



Kernelized Fuzzy Rough Sets-Based Three-Way Feature Selection

Xingchen Liu¹, Liuxin Wang¹, Linchao Pan¹, and Can Gao^{1,2,3}(✉)

¹ College of Computer Science and Software Engineering, Shenzhen University,
Shenzhen 518060, Guangdong, People's Republic of China
2005gaocan@163.com

² Guangdong Key Laboratory of Intelligent Information Processing, Shenzhen
518060, Guangdong, People's Republic of China

³ SZU Branch, Shenzhen Institute of Artificial Intelligence and Robotics for Society,
Shenzhen 518060, Guangdong, People's Republic of China

Abstract. Feature selection is the process of selecting important features from a dataset. The feature subset formed by important features represents the features of the entire dataset to reduce the complexity of subsequent computations. In recent years, feature selection methods based on rough set theory have been continuously developed, and the approximate quality of kernelized fuzzy rough sets is a better method for evaluating features. However, the heuristic greedy strategy adopted by traditional methods is difficult to guarantee the quality of feature subsets. Based on the idea of three-way decision, this paper proposes fuzzy dependency-based three-way feature selection method. We expand the three potential feature subsets through a differentiated approach and reduce the redundancy among them. Ensemble learning is performed on the three feature subsets to improve the classification performance. The experimental results show that compared with the traditional greedy feature selection method, the proposed feature selection method produces better classification performance, which demonstrates its effectiveness.

Keywords: Kernelized fuzzy rough sets · Three-way decision · Feature Selection

1 Introduction

With the continuous growing of the scale of datasets in recent years, a given learning problem and classification task contains a large number of features, and these features are often irrelevant or redundant. Such features will lead to the problems of high computational complexity, weak generalization ability and poor interpretability. Feature selection is an effective technique to alleviate these problems. It reserves highly correlated features and removes redundant and irrelevant features to find the optimal feature subset, and thus improving the performance of models [1]. Therefore, it becomes one of important preprocesses for machine learning, data mining, and pattern recognition etc. [2].

Rough set theories [3] provides an effective method for modeling vague, uncertain, or imprecise data. It uses a pair of exact sets (upper approximation and lower approximation) to describe the uncertainty within the data set.

The attribute reduction methods in this theory remove redundant features in the dataset while maintaining the correlation between feature subsets and decision classes, which coincides with the purpose of feature selection [4]. Based on the rough set theory, the correlation information within data is found without the need for supplementary information, and the number of attributes contained in the data set is reduced, which realizes the feature selection based on rough sets. Presently, there are many extensions in rough sets theory, such as probabilistic rough sets [5], neighborhood rough sets [6], and fuzzy rough sets [7, 8], etc.

To deal with the information loss caused by discretizing data, Dubois and Prade defined fuzzy rough sets [7, 8] by introducing fuzzy membership functions and extending the membership of elements to $[0,1]$, which provides a high degree of flexibility when dealing with continuous data in fields such as medicine, industry and finance, and can effectively model the ambiguity and uncertainty that exist in the data. On the premise of reserving the advantages of rough sets-based set feature selection for processing high dimensional data, fuzzy rough sets-based feature selection is realized by the fuzzy division of each feature by fuzzy set theory [9]. This method can effectively reduce discrete or continuous noise data, without the cost of adding extra information.

Aiming at the linear inseparability of the data obtained in the real world, that is, there is no dividing hyperplane that can correctly classify the training samples, we use the kernel methods to map the samples from the original space to a higher-dimensional feature space to solve the problem, which makes the samples linearly separable in this feature space. And for a limited-dimensional sample space, there must be a high-dimensional feature space that makes the mapped samples linearly separable. Hu integrated kernel functions with fuzzy rough sets and proposed the model of kernelized fuzzy rough sets, which forms a bridge between kernel machines and rough set-based data analysis [10]. Some generalized feature evaluation functions and attribute reduction algorithms based on the proposed model are shown and the effectiveness of the proposed technique is validated.

The three-way decision [11] theory extends the traditional two-way decision theory and is a decision-making method that conforms to human thinking. In two-way decision, the judgment of objects only stays in two results: acceptance and rejection. However, in practice, people often delay the judgment and decide on objects that they are confident to accept or reject instead of making decisions immediately for uncertain or incomplete information. The three-way decision divides objects into three domains (positive domain, negative domain and boundary domain) according to the decision-making state value by defining the decision function and the threshold of the domain, then constructs the corresponding three-way decision rules [12].

In this paper, we introduce a feature selection method based on kernelized fuzzy rough sets and three-way decision. When constructing feature subsets, how to maintain the maximum relevance for the decision class while minimizing the redundancy between feature subsets is a key issue in feature selection. The three-way strategy we employ is to construct three differentiated subsets of features and to expand the features in the subset from different perspectives. The feature subsets constructed by this strategy tend to be smaller than those constructed

by traditional methods. Dependency is an important metric in rough set theory to measure the relevance of features with respect to decision classes. Thus, the dependency gained from new features is used as a reference for our feature subset expansion strategy. Finally, we consider the idea of ensemble learning to construct a multiple feature subsets-based co-classification model.

The rest of the paper is organized as follows. Section 2 presents the notions and properties of the fuzzy rough set model and feature selection. Section 3 shows the three-way attribute reduction algorithm. Experimental analysis is given in Sect. 4. Conclusions come in Sect. 5.

2 Preliminaries

In this section, we will first give some basic definitions, and then review the related work of rough sets, fuzzy rough sets, and kernelized fuzzy rough sets.

2.1 The Notations

Let $I = (\mathbb{U}, \mathbb{A})$ be an information system, where $\mathbb{U} = \{x_1, x_2, \dots, x_n\}$ is a nonempty set of finite objects called the universe of discourse and \mathbb{A} is a nonempty finite set of attributes $a : \mathbb{U} \rightarrow V_a$ for every $a \in \mathbb{A}$. For decision systems, $\mathbb{A} = (\mathbb{C}, \mathbb{D})$, where \mathbb{C} is the set of input features and \mathbb{D} is the set of output features. Additionally $a(x), a \in \mathbb{C}, x \in \mathbb{U}$ represents the value of the object x under the attribute a .

2.2 Rough Sets

With any $P \subseteq \mathbb{A}$ there is an associated equivalence relation $IND(P)$:

$$IND(P) = \{(x, y) \in \mathbb{U}^2 \mid \forall a \in P, a(x) = a(y)\} \tag{1}$$

An associated equivalence relation is reflexive, symmetric and transitive. The family of all equivalence classes of $IND(P)$ are denoted by $\mathbb{U}/IND(P)$ or \mathbb{U}/P for short, which is simply the set of equivalence classes generated by $IND(P)$:

$$\mathbb{U}/IND(P) = \otimes\{\mathbb{U}/IND(\{a\}) \mid a \in P\} \tag{2}$$

where

$$A \otimes B = \{X \cap Y \mid X \in A, Y \in B, X \cap Y \neq \emptyset\} \tag{3}$$

The equivalence classes of the indiscernibility relation with respect to P are denoted $[x]_P, x \in \mathbb{U}$. Let $X \subseteq \mathbb{U}$, X can be approximated using only the information contained within P by constructing the P -lower and P -upper approximations of the classical crisp set X :

$$\underline{P}X = \{x \mid [x]_P \subseteq X\} \tag{4}$$

$$\bar{P}X = \{x \mid [x]_P \cap X \neq \emptyset\} \tag{5}$$

Let P and Q be subsets of condition attributes and decision attributes, respectively, then according to the upper approximation and the lower approximation, then the positive, negative, and boundary regions are defined as:

$$POS_P(Q) = \bigcup_{X \in \mathbb{U}/Q} \underline{P}X \tag{6}$$

$$NEG_P(Q) = \mathbb{U} - \bigcup_{X \in \mathbb{U}/Q} \bar{P}X \tag{7}$$

$$BND_P(Q) = \bigcup_{X \in \mathbb{U}/Q} \bar{P}X - \bigcup_{X \in \mathbb{U}/Q} \underline{P}X \tag{8}$$

All objects in the positive region $POS_P(Q)$, must belong to the set X . All objects in the negative region $NEG_P(Q)$, must not belong to the set X . And the objects in the boundary region $BND_P(Q)$, may belong to X . The model of attribute reduction in rough set requires that the positive region of the decision attribute remains unchanged.

If $IND(P) = IND(P - a)$, the attribute $a \in P$ is dispensable in the feature set, otherwise it is indispensable. To achieve attribute reduction, that is, to find the smallest subset P of the conditional attribute set. The minimum subset P needs to satisfy the following two conditions:

- (1) $POS_P(Q) = POS_C(Q)$
- (2) $\forall a \in P, POS_{P-\{a\}}(Q) = POS_C(Q)$

Then the subset P is a reduct of C .

2.3 Fuzzy Rough Sets

The membership of an object $x \in \mathbb{U}$, belonging to the fuzzy positive region can be defined by

$$\mu_{\underline{R}_P X}(x) = \inf_{y \in \mathbb{U}} I(\mu_{R_P}(x, y), \mu_X(y)) \tag{9}$$

Here I is a fuzzy implicator and T is a t-norm. R_P is the fuzzy similarity relation induced by the subset of features P :

$$\mu_{R_P}(x, y) = \bigcap_{a \in P} \{\mu_{R_a}(x, y)\} \tag{10}$$

Many fuzzy similarity relations can be constructed to represent the similarity between objects x and y for feature a , such as

$$\mu_{R_a}(x, y) = 1 - \frac{|a(x) - a(y)|}{|a_{\max} - a_{\min}|} \tag{11}$$

$$\mu_{R_a}(x, y) = \exp\left(-\frac{(a(x) - a(y))^2}{2\sigma_a^2}\right) \tag{12}$$

where σ_a^2 is the variance of feature a . The fuzzy positive region can be defined as

Table 1. Selected t-Norms and their duals (S Conorms)

	Operators T	Operators S
1	$T_M(a, b) = \min(a, b)$	$S_M(a, b) = \max(a, b)$
2	$T_P(a, b) = a \times b$	$S_P(a, b) = a + b - ab$
3	$T_L(a, b) = \max(a + b - 1, 0)$	$S_L(a, b) = \min(a + b, 1)$
4	$T_{\cos}(a, b) = \max(ab - \sqrt{1 - a^2}\sqrt{1 - b^2}, 0)$	$S_{\cos}(a, b) = \min(a + b - ab + \sqrt{2a - a^2}\sqrt{2b - b^2}, 1)$

$$\mu_{POS_{R_P}(Q)}(x) = \sup_{X \in \mathbb{U}/Q} \underline{\mu}_{R_P}(x) \tag{13}$$

Using the definition of the fuzzy positive region, the new dependency function can be defined as follows:

$$\gamma'_P(Q) = \frac{\sum_{x \in \mathbb{U}} \mu_{POS_{R_P}(Q)}(x)}{|\mathbb{U}|} \tag{14}$$

A fuzzy-rough reduct R can be defined as a subset of features that preserves the dependency degree of the entire dataset, that is, $\gamma'_R(\mathbb{D}) = \gamma'_C(\mathbb{D})$.

2.4 Kernelized Fuzzy Rough Set

Some widely encountered kernel functions satisfying reflexivity, symmetry, and transitivity are:

1. Gaussian kernel: $k_G(x, y) = \exp\left(-\frac{\|x-y\|^2}{\delta_u}\right)$
2. Exponential kernel: $k_E(x, y) = \exp\left(-\frac{\|x-y\|}{\delta}\right)$
3. Rational quadratic kernel: $k_R(x, y) = 1 - \frac{\|x-y\|^2}{\|x-y\|^2 + \delta}$

With the kernel function and the fuzzy operator in Table 1 and Table 2, we can substitute fuzzy relations in fuzzy rough sets. The kernelized fuzzy lower and upper approximation operators are defined as:

1. S -kernel fuzzy lower approximation operator: $\underline{k}_S X(x) = \inf_{y \in U} S(N(k(x, y)), X(y))$;
2. θ -kernel fuzzy lower approximation operator: $\underline{k}_\theta X(x) = \inf_{y \in U} \theta(k(x, y), X(y))$;
3. T -kernel fuzzy upper approximation operator: $\overline{k}_T X(x) = \sup_{y \in U} T(k(x, y), X(y))$;
4. σ -kernel fuzzy upper approximation operator: $\overline{k}_\sigma X(x) = \sup_{y \in U} \sigma(N(k(x, y)), X(y))$

Let the classification be formulated as $\langle U, A, D \rangle$, where U is thenonempty and finite set of samples, A is the set of features characterizing the classification, D is the class attribute which divides the samples into subset $\{d_1, d_2, \dots, d_K\}$. For $\forall x \in U$,

$$d_i(x) = \begin{cases} 0, & x \notin d_i \\ 1, & x \in d_i \end{cases}$$

We construct the algorithms for computing the fuzzy lower and upper approximations for a given kernel function.

1. $\underline{k}_S d_i(x) = \inf_{y \notin d_i} (1 - k(x, y));$
2. $\underline{k}_\theta d_i(x) = \inf_{y \notin d_i} \left(\sqrt{1 - k^2(x, y)} \right);$
3. $\overline{k}_T d_i(x) = \sup_{y \in d_i} k(x, y);$
4. $\overline{k}_\sigma d_i(x) = \sup_{y \in d_i} \left(1 - \sqrt{1 - k^2(x, y)} \right).$

The kernelized dependency function is defined as follows:

$$\gamma_B^S(D) = \frac{|\cup_{i=1}^I \underline{k}_S d_i|}{|U|} \text{ or } \gamma_B^\theta(D) = \frac{|\cup_{i=1}^I \underline{k}_\theta d_i|}{|U|} \tag{15}$$

The coefficients of classification quality reflect the approximation ability of the approximation space or the ability of the granulated space induced by attribute subset B to characterize the decision.

3 Kernelized Fuzzy Rough Set-Based Three-Way Decision Feature Selection

This section first expounds the problems existing in the heuristic kernelization dependency feature selection strategy, and then describes the feature selection method using the idea of three-way decision.

3.1 Heuristic Feature Selection

Since finding the minimum subset is an NP-hard problem, a heuristic search algorithm is generally used to obtain feature subsets. The maximal dependency(MD) strategy is designed in [13], and its heuristic feature evaluation function is

$$\max_{f \in C-S} \Psi(f, S, D) \tag{16}$$

where $\Psi(f, S, D) = \gamma_B^{S \cup \{f\}}(D) - \gamma_B^S(D)$, C is the initial feature set, S is the selected feature subset, D is the decision feature, and F is a candidate feature.

The purpose of feature selection is to obtain the feature subset with the fewest features under the condition of maintaining the descriptive ability of the feature subset. MD adopts a greedy strategy, that is, adding a candidate feature that maximizes Ψ in each step, so that the dependency of the selected feature subset increases as quickly as possible, and its search can only guarantee a local optimum. The selected feature subset may be too large and redundant, and the quality of the feature subset is difficult to guarantee.

3.2 Feature Selection Based on Three-Way Decision

In order to avoid the problems caused by the greedy strategy and make the feature subset more concise and informative, this paper proposes a three-way decision-based feature selection strategy. In the three-way search, generally each layer maintains three feature subsets, which are used to generate the top three new feature subsets respectively, totaling 9 candidate feature subsets. Then, the top three are selected from the 9 feature subsets, and they are constrained from not originating from the same branch as the 3 feature subsets of the next layer. Three-way feature selection will eventually generate 3 better feature subsets. The method of feature selection and generation of successor is as follows:

$$\bigcup_{i=1}^w \max_{f_i \in C - S_i} \Psi(f_i, S_i, D), \tag{17}$$

$$\Psi(f_i, S_i, D) = \gamma_B^{S_i \cup \{f_i\}}(D) - \gamma_B^{S_i}(D), \tag{18}$$

C is the conditional feature set, i represents the sequence number of the branch, S_i represents the feature selected by the i th branch, and f_i represents the candidate feature of the i th branch.

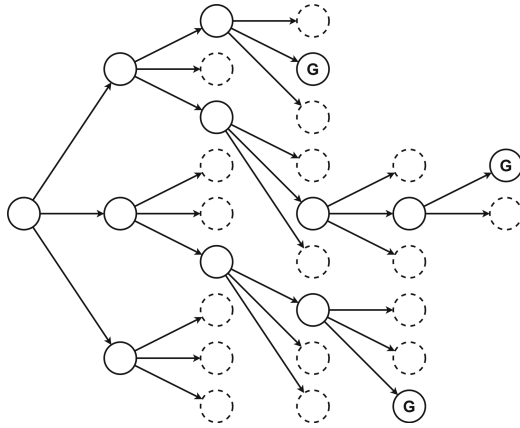


Fig. 1. Three-way feature selection

The idea of three feature selection is shown in Fig. 1. The solid and dashed circle nodes in the figure represent a subset of features. The solid circle indicates that the feature subset will continue to expand, and the dashed circle indicates that the feature subset will not expand. Node G indicates that the feature subset has reached the stopping condition.

The specific descriptions of the three feature selection algorithms are as Algorithm 1.

Algorithm 1. Kernelized fuzzy rough set-based three-way decision feature selection

Input: A kernelized fuzzy rough set-based three-way decision system $KFDS = \langle U, A, D \rangle$, Cutoff threshold θ

Output: Three reduced feature subsets R

```

1:  $subset = \{\{\emptyset\}\}, R = \{\emptyset\}$ 
2:  $k = 3$ 
3: while  $flag$  do
4:    $flag = FALSE$ 
5:   for all  $i$  in  $subset$  do
6:     if  $i.dependency > \theta$  then
7:        $R.add(i)$ 
8:        $subset.remove(i)$ 
9:        $k = k - 1$ 
10:       $continue$ 
11:    end if
12:     $bestAttrs = getMaxDependencyGain.Attrs(A - i)$ 
13:    for all  $j$  in  $bestAttrs$  do
14:       $subset.add(i \cup j)$ 
15:       $flag = TRUE$ 
16:    end for
17:  end for
18:   $subset = subset.getTopK(k)$ 
19: end while

```

The algorithm first starts with an empty set, and selects the top three features of dependency to form a feature subset of size 1. Next, test whether the current feature subset reaches the threshold. If it reaches the threshold, terminate the expansion of the subset and add it to the output subset set. Otherwise, continue to select the top three features of dependency to expand the subset until the subset There are three feature subsets in the set. In order to maintain the difference of feature subsets, the algorithm constrains that all subsets selected in each round cannot come from the same branch, and existing subsets cannot be selected.

Let the size of the original feature set A in the dataset be N . In the k th round, a feature subset has selected k features, and the time complexity of calculating the dependency gain of the remaining $(N - k)$ features is $O(N - k)$. Then in the worst case, that is, when all features are selected, the total complexity of one feature subset is $O\left(\sum_{k=1}^N (N - k)\right) = O(N^2)$, and the total complexity of three feature subsets is approximately $O(N^2)$.

After obtaining 3 feature subsets, the 3 feature subsets are respectively constructed as homogeneous learners to form three collaborative decision-making models to obtain better learning performance.

3.3 Computational Complexity

The main computational cost of three-branch decision feature selection is from the computation of kernelized dependency with different feature subsets and the selection of features with different branches. Compared with the traditional fuzzy rough set, the usage of kernel functions greatly reduces the storage space and computational cost. With M features, the time complexity of computing the Euclidean distance between a pair of samples is $O(M)$. With N samples, it first spends $O(N)$ to calculate the kernelized lower approximation of each sample, and then merges the lower approximation of all samples by $O(N)$ to obtain the kernelized dependency to measure the quality of feature subsets. In the feature selection process of the three-branch decision, each branch evaluates M features at most and the size of the branch is at most $M - 2$ features. Therefore, the time complexity of computing the kernelized dependency and the feature selection process at different branches are $O(N^2M)$ and $O(M^2)$, respectively. However, the actual computation cost will be much smaller than the theoretical computation cost due to the branch size and the cutoff threshold.

4 Experiment

This part mainly includes the experiment steps and presents an analysis of the model with classification accuracy. We compare the three-way decision model based on kernelized fuzzy rough sets with the traditional greedy algorithm. At the same time, we also make the comparison between soft voting and hard voting for the model in this paper. For each sample, we obtained its three feature subsets obtained, and the closest distance from each feature subset to each class in the data set is calculated and voted, then the closest distance is selected as the feature subset described. Finally, the class to which the majority of feature subsets belong is taken as the class of the sample, this method is called hard voting, while soft voting corresponds to it, the sum of the three feature subsets to the nearest samples of a certain class is taken as the total distance, then the class to which the minimum value belongs can be taken as the class to which it belongs by comparing all distances. In order to facilitate the following representation of the experiment, 'KFRS-FS(S)' is used for soft voting, and 'KFRS-FS(H)' is used for hard voting.

4.1 Datasets and Settings

In this experiment, the specific information of the datasets is shown in Table 3. We summarize the basic information of each dataset as dataset name, number of features and number of samples. At the same time, for the experiment results of each dataset, the average performance of the ten-fold cross-validation method is used as the final performance of our model on the dataset, in order to eliminate the adverse effects of accidental errors in the experiments.

Tests on small-scale datasets show that the kernel-based fuzzy rough set method can extract better feature subsets when the dependency value belongs

to $[0.5, 1.0]$. The performance of the algorithm in this paper is compared with that of the greedy algorithm with dependency in $[0.5, 1.0]$. At the same time, for each dataset, the performance difference between soft voting and hard voting is compared. The specific experimental data are shown in Table 5 and Table VI.

Table 2. Residual Implication Induced by the t-Norms and Their Duals

	Residual implication θ	Operator σ
1	$\theta_M(a, b) = \begin{cases} 1, & a \leq b \\ b, & a > b \end{cases}$	$\sigma_M(a, b) = \begin{cases} 0, & a \geq b \\ b, & a < b \end{cases}$
2	$\theta_p(a, b) = \begin{cases} 1, & a = 0 \\ \min(1, b/a), & \text{otherwise} \end{cases}$	$\sigma_p(a, b) = \begin{cases} 1, & a = 0 \\ \max(0, \frac{b-a}{1-a}), & \text{otherwise} \end{cases}$
3	$\theta_L(a, b) = \min(b - a + 1, 1)$	$\sigma_L(a, b) = \min(0, b - a)$
4	$\theta_{\cos}(a, b) = \begin{cases} 1, & a \leq b \\ ab + \sqrt{1-a^2}\sqrt{1-b^2}, & a > b \end{cases}$	$\sigma_{\cos}(a, b) = \begin{cases} 0, & a > b \\ a + b - ab - \sqrt{2a-a^2}\sqrt{2b-b^2}, & a \leq b \end{cases}$

4.2 Algorithm Performance Comparison

The performance comparison results of the two algorithms on the selected dataset are shown in Table 4. In Table 4, the second column represents the performance of the model in this paper, which is represented by KFRS-FS here, and the third column represents the performance of the classic greedy algorithm, which is represented by GA. It can be seen that the performance of the algorithm in this paper is generally higher than that of the greedy algorithm. Among them, there are more than 5% points of performance improvement in australian, bupa, dnatest, mammographic, spect-train or other datasets, and the improvement is more significant. From overall view, the KFRS-FS algorithm proposed in this

Table 3. Experiment datasets

Dataset	Features	Objects
appendicitis	8	106
australian	15	690
bupa	7	345
dnatest	181	1186
fetal-state	21	2126
german	7	345
haberman	4	306
mammographic	5	748
spectf-train	22	267
vehicle	18	946
wdbc	31	569
weather	5	22

Table 4. Classification accuracy by KFRS-FS and GA

Datasets	KFRS-FS-W1	KFRS-FS-W2	KFRS-FS-W3	KFRS-FS	GA
appendicitis	84.91	86.82	85.73	90.00	87.27
australian	85.22	84.93	85.22	84.78	80.86
bupa	61.71	64.30	66.06	63.42	57.71
dnatest	39.14	38.87	36.76	46.38	38.31
fetal-state	92.05	91.53	91.91	91.97	91.39
german	71.50	71.80	74.00	72.80	71.30
haberman	64.63	71.85	69.57	67.10	64.19
mammographic	73.52	75.00	75.27	75.60	70.66
spectf-train	78.75	78.75	77.75	82.50	76.50
vehicle	70.58	70.58	70.81	73.17	69.27
wdbc	97.19	97.19	97.37	97.54	96.49
weather	85.00	91.67	86.33	90.00	86.66

paper has different degrees of increase in algorithm performance compared with the classical greedy algorithm according to different datasets.

4.3 Analysis of KFRS-FS

This part is mainly aimed at the comparison between soft voting and hard voting inside the KFRS-FS algorithm introduced in this paper, as shown in Table 5 and Table 6, where the second column represents the feature subset distribution obtained by soft voting and hard voting, and the third column represents the performance of soft voting and hard voting. It can be seen that for datasets with fewer features, the performance of soft voting is higher than that of hard voting. On the contrary, the performance of datasets with more features is better than hard voting. It proves that hard voting, which first finds the class to which each feature subset belongs, will have more advantages in the comparison of model performance in the sample space with high dimension while soft voting will ignore the performance of individual feature subsets and try to find an overall performance, this gives soft voting a poor effects in higher dimensions. However, in low-dimensional space, the overall performance will have a better model performance.

In addition, experiments show that the most appropriate cutoff threshold varies in different datasets. When the cutoff threshold is too low, the model can not fully exploit the information in the feature space. When the cutoff threshold is too high, the feature subset may have high redundancy. Both of these will lead to degradation in the performance of the model.

Table 5. Classification accuracy by soft voting

Datasets	Feature subsets by KFRS-FS(S)	Performance of KFRS-FS(S)
appendicitis	[[2, 3], [6, 3], [4, 6]]	85.45
australian	[[6, 2, 4, 7, 1, 5, 8, 10, 3, 0], [13, 6, 2, 4, 7, 1, 5, 8, 10, 3], [9, 6, 1, 4, 7, 2, 5, 8, 10, 3]]	81.15
bupa	[[5, 3, 1, 0, 4], [2, 5, 1, 0, 3], [3, 2, 5, 1, 4]]	63.42
dnatest	[[0, 1], [1, 2], [2, 0]]	39.66
fetal-state	[[6, 1, 7, 0, 12, 3, 9, 13, 14, 20, 4, 16, 10, 11], [1, 12, 7, 0, 3, 13, 9, 20, 6, 4, 16, 10, 11, 17, 8, 15, 2], [16, 1, 12, 7, 3, 14, 13, 9, 6, 20, 0, 10, 17, 4, 8]]	91.97
german	[[9, 5, 4, 0, 8, 7, 6, 2], [1, 3, 9, 4, 5, 0, 8, 7, 6], [3, 9, 4, 2, 5, 0, 8, 7, 6]]	72.29
haberman	[[0, 1], [2, 0], [1, 2]]	67.09
mammographic	[[0, 3, 2], [1, 0, 3], [2, 0, 1]]	68.53
spectf-train	[[40, 26, 1, 5, 4, 2, 7, 21, 33, 24, 3, 22], [42, 40, 3, 8, 1, 33, 21, 4, 7, 28, 2, 14], [41, 40, 28, 1, 4, 21, 8, 3, 7, 33, 2, 24]]	82.50
vehicle	[[8, 17, 14, 15, 9, 0, 12, 2, 16, 7, 3], [11, 17, 14, 15, 9, 0, 2, 12, 16, 3, 7, 1], [6, 17, 14, 15, 9, 0, 2, 12, 16, 3, 7, 1]]	70.23
wdbc	[[22, 27, 21, 11, 24, 20, 8, 7, 18], [27, 0, 21, 11, 24, 20, 8, 6, 18], [20, 27, 21, 11, 24, 18, 22, 8, 9]]	97.01
weather	[[1, 0], [0], [2, 1, 3]]	90.00

Table 6. Classification accuracy by hard voting

Datasets	Feature subsets by KFRS-FS(H)	Performance of KFRS-FS(H)
appendicitis	[[2, 3], [6, 3], [4, 6]]	90.00
australian	[[6, 2, 4, 7, 1, 5, 8], [13, 6, 2, 4, 7, 1, 5], [9, 6, 1, 4, 7, 2, 5]]	84.78
bupa	[[5, 3, 1, 0], [2, 5, 1, 0], [3, 2, 5, 1]]	62.85
dnatest	[[0, 1], [1, 2], [2, 0]]	46.38
fetal-state	[[6, 1, 7, 0, 12, 3, 9, 13, 14, 20, 4, 16, 10, 11, 17, 8], [1, 12, 7, 0, 3, 13, 9, 20, 6, 4, 16, 10, 11, 17, 8, 15, 2], [16, 1, 12, 7, 3, 14, 13, 9, 6, 20, 0, 10, 17, 4, 8, 15]]	91.78
german	[[9, 5, 4, 0, 8, 7, 6, 2, 13, 23, 10, 1], [1, 3, 9, 4, 5, 0, 8, 7, 6, 2, 10, 23], [3, 9, 4, 2, 5, 0, 8, 7, 6, 10, 23, 13]]	72.80
haberman	[[0, 1], [2, 0], [1, 2]]	59.03
mammographic	[[0, 3, 2], [1, 0, 3], [2, 0, 1]]	75.60
spectf-train	[[40, 26, 1, 5, 4], [42, 40, 3, 8, 1], [41, 40, 28, 1, 4]]	73.75
vehicle	[[8, 17, 14, 15, 9, 0, 12, 2, 16, 7, 3, 1, 4], [11, 17, 14, 15, 9, 0, 2, 12, 16, 3, 7, 1, 4], [6, 17, 14, 15, 9, 0, 2, 12, 16, 3, 7, 1, 4]]	73.17
wdbc	[[22, 27, 21, 11, 24], [27, 0, 21, 11], [20, 27, 21, 11]]	97.54
weather	[[1], [0], [2, 1, 3]]	46.66

5 Conclusions

In this paper, the idea of three-way decision is introduced into feature selection based on kernelized fuzzy dependency. From the perspective of multi-branch, multiple feature subsets containing sufficient information and complementarity are obtained, and the classification performance of this method is further improved through ensemble learning. The algorithm proposed in this paper has been performed on benchmark datasets and compared with traditional methods. The experimental results show that the scale of the three feature subsets calculated by the new method is much smaller than the original number of features, which reduces the computational complexity of classification. Moreover, the ensemble learning based on three feature subsets has better classification accuracy on multiple datasets than the traditional kernelized fuzzy rough set feature selection method, indicating that the new method has better classification accuracy. Further research topics include how to extend the three-way decision to the semi-supervised domain, so that the method can be used in more practical situations.

Acknowledgements. The work was supported by the National Natural Science Foundation of China (Nos. 61806127).

References

1. Cervante, L., Xue, B., Shang, L., Zhang, M.: A multi-objective feature selection approach based on binary PSO and rough set theory. In: Middendorf, M., Blum, C. (eds.) *EvoCOP 2013*. LNCS, vol. 7832, pp. 25–36. Springer, Heidelberg (2013). https://doi.org/10.1007/978-3-642-37198-1_3
2. Li, Y., Li, T., Liu, H.: Recent advances in feature selection and its applications. *Knowl. Inf. Syst.* **53**(3), 551–577 (2017). <https://doi.org/10.1007/s10115-017-1059-8>
3. Pawlak, Z.: Rough sets. *Int. J. Comput. Inf. Sci.* **11**(5), 341–356 (1982)
4. Cervante, L., Xue, B., Shang, L., Zhang, M.: A dimension reduction approach to classification based on particle swarm optimisation and rough set theory. In: Thielscher, M., Zhang, D. (eds.) *AI 2012*. LNCS (LNAI), vol. 7691, pp. 313–325. Springer, Heidelberg (2012). https://doi.org/10.1007/978-3-642-35101-3_27
5. Qu, Y., Shen, Q., Parthaláin, N.M., Shang, C., Wu, W.: Fuzzy similarity-based nearest-neighbour classification as alternatives to their fuzzy-rough parallels. *Int. J. Approx. Reason.* **54**(1), 184–195 (2013)
6. Cattaneo, G., et al.: Abstract approximation spaces for rough theories. *Rough Sets Knowl. Discov.* **1**, 59–98 (1998)
7. Dubois, D., Prade, H.: Rough fuzzy sets and fuzzy rough sets. *Int. J. Gen. Syst.* **17**(2–3), 191–209 (1990)
8. Dubois, D., Prade, H.: Putting rough sets and fuzzy sets together. In: Slowiński, R. (ed.) *Intelligent Decision Support*, vol. 11, pp. 203–232. Springer, Dordrecht (1992). https://doi.org/10.1007/978-94-015-7975-9_14
9. Gong, Z., Sun, B., Chen, D.: Rough set theory for the interval-valued fuzzy information systems. *Inf. Sci.* **178**(8), 1968–1985 (2008)

10. Qinghua, H., Daren, Yu., Pedrycz, W., Chen, D.: Kernelized fuzzy rough sets and their applications. *IEEE Trans. Knowl. Data Eng.* **23**(11), 1649–1667 (2010)
11. Yao, Y.: Tri-level thinking: models of three-way decision. *Int. J. Mach. Learn. Cybern.* **11**(5), 947–959 (2020)
12. Gao, C., Zhou, J., Miao, D., Wen, J., Yue, X.: Three-way decision with co-training for partially labeled data. *Inf. Sci.* **544**, 500–518 (2021)
13. Qinghua, H., Zhang, L., Zhang, D., Pan, W., An, S., Pedrycz, W.: Measuring relevance between discrete and continuous features based on neighborhood mutual information. *Expert Syst. Appl.* **38**(9), 10737–10750 (2011)