COL	83.64 (5,0.7)	76.12*	88.90	75.95*	77.86*	73.89*	78.55	62.90*	78.31*	73.83*	82.45*	80.93*
PRO	93.39 (5,0.3)	89.30	91.43	88.64	89.62	88.84	85.01*	89.54	89.80	90.39	91.01	90.51
ALL	93.16 (5,0.1)	92.93	93.93	84.11	94.02	92.11	92.59	94.43	92.82	92.38	92.46	92.38
GEN	57.02 (5,0.9)	54.88	54.02	53.66	54.32	54.92	51.86	51.02	55.28	56.06	55.04	56.56
WAR	81.14 (9,0.1)	91.95v	76.73	75.50	77.76*	91.19	65.86*	90.10	80.43	80.90	64.67*	71.38*
Summary	(v//*)	1/29/1	0/30/1	0/28/3	0/28/3	0/23/8	0/24/7	0/28/3	0/27/4	0/29/2	0/24/7	0/26/5
AVE_ACC	83.17	82.29	81.78	80.60	81.04	77.59	79.53	81.30	80.99	81.20	79.55	80.23
AVE_RANK	2.69	6.16	6.24	6.77	7.13	8.08	7.76	6.24	6.40	6.71	7.61	6.10
#BEST	16	4	2	2	2	2	2	2	3	0	2	3

TABLE VIII
JRIP CLASSIFICATION ACCURACY (%) WITH REDUCTS RETURNED BY DIFFERENT FS METHODS

Dataset	**LRM-RPFRFS** (k,λ)	RAW-DATA	FRFG	GBFG	EL-TSFRFS	FGS-RFRAS	FRFM	HARCM	GSA	GWO	PCC	RELIEF
PAB	97.72 (5,0.3)	96.96	96.96	96.96	96.65	96.45	96.08	96.46	92.76	93.63	95.49	94.09
LEA	53.32 (1,0.9)	51.21	45.94	52.71	51.53	16.74*	46.35*	52.97	52.68	48.26	49.47	49.12
DIA	64.43 (3,0.1)	63.41	62.75	63.41	63.25	61.79*	64.22	63.41	64.29	62.20	63.62	64.32
VEH	67.87 (3,0.3)	68.46	64.10	66.37	60.96*	57.54*	52.15*	68.61	67.91	67.86	66.16	64.95
SOB	91.34 (9,0.1)	83.36*	84.91*	86.66	88.38	84.37*	77.48*	83.96*	79.87*	88.45	88.80	89.48
PAR	90.58 (9,0.7)	87.89	86.87	88.47	87.37	88.67	85.37	86.64	85.77	80.63*	85.93	89.86
WPB	77.84 (1,0.1)	72.84	76.15	75.74	77.11	74.65	74.58	75.15	74.63	75.26	74.73	74.73
ION	87.91 (3,0.3)	87.09	87.91	87.96	87.13	84.22	87.04	87.52	83.65	83.65	83.13	87.17
CEL	70.90 (1,0.1)	70.56	70.47	70.52	70.44	70.82	70.68	70.66	70.69	70.78	70.65	70.56
NAS	99.59 (1,0.3)	99.61	99.46	88.59*	99.52	85.70*	84.36*	83.80*	99.46	99.51	99.59	99.56
STA	92.51 (9,0.9)	92.57	92.74	91.48	92.89	92.34	90.26*	92.97	91.77	92.77	90.03	89.27
CM1	87.74 (9,0.1)	85.56	85.71	86.90	85.65	87.51	86.83	85.53	87.07	86.73	86.69	87.57
PC1	91.88 (9,0.1)	91.39	91.13	91.00	91.27	90.97	91.26	91.32	91.59	91.40	91.85	91.28
MC2	73.06 (9,0.5)	68.42	68.33	67.20	64.04	71.39	72.92	62.87*	72.21	70.28	68.60	67.85
KC3	84.60 (3,0.1)	81.40	83.10	79.45	82.60	79.25	83.25	80.50	83.75	80.80	81.80	80.80
SPE	78.25 (7,0.1)	76.93	76.18	74.46	76.97	77.96	80.95	79.86	78.62	78.73	79.48	76.45
LIB	50.86 (5,0.1)	55.69	49.06	44.89	50.05	43.42*	52.81	53.25	53.67	53.72	27.75*	13.03*
BAK	96.77 (1,0.1)	96.51	96.45	96.43	96.64	96.46	96.87	96.42	96.51	96.58	96.37	96.76
LEA_N	53.65 (1,0.9)	52.74	46.09*	53.09	52.56	15.24*	49.00	52.38	42.91*	49.76	48.94	44.94*
DIA_N	66.44 (1,0.3)	63.27	62.57	63.27	63.63	62.35	64.24	62.74	63.20	63.70	65.04	61.48
VEH_N	63.12 (5,0.3)	60.53	64.60	65.08	65.53	49.80*	56.79*	68.48	66.32	65.89	50.88*	58.27*
SOB_N	91.34 (3,0.1)	83.36*	84.91*	79.45*	89.46	84.37*	78.64*	83.96*	82.55*	78.93*	78.36*	89.48
PAR_N	90.99 (7,0.1)	87.81	87.06	87.58	87.56	88.27	85.72	88.12	85.26*	83.94*	86.10*	88.85
D1S	95.70 (1,0.1)	93.40	95.20	95.90	96.00	95.40	95.20	94.90	80.30*	95.40	95.50	96.00
D3S	93.20 (3,0.3)	90.20	90.10	90.60	89.10	92.70	90.90	93.20	85.60	90.50	84.40*	92.80
D5S	85.50 (1,0.3)	82.00	75.50*	75.50*	69.40*	59.30*	73.90*	83.00	84.50	76.40*	73.60*	83.80
COL	84.67 (7,0.1)	76.33*	90.10v	77.57*	79.48*	75.84*	80.33*	60.07*	70.74*	70.02*	83.90*	81.93*
PRO	93.11 (5,0.3)	84.44*	93.38	88.96	90.75	86.78*	83.25*	89.80	86.00*	86.84*	89.75	90.29
ALL	93.86 (5,0.1)	86.55*	93.86	80.61*	95.29	88.27	91.63	94.34	88.11	88.20	89.98	91.46
GEN	57.54 (5,0.9)	53.24	54.44	53.10	54.14	53.64	56.44	54.07	53.56	54.98	56.30	54.60
WAR	71.90 (9,0.1)	77.67	65.67	64.67	63.71*	77.10	51.43*	77.90v	72.62	72.71	52.57*	63.19*
Summary	(v//*)	0/26/5	1/26/4	0/26/5	0/27/4	0/19/12	0/20/11	1/25/5	0/14/7	0/25/6	0/23/8	0/26/5
AVE_ACC	80.59	78.11	78.12	76.92	78.03	73.85	75.84	77.90	77.05	77.37	75.98	76.90
AVE_RANK	2.32	6.76	6.92	7.13	6.18	8.05	7.37	6.05	7.08	6.63	7.32	6.19
#BEST	18	2	2	1	2	0	2	4	0	0	0	1

and ranks of the 31 datasets are displayed in the "AVE_ACC" and "AVE_RANK" row, respectively, and the number of the best performances is summarized in the "#BEST" row.

1) It can be seen that in conjunction with the use of either J48, Bagging, Jrip, or Part, across all datasets (the number of samples ranges from 62 to 12 000 and the number of features ranges from 11 to 7130), the classification performance achieved by the proposed method is superior to those attainable by the others for a great majority of cases (on 19 datasets for J48; 16 datasets for Bagging; 18 datasets for Jrip; 16 datasets for Part). In addition, the average accuracy rates and the average ranks gained by LRM-RPFRFS are the best for different classifiers. Importantly, such an outstanding performance is achieved with the utilization of the smaller feature subsets, forming a sharp contrast with the rest. For those datasets, where the utilization of features returned by LRM-RPFRFS does not result in the highest accuracy, the performance remains

TABLE IX
PART CLASSIFICATION ACCURACY (%) WITH REDUCTS RETURNED BY DIFFERENT FS METHODS

Dataset	**LRM-RPFRFS** (k,λ)	RAW-DATA	FRFG	GBFG	EL-TSFRFS	FGS-RFRAS	FRFM	HARCM	GSA	GWO	PCC	RELIEF
PAB	97.70 (5,0.3)	96.93	96.93	96.93	94.77	96.40	95.83	96.39	92.84	93.64	95.91	93.86
LEA	62.06 (1,0.9)	61.85	56.71	61.06	62.18	27.71*	59.85	62.21	62.50	53.91*	57.71	62.06
DIA	65.02 (7,0.9)	64.37	63.54	64.37	64.93	61.79	65.00	63.65	64.92	63.31	62.97	64.76
VEH	71.84 (9,0.1)	72.21	66.56*	69.49	66.29	62.87*	58.05*	70.89	70.26	69.25	71.27	69.31
SOB	93.54 (9,0.1)	86.42*	85.55*	87.25	92.14	84.79*	81.43*	84.39*	81.73*	90.29	85.55	86.02
PAR	86.57 (9,0.7)	84.94	86.25	84.89	86.13	86.64	85.15	86.23	87.20	80.86	84.12	87.66
WPB	74.65 (7,0.1)	74.82	74.96	76.19	73.68	75.98	75.72	73.25	75.86	74.69	74.18	76.53
ION	87.91 (3,0.3)	87.39	88.61	89.61	85.87	84.09	87.09	88.17	83.39	85.30	85.09	88.26
CEL	70.75 (9,0.1)	70.24	70.24	70.43	70.19	70.81	70.70	70.12	71.08	70.92	70.86	70.71
NAS	99.64 (1,0.3)	99.60	99.55	88.32*	99.56	85.51*	84.21*	83.89*	99.45	99.56	99.59	99.62
STA	91.65 (1,0.1)	91.41	92.55	90.81	92.55	91.24	90.26	92.38	91.18	92.51	89.75	88.59
CM1	87.45 (9,0.1)	84.25	86.72	86.98	86.32	87.34	87.36	85.38	87.04	86.86	86.81	87.42
PC1	91.99 (9,0.1)	90.30	90.26	90.31	90.15	90.86	91.69	90.45	91.73	90.62	91.38	91.70
MC2	72.68 (9,0.5)	67.97	68.83	69.05	64.47*	65.16*	70.96	65.19*	70.46	68.72	71.68	66.33
KC3	84.60 (3,0.1)	79.40	80.75	80.45	80.85	80.05	82.55	80.65	82.20	80.90	81.10	80.90
SPE	77.13 (1,0.1)	75.62	76.65	77.09	74.75	77.14	74.85	77.53	75.78	75.72	76.88	75.82
LIB	61.78 (5,0.1)	68.41	62.61	58.42	61.97	56.69	65.09	69.19	64.96	64.67	41.14*	27.06*
BAK	96.77 (1,0.1)	96.36	96.74	96.65	96.75	96.76	96.84	96.59	96.57	96.64	96.74	96.69
LEA_N	61.24 (1,0.9)	60.62	56.15	60.38	60.35	26.76*	59.68	60.79	55.26	59.06	53.71*	51.56*
DIA_N	64.79 (1,0.3)	64.17	63.46	64.17	64.88	62.24	63.36	62.77	63.68	63.99	62.81	62.11
VEH_N	67.77 (5,0.3)	67.20	66.57	69.03	70.08	54.77*	61.65*	70.85	71.29	71.03	56.06*	63.21
SOB_N	93.54 (3,0.1)	86.41*	85.55*	77.16*	90.91	84.79*	80.57*	84.39*	82.30*	81.39*	75.43*	86.02
PAR_N	87.67 (7,0.1)	86.12	87.16	84.86	85.96	84.77	86.69	86.19	87.12	83.89	84.23	86.62
D1S	98.80 (1,0.1)	98.40	96.50	98.70	98.20	98.30	98.00	98.00	85.10*	96.90	97.20	97.10
D3S	96.60 (1,0.5)	95.40	93.50	95.00	94.30	92.40	92.90	94.20	89.40*	94.40	92.20	94.10
D5S	94.50 (3,0.1)	90.60	80.40*	80.40*	70.70*	62.50*	77.50*	88.60*	89.70	79.90*	78.70*	93.60
COL	84.31 (7,0.1)	82.21	91.71	78.74	77.90*	76.67*	81.48	61.39*	85.62	78.86*	80.02*	79.33*
PRO	93.39 (5,0.3)	80.67*	93.27	91.20	89.92	82.07*	82.53*	88.48	82.69*	86.24	89.66	88.65
ALL	93.57 (5,0.1)	80.73*	92.82	80.71*	96.52	89.07	90.95	94.21	82.95*	87.21	89.46	92.55
GEN	59.90 (5,0.9)	54.32*	52.32*	49.86*	51.98*	52.04*	52.64*	51.06*	52.16*	50.90*	52.82*	54.82*
WAR	81.76 (9,0.1)	78.19	76.48	72.52	74.24	84.71	64.71*	83.71	80.71	80.62	62.14*	69.00*
Summary	(v//*)	0/26/5	1/25/5	0/26/5	0/27/4	0/19/12	0/22/9	0/24/7	0/24/7	0/26/5	0/22/9	0/24/7
AVE_ACC	82.31	79.92	79.80	78.74	79.04	75.22	77.91	79.40	79.26	79.12	77.33	78.45
AVE_RANK	2.56	6.31	6.29	6.63	6.68	8.16	7.02	6.73	6.19	7.42	7.79	6.19
#BEST	16	1	2	1	3	1	1	2	3	0	0	2

comparable to alternative methods while predominantly involving significantly fewer features. Particularly, the averaged differences between the accuracies obtained by LRM-RPFRFS and the highest accuracies are approximately 2.23% (on 12 datasets for J48), 2.31% (on 15 datasets for Bagging), 2.12% (on 13 datasets for Jrip), and 2.14% (on 15 datasets for Part), respectively. Furthermore, the reducts returned by LRM-RPFRFS does not lead to any weakest performance across all examined cases.

2) To investigate the antinoise ability of LRM-RPFRFS, the experiment is carried out on five noisy datasets (i.e., LEA_N, DIA_N, VEH_N, SOB_N, and PAR_N), artificially generated by adding 10% noise randomly. As can be seen from Tables VI–IX, LRM-RPFRFS obtains highest classification performance among 11 feature reducts obtained by 11 FS methods and the raw data in terms of accuracy on 3, 3, 4, and 3 datasets, respectively, by using four different classifiers. Moreover, for the remaining classification results, the gaps between the accuracy of LRM-RPFRFS and the best one are 3.12%, 0.78%; 1.14%, 2.69%; 5.36%; 3.42%, 0.21% for each classifier, where LRM-RPFRFS not only avoids the weakest performance, but also demonstrates results comparable to the best-performing methods. Therefore, the LRM-RPFRFS algorithm is robust to noisy information in data.

3) To investigate the adaptation of LRM-RPFRFS for different distributions, the experiment is carried out on three synthetic datasets (i.e., $D1S$, $D2S$, and $D3S$) containing four classes but each with 1, 3, and 5 standard deviation of the clusters (denoted as cluster_std), respectively (shown in Fig. 6). As can be seen from Tables VI–IX, LRM-RPFRFS can achieve higher classification accuracies than the others on the three datasets. In particular, it exhibits obvious superiorities on $D2S$ and $D3S$. Therefore, the LRM-RPFRFS algorithm is effectively applicable to the datasets with large class density difference.

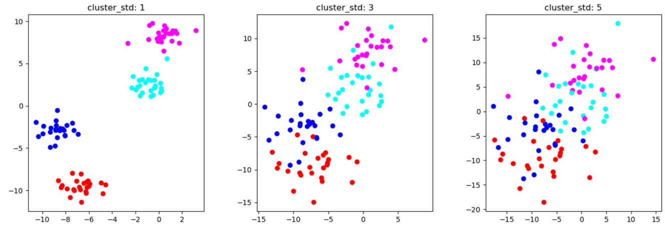
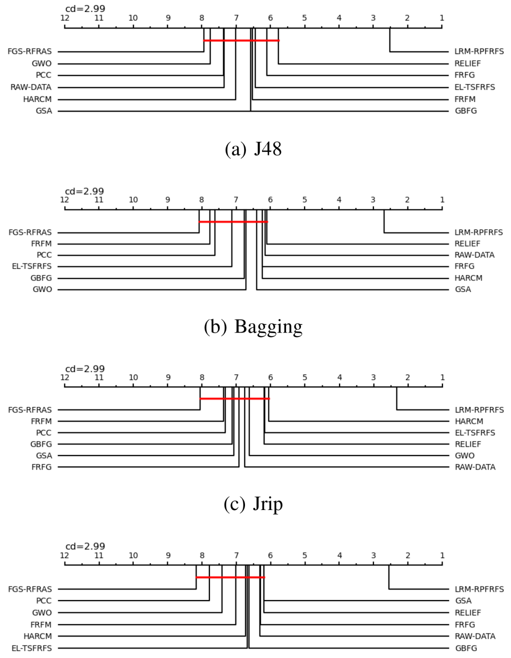


Fig. 6. Synthetic datasets with different distributions.

4) To investigate the performance of LRM-RPFRFS in practical applications, the experiment is carried out on 3 biological datasets (i.e., COL, PRO and ALL) and two face datasets (i.e., GEN and WAR). As can be seen from Tables VI–IX, although LRM-RPFRFS does not achieve the highest classification accuracy in most cases, it consistently obtains classification results comparable to the best ones while the comparison approaches only occasionally achieve better performance. Therefore, the LRM-RPFRFS algorithm has better generalization performance in the practical applications, such as, cancer diagnosis and face recognition.

Together, all of the above results illustrate that the LRM-RPFRFS algorithm entails an overall stronger performance, due to the following reasons.

1) Compared with the above six FRS-FS methods, LRM-RPFRFS constructs a more robust pseudo FRS model where the distribution-aware linear reconstruction relation is developed by fully considering the distribution information of samples and density information of classes



(a) J48

(b) Bagging

(c) Jrip

(d) Part

Fig. 7. Comparisons of the classification accuracy of eleven approaches against each other with the Nemenyi test using four classifiers. (a) J48. (b) Bagging. (c) Jrip. (d) Part.

to better fit different data distributions, and the pseudo-fuzzy lower approximation is calculated by determining the kNN granules in terms of linear reconstruction coefficients to empower the antinoise ability.

2) Compared with the above four state-of-the-art FS methods, LRM-RPFRFS is a filter strategy-based algorithm guided from the perspective of *RF*, *SRP,* and *DS*. However, existing filter strategy-based FS methods (e.g., PCC and RelieffF) primarily focuses on the relationships between features, neglecting the insight of meaningful information including direct relevance between conditional features and the decision feature, and redundancy between features. Wrapped strategy-based FS methods (e.g., GWO and GSA) are distinct from the filter strategy-based FS methods in that they usually select feature subset by evaluating the performance associated with the final predictive model, thereby seriously impacting generalization.

In conclusion, LRM-RPFRFS can achieve significant improvements in terms of effectiveness, generalization, and robustness. Particularly, on noisy datasets and synthetic datasets with different distributions, the classification accuracies of the LRM-RPFRFS algorithm demonstrates the outstanding robustness of utilizing the proposed robust LRM-k-PFRS model. Moreover, the systematic experiments have been carried out on both benchmark datasets and practical applications (including cancer diagnosis and face recognition), where LRM-RPFRFS can achieve better performance than the others on most cases. Therefore, these experimental results collectively demonstrate both the effectiveness and generalization.

3) Statistical Test: The statistical tests are conducted to validate the above experimental results.

T-Test: In order to prove that experimental results are not obtained accidentally, a paired T-test is conducted to provide statistical analysis of the classification accuracies on each dataset under the significance level of 0.05. For each dataset, the classification accuracy obtained by LRM-RPFRFS is the baseline reference for other methods in the tests. In Tables VI–IX, the summary of the statistical outcomes is displayed in the "Summary" row within each of these tables, where the count of the number of statistically better (v), equivalent (), or worse (*) cases for each method on all the datasets compared to LRM-RPFRFS is displayed. For example, in Table VI, (1/26/4) in the "FRFG" column indicates that this method performs better than LRM-RPFRFS on 1 datasets, performs equivalently to it on 26 datasets, and performs worse than it on 4 datasets. As can be seen from Tables VI–IX, the number of occurrences of "*" is much higher than that of "v." That is, the employment of the reduct returned by LRM-RPFRFS leads to statistically better or equal results as compared to the application of the reducts produced by the alternative methods in most cases.

Friedman Test: The Friedman test is performed to compare the performance among multiple methods in terms of the classification accuracy. Under the null hypothesis that "the multiple methods have the same performance," the Friedman statistics can be calculated by using the ranks of the classification accuracies on 31 dataset of Tables VI–IX. The Friedman statistics (and p-value) of the four classifiers including J48, Bagging, Jrip, and Part are 52.67 (2.06×10^{-7}), 50.47 (5.14×10^{-7}), 54.21 (1.07×10^{-7}), and 51.18 (3.82×10^{-7}), respectively. Moreover, corresponding critical difference (CD) value under $\alpha = 0.05$ is 2.99. Thus, the null hypothesis that "the multiple methods have the same performance" should be rejected with a 0.95 confidence interval. Further, the performance of these approaches is significantly different.

Nemenyi's Post-Hoc Test: The Nemenyi's post-hoc test is performed to distinguish the differences of the eleven FS methods using pairwise comparisons. The CD is 2.99 for a confidence level of $\alpha = 0.05$. The hypothesis that "the performance of the two approaches is the same" is rejected with the corresponding confidence if the difference between the average ranks shown in the "AVE_RANK" row in Tables VI–IX of the two approaches exceeds the CD. As can be seen from Fig. 7, the employment of either J48, Bagging, Jrip, or Part, the accuracies of LRM-RPFRFS is statistically better than those with the comparison approaches.

VI. Conclusion

This article has presented a novel FS approach, entitled LRM-RPFRFS, which is proposed from the perspective of *RF*, *SRP,* and *DS* to determine the final feature subset. Moreover, a robust pseudo FRS model called LRM-k-PFRS is proposed where the distribution-aware linear reconstruction relation is constructed by considering the insight of meaningful information (i.e., distribution information of samples and density information of classes) to enhance the robustness and the pseudofuzzy lower approximation is calculated based on kNN granules to empower the antinoise ability. Experimental results have demonstrated in general that LRM-RPFRFS can achieve significant improvements in terms of effectiveness, generalization, and robustness While promising, the work also opens up an avenue for further development. For instance, it would be useful to investigate how to construct a robust pseudo FRS model with a sparse LRM model to strengthen the interpretability and generalization.

References

[1] W. Li, H. Zhou, W. Xu, X. Wang, and W. Pedrycz, "Interval dominance-based feature selection for interval-valued ordered data," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 34, no. 10, pp. 6898–6912, Oct. 2023.

[2] W. Xu, M. Huang, Z. Jiang, and Y. Qian, "Graph-based unsupervised feature selection for interval-valued information system," *IEEE Trans. Neural Netw. Learn. Syst.*, 2023, to be published, doi: 10.1109/TNNLS.2023.3263684.

[3] D. Dubois and H. Prade, "Rough fuzzy sets and fuzzy rough sets," *Int. J. Gen. Syst.*, vol. 17, no. 2/3, pp. 191–209, 2007.

[4] D. Dubois and H. Prade, *Putting Rough Sets and Fuzzy Sets Together*. Berlin, Germany: Springer, 1992, pp. 203–232.

[5] L. Zadeh, "Fuzzy sets," *Inf. Control*, vol. 8, no. 3, pp. 338–353, 1965.

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

[7] R. Jensen and Q. Shen, "New approaches to fuzzy-rough feature selection," *IEEE Trans. Fuzzy Syst.*, vol. 17, no. 4, pp. 824–838, Aug. 2009.

[8] R. Jensen and Q. Shen, "Fuzzy rough attribute reduction with application to web categorization," *Fuzzy Sets Syst.*, vol. 141, no. 3, pp. 469–485, 2004.

[9] R. Jensen and Q. Shen, "Semantics-preserving dimensionality reduction: Rough and fuzzy-rough-based approaches," *IEEE Trans. Knowl. Data Eng.*, vol. 16, no. 12, pp. 1457–1471, Dec. 2004.

[10] R. Jensen and Q. Shen, "Fuzzy-rough data reduction with ant colony optimization," *Fuzzy Sets Syst.*, vol. 149, no. 1, pp. 5–20, 2005.

[11] R. Jensen and Q. Shen, "Fuzzy-rough sets assisted attribute selection," *IEEE Trans. Fuzzy Syst.*, vol. 15, no. 1, pp. 73–89, Feb. 2007.

[12] Q. Hu, D. Yu, Z. Xie, and J. Liu, "Fuzzy probabilistic approximation spaces and their information measures," *IEEE Trans. Fuzzy Syst.*, vol. 14, no. 2, pp. 191–201, Apr. 2006.

[13] Q. Hu, Z. Xie, and D. Yu, "Hybrid attribute reduction based on a novel fuzzy-rough model and information granulation," *Pattern Recognit.*, vol. 40, no. 12, pp. 3509–3521, 2007.

[14] Q. Hu, D. Yu, and Z. Xie, "Information-preserving hybrid data reduction based on fuzzy-rough techniques," *Pattern Recognit. Lett.*, vol. 27, no. 5, pp. 414–423, 2006.

[15] T. Yang, Y.-J. Li, Y. Qian, and F.-Y. Wang, "Consistent matrix: A feature selection framework for large-scale datasets," *IEEE Trans. Fuzzy Syst.*, vol. 31, no. 11, pp. 4024–4038, Nov. 2023.

[16] C. Wang, Y. Qian, W. Ding, and X. Fan, "Feature selection with fuzzy-rough minimum classification error criterion," *IEEE Trans. Fuzzy Syst.*, vol. 30, no. 8, pp. 2930–2942, Aug. 2022.

[17] R. Jensen, N. M. Parthaláin, and C. Cornells, "Feature grouping-based fuzzy-rough feature selection," in *Proc. IEEE Int. Conf. Fuzzy Syst.*, 2014, pp. 1488–1495.

[18] L. Zheng, F. Chao, N. M. Parthaláin, D. Zhang, and Q. Shen, "Feature grouping and selection: A graph-based approach," *Inf. Sci.*, vol. 546, pp. 1256–1272, 2021.

[19] J. Wan, H. Chen, T. Li, B. Sang, and Z. Yuan, "Feature grouping and selection with graph theory in robust fuzzy rough approximation space," *IEEE Trans. Fuzzy Syst.*, vol. 31, no. 1, pp. 213–225, Jan. 2023.

[20] Y. Qu, L. Qiu, C. Shang, and Q. Shen, "Exclusive lasso assisted two-stage fuzzy-rough feature selection," TechRxiv, 2023, doi: 10.36227/techrxiv.24243172.v1.

[21] J. Fernández Salido and S. Murakami, "Rough set analysis of a general type of fuzzy data using transitive aggregations of fuzzy similarity relations," *Fuzzy Sets Syst.*, vol. 139, no. 3, pp. 635–660, 2003.

[22] C. Cornelis, M. De Cock, and A. M. Radzikowska, *Vaguely Quantified Rough Sets*. Berlin Heidelberg, Germany: Springer, 2007, pp. 87–94.

[23] A. Mieszkowicz-Rolka and L. Rolka, "Variable precision fuzzy rough sets," *Trans. Rough Sets I*, vol. 3100, pp. 144–160, 2004.

[24] S. Zhao, E. C. C. Tsang, and D. Chen, "The model of fuzzy variable precision rough sets," in *2007 Int. Conf. Mach. Learn. Cybern.*, 2007, vol. 6, pp. 3057–3062.

[25] S. An, Q. Hu, W. Pedrycz, P. Zhu, and E. C. 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.

[26] N. Verbiest, C. Cornelis, and F. Herrera, "OWA-FRPS: A prototype selection method based on ordered weighted average fuzzy rough set theory," in *Proc. Conf. Rough Sets, Fuzzy Sets, Data Mining Granular Comput.*, 2013, vol. 8170, pp. 180–190.

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

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

[29] C. Wang, Y. Huang, M. Shao, and X. Fan, "Fuzzy rough set-based attribute reduction using distance measures," *Knowl. Based Syst.*, vol. 164, pp. 205–212, 2019.

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

[31] Q. Hu, L. Zhang, D. Chen, W. Pedrycz, and D. Yu, "Gaussian kernel based fuzzy rough sets: Model, uncertainty measures and applications," *Int. J. Approx. Reasoning*, vol. 51, no. 4, pp. 453–471, 2010.

[32] Q. Hu, L. Zhang, Y. Zhou, and W. Pedrycz, "Large-scale multimodality attribute reduction with multi-kernel fuzzy rough sets," *IEEE Trans. Fuzzy Syst.*, vol. 26, no. 1, pp. 226–238, Feb. 2018.

[33] C. Wang, Y. Huang, W. Ding, and Z. Cao, "Attribute reduction with fuzzy rough self-information measures," *Inf. Sci.*, vol. 549, pp. 68–86, 2021.

[34] Y. Qu, R. Li, A. Deng, C. Shang, and Q. Shen, "Non-unique decision differential entropy-based feature selection," *Neurocomputing*, vol. 393, pp. 187–193, 2020.

[35] X. Zhang, C. Mei, D. Chen, Y. Yang, and J. Li, "Active incremental feature selection using a fuzzy-rough-set-based information entropy," *IEEE Trans. Fuzzy Syst.*, vol. 28, no. 5, pp. 901–915, May 2020.

[36] B. Sang, W. Xu, H. Chen, and T. Li, "Active antinoise fuzzy dominance rough feature selection using adaptive k-nearest neighbors," *IEEE Trans. Fuzzy Syst.*, vol. 31, no. 11, pp. 3944–3958, Nov. 2023.

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

[38] S. An, M. Zhang, C. Wang, and W. Ding, "Robust fuzzy rough approximations with kNN granules for semi-supervised feature selection," *Fuzzy Sets Syst.*, vol. 461, 2023, Art. no. 108476.

[39] W. Qian, F. Xu, J. Huang, and J. Qian, "A novel granular ball computing-based fuzzy rough set for feature selection in label distribution learning," *Knowl.-Based Syst.*, vol. 278, 2023, Art. no. 110898.

[40] J. Xu, X. Meng, K. Qu, Y. Sun, and Q. Hou, "Feature selection using relative dependency complement mutual information in fitting fuzzy rough set model," *Appl. Intell.*, vol. 53, pp. 18239–18262, 2023.

[41] W. Xu, K. Yuan, W. Li, and W. Ding, "An emerging fuzzy feature selection method using composite entropy-based uncertainty measure and data distribution," *IEEE Trans. Emerg. Topics Comput. Intell.*, vol. 7, no. 1, pp. 76–88, Feb. 2023.

[42] J. Zhang and J. Yang, "Linear reconstruction measure steered nearest neighbor classification framework," *Pattern Recognit.*, vol. 47, no. 4, pp. 1709–1720, 2014.

[43] R. Guha, M. Ghosh, A. Chakrabarti, R. Sarkar, and S. Mirjalili, "Introducing clustering based population in binary gravitational search algorithm for feature selection," *Appl. Soft Comput.*, vol. 93, 2020, Art. no. 106341.

[44] S. Dhargupta, M. Ghosh, S. Mirjalili, and R. Sarkar, "Selective opposition based grey wolf optimization," *Expert Syst. Appl.*, vol. 151, 2020, Art. no. 113389.

[45] R. J. Urbanowicz, M. Meeker, W. La Cava, R. S. Olson, and J. H. Moore, "Relief-based feature selection: Introduction and review," *J. Biomed. Informat.*, vol. 85, pp. 189–203, 2018.

[46] R. Guha, K. K. Ghosh, S. Bhowmik, and R. Sarkar, "Mutually informed correlation coefficient (MICC) – A new filter based feature selection method," in *2020 IEEE Calcutta Conf.*, 2020, pp. 54–58.

[47] J. R. Quinlan, *C4. 5: Programs for Machine Learning*. Amsterdam, The Netherlands: Elsevier, 2014.

[48] L. Breiman, "Bagging predictors," *Mach. Learn.*, vol. 24, pp. 123–140, 1996.

[49] W. W. Cohen, "Fast effective rule induction," in *Proc. 12th Int. Conf. Mach. Learn.*, Elsevier, 1995, pp. 115–123.

[50] E. Frank and I. H. Witten, "Generating accurate rule sets without global optimization," 1998.