Knowledge-Based Systems 255 (2022) 109795

Contents lists available at ScienceDirect

Knowledge-Based Systems

journal homepage: www.elsevier.com/locate/knosys

Feature selection based on robust fuzzy rough sets using kernel-based similarity and relative classification uncertainty measures

Pei Liang [a,b], Dingfei Lei [a], KwaiSang Chin [b], Junhua Hu [a,*]

[a] School of Business, Central South University, Changsha 410083, China
[b] Department of Advanced Design and Systems Engineering, City University of Hong Kong, Hong Kong

ARTICLE INFO

Article history:
Received 9 February 2022
Received in revised form 24 August 2022
Accepted 25 August 2022
Available online 31 August 2022

Keywords:
Feature selection
Fuzzy rough set
Similarity measure
Robustness
Noise

ABSTRACT

The current research on fuzzy rough sets (FRSs) for feature selection has two major problems. On the one hand, most existing methods employ multiple intersection operations of fuzzy relations to define fuzzy dependency functions applied to feature selection. These operations can make the evaluation of the significance of feature subsets less identifiable in high-dimensional data space. On the other hand, the classical FRS implemented for feature selection is highly sensitive to noisy information. Thus, improving the robustness of the FRS model is critical. To address the above issues, first, we propose a radial basis function kernel-based similarity measure for computing fuzzy relations. The value difference metric and Euclidean metric are utilized to measure the distance values of the mixed symbolic and real-valued features. Hereafter, a novel robust FRS model is proposed by introducing the relative classification uncertainty (RCU) measure. *k*-nearest neighbours and Bayes rules are employed to yield an RCU level. Relative noisy information is detected in this way. Finally, extensive experiments are conducted to illustrate the effectiveness and robustness of the proposed model.

1. Introduction

In medical diagnosis [1], text mining [2], image annotation [3] and gene expression [4] domains, data typically involve an enormous number of features. However, some features engaged in a given learning task are generally redundant. Feature selection, also known as attribute reduction [5], pertains to finding a reduct under the premise of holding the identical performance as the family feature set in the learning task. Moreover, feature selection not only helps remove redundant features but also improves storage, speed and accuracy. Two key steps are included in feature selection: construct the feature evaluation function and design the feature searching strategy. The feature evaluation function delimits the quality of the candidate feature subset, such as dependency [6], neighbourhood dependency [7,8] and fuzzy dependency [9] in rough set theory; mutual information [10,11] in information theory; sample margin [12] in statistical theory; and some machine learning-based methods [13–16]. The feature searching strategy is used to speed up the optimal subset searching process and mainly includes sequential forward search iteration [17], sequential backward search iteration [18] and some intelligent optimization-based searching algorithms such as genetic algorithm [19], particle swarm optimization algorithm [20], and chaotic cuckoo optimization algorithm [21].

Among these theories, rough sets have been successfully applied to feature selection, and multi-source data can be addressed [22]. Pawlak [13] proposed the classical rough set, which is mainly used to handle symbolic or discrete data. Real-valued data need to be discretized before feature selection. However, discretization might cause mutual information loss in the real-valued features. Accordingly, Dubois and Prade [9] proposed the classical fuzzy rough set (FRS), which can deal directly with real-valued features. Crisp equivalence relations are converted into membership degrees of objects corresponding to decisions in FRS. Hereafter, a series of extensions of FRSs have been proposed [23–25], where varying operators are introduced to define the fuzzy similarity relations, such as the T-similarity measure, R-implication measure [24] and kernel-based similarity measure [25,26]. Applying FRS models for feature selection draws more attention as well. To date, many critical studies have been presented to implement FRS-based feature selection [27–33]. For example, Chen et al. [28] developed the attribute reduction algorithm using the discernibility matrix to find the minimal elements. Similarly, Chen and Yang [27] combined the classical rough set and FRSs to define a discernibility relation for every symbolic and real-valued condition attribute, where intersection operators were used. Wang et al. [30] introduced the parameterized fuzzy relation for characterizing fuzzy information granules

* Corresponding author.
E-mail addresses: shirley_lp@csu.edu.cn (P. Liang), 181611128@csu.edu.cn (D. Lei), mekschin@cityu.edu.hk (K. Chin), hujunhua@csu.edn.cn (J. Hu).

https://doi.org/10.1016/j.knosys.2022.109795
0950-7051/

to analyse real-valued data. They used the relationship between the fuzzy neighbourhood and fuzzy decision to construct the fuzzy neighbourhood rough set model. Yang et al. [31] studied incremental attribute reduction with FRSs. The discernibility relations were updated dynamically with the number of attributes increasing or decreasing. Readers can refer to these studies for more methods to compute the fuzzy relations.

However, one limitation can be found in these studies. The intersection operation used to calculate fuzzy similarity relations can easily result in a slight value difference in a high-dimensional data space. For example, let B be a subset of the features, R_a is the fuzzy similarity relation induced by the feature a and R_B is computed by the multiple intersection operations of R_a, such as $R_B = \cap_{a\in B} R_a$. This operation is usually equal to deriving the minimum value of all R_a in the subset B. In high dimensionality, obtaining very small values of R_a is very common and leads to R_B being less discriminated. Since the fuzzy dependency membership is derived by fuzzy similarity relations through fuzzy rough approximations, the little-discriminated fuzzy similarity relations will hinder the fuzzy dependency membership from truly reflecting the relationship among different objects. Hence, the classification ability of varying feature subsets dependent on the fuzzy dependency membership is affected. This problem has also been stated in [26,34].

To overcome this problem, Wang et al. [34] used distance measures of Manhattan distance, and Yuan et al. [26] employed a hybrid kernel function to compute the fuzzy similarity relations, which could surmount the discrimination bottleneck encountered in the intersection operations. Kernel functions such as the linear kernel [35], wavelet kernel [36], Hermite kernel [37] and Gaussian kernel [38] are commonly used. The linear kernel can be used to deal with simple and extensive linear relations, the wavelet kernel can ensure global optimality of the relations, the Hermite kernel performs well in complex relations, and the Gaussian kernel is the most popular function in the support vector machine (SVM) classifier. All of them can be used in a similarity measure. In contrast, we present a solution using a radial basis function (RBF) kernel-based similarity measure with mixed distance metrics to compute the fuzzy relations (RBF kernel and Gaussian kernel are considered interchangeable.). The RBF kernel-based similarity measure [39] takes advantage of addressing nonlinear feature relations in high-dimensional feature space. Particularly, we use the value difference metric (VDM) [40] and Euclidean metric [41] to measure the distance of objects under the specific symbolic feature subset and real-valued feature subset, respectively. Euclidean distance is a desired metric for measuring the distance between real-valued features [38]. In addition, the VDM uses conditional probability terms to estimate the distance of symbolic values and considers the connections between symbolic values and output classes. Jia et al. [42] stated that the VDM is a commonly used distance metric for dealing with discrete attributes. They used it to calculate the distance between all discretized attribute values for similarity-based attribute reduction in rough set theory. Luo et al. [43] indicated that the inherent ordered relationships and statistical information from nominal values can be captured through VDM, which improved the accuracy and validity of data representation in the neighbourhood rough set model. Similarly, Hamed et al. [44] exploited the usefulness of the VDM to address categorical features and took missing values into account. Existing research reaps the benefits from the VDM metric. With VDM, symbolic feature values are not simply given equal distance from each other, and the differing degrees of similarities of value pairs are considered. Thus, we propose VDM combined with Euclidean distance metrics to manage the symbolic and real-valued features and eliminate the need to transform mixed data types into one kind. More advantages of this mixed distance metric can also be found in [38]. The difference in the current study is that we consider the influence of dimension and normalize the distances of different types of features before mixing. This normalized distance value can balance the distance of the two feature types and facilitate the synthesis of the value more reasonably. Subsequently, the normalized distance value is fed into the RBF kernel to obtain the fuzzy similarity relation membership. Most importantly, the discrimination level of different fuzzy dependency memberships is more noticeable because of the replacement of multi-intersection operations.

The problem of learning in noisy environments has also attracted much attention in feature selection studies. Noise can be physically distinguished into two categories: (1) feature noise and (2) class noise [45]. The former mainly refers to incomplete, unknown or missed information in features. The latter is represented as misclassifications. In classical FRS, fuzzy rough approximations are highly sensitive to class noise since the values are associated with the nearest neighbour objects to the given target object, where different class labels are needed. Once the class label of the nearest neighbour is misclassified, the quality of the fuzzy approximation is reduced. Thus, some robust FRS models have been proposed to address the influence of class noise. Robust FRS models are roughly divided into two groups. In the first group, classification boundary objects can be seen as noisy objects, such as β-precision FRS [46], soft FRS [47], k-trimmed FRS [48], nested topological FRS [49] and different classes' ratio FRS (DC_ratio FRS) [50]. A crucial characteristic of these models is how they select nearest neighbour objects. The second group employs robust approximation operators, such as k-nearest neighbours (KNN) FRS [34], k-means FRS [48], k-median FRS [48] and ordered weighted average FRS [51]. Next, we briefly explain the related work. Ref. [50] considered the influence of the nearest object for a specific target object. A definition of different class ratios of the nearest object was proposed to detect the noisy data. If the ratio value is larger than a predefined threshold, then the nearest object is seen as noisy data and abandoned to compute the fuzzy rough approximations. In [34], the average value of $1-R_B$ (R_B represents the fuzzy similarity relations induced by the feature subset B) of the k nearest objects for the target objects was employed to compute the fuzzy rough approximations. Among these robust models, [50] verified that the DC_ratio FRS model is generally superior to β-precision FRS [46], soft FRS [47], k-trimmed FRS [48], k-means FRS and k-median FRS [48]. The distance-based robust FRS model proposed in [34] outperformed the robust model in [49].

Nevertheless, some shortcomings are found in [34,50]. First, detecting the noisy information under all the individual features for each object in the universe is too trivial. Second, as the different class ratios are computed in a circular range with a small radius, a large deviation will occur if initial noisy information exists in the circular range. Third, a simple aggregation of the neighbours seems too crude to handle the robustness issue. To overcome these deficiencies, this study introduces a robust FRS model that considers the influence of the nearest object of a target object and considers the neighbours' insight distribution of the nearest object by calculating the relative classification uncertainty (RCU) value. An RCU measure is defined to detect noisy objects. The KNN algorithm and Bayes theory are employed to compute the RCU, which can eliminate the side effects caused by the initial noisy information existing in the circular range of the nearest object to the target object. Moreover, extensive experiments are conducted to show the robustness and effectiveness of the proposed model.

With regard to the feature searching strategy, we design the heuristic forward coupled with the heuristic backward searching algorithm to explore the feature reduct. It aims to avoid becoming

stuck in a local optimum. Currently, the development of FRS works bridges some new research perspectives for its derived concepts, such as some three-way decision-making models [52–55]. Particularly, our work has practical implications for managers focused on enabling, producing or consuming analytics in a broad array of contexts where the optimal feature subset selection for complex data sets may facilitate enhanced insight and foresight. The main contributions of our work are threefold. First, we propose a novel FRS model for feature selection. Second, as part of the proposed model, we design the kernel-based fuzzy relations measure and the RCU measure. The former takes advantage of computing fuzzy relations of mixed features, which can overcome the feature subset's significance identification bottleneck encountered in existing intersection operations. The latter employs KNN coupled with Bayes rules to capture the RCU level of certain objects. Relative and absolute noisy objects appearing in the fuzzy approximation process are detected to enhance the robustness of the FRS model. Third, extensive experimental results show that our proposed model outperforms some other models in classification performance and statistical analysis.

This paper is further organized as follows. Section 2 presents some related notions of classical FRS. Section 3 discusses the distance and similarity measures for mixed-type features in the FRS model and presents the associated properties. Section 4 offers the RCU FRS model and analyses the related property. Section 5 designs the feature selection algorithm with the RCU FRS model based on two greedy searching algorithms. Section 6 discusses the experimental results. Finally, Section 7 concludes the study.

2. Basic notions

This section reviews some basic notions regarding the classical FRS [9], including fuzzy relations, fuzzy lower and upper approximations and the fuzzy dependency function. First, the fuzzy relations used to divide equivalence classes are introduced.

Let U be a nonempty universe, and let R be a fuzzy binary relation on U. We say that R is a fuzzy equivalence relation if it satisfies the following:

(1) Reflexivity: $R(x,x)=1, \forall x \in U$.
(2) Symmetry: $R(x,y)=R_a(y,x), \forall x,y \in U$.
(3) Sup-min transitivity: $R(x,y) \geq \sup\min_{z\in U}\{R(x,z), R(z,y)\}$.

For any $x \in U$, let $[x]_R(y)=R(x,y), y \in U$, where $[x]_R$ is the fuzzy similarity or equivalence class associated with x and R on U, which is also called the fuzzy neighbourhood of x.

Definition 1 ([9]). Let $DT=\langle U, A\cup D\rangle$ be the fuzzy decision system with a mapping $a|U \to V_a$ for each $a \in A$, where $U=\{x_1,x_2,\ldots,x_n\}$ is a finite set of objects and $A=\{a_1,a_2,\ldots,a_m\}$ is the condition attribute set. A fuzzy relation R_a is defined for each condition attribute $a \in A$, V_a is the domain of a and $D=\{d\}$ is the decision attribute set where a mapping $d|U \to V_d$ is defined. V_d is the domain of the decision attribute d with nominal values. D partitions the sample set U into r crisp equivalence classes $U/D=\{D_1,D_2,\ldots,D_r\}\ (1\leq j\leq r)$. B is a subset of A, denoted as $B\subseteq A$, and R_a is the fuzzy similarity relation for each $a \in B$. Let

$$R_B=\bigcap_{a\in B} R_a. \tag{1}$$

Then, R_B is a fuzzy similarity relation on U. Furthermore, the membership function of the decision class D_j is

$$D_j(x)=\begin{cases}1, x\in D_j;\\ 0, otherwise.\end{cases} \tag{2}$$

To address the uncertainty of the decision feature, the classical FRSs introduce the definitions of the fuzzy rough approximations as follows:

$$\underline{R_B}\left(D_j\right)(x)=\inf_{y\notin D_j}\{1-R_B(x,y)\}, D_j\in U/D, \tag{3}$$

$$\overline{R_B}\left(D_j\right)(x)=\sup_{y\in D_j}\{R_B(x,y)\}, D_j\in U/D, \tag{4}$$

where $\underline{R_B}\left(D_j\right)(x)$ and $\overline{R_B}\left(D_j\right)(x)$ are called the fuzzy lower and upper approximations of D_j, respectively. $\underline{R_B}\left(D_j\right)(x)$ denotes that the membership of x certainly belonging to D_j is equal to the minimum of the dissimilarities between x and all objects from the sample domain $U-D_j$. $\overline{R_B}\left(D_j\right)(x)$ means that the membership of x possibly belonging to D_j is equal to the maximum of the similarities between all objects from D_j.

The fuzzy positive region $POS_B(D)$ of the condition attribute subset concerning decision attribute D is defined as follows:

$$POS_B(D)=\bigcup_{j=1}^{r}\underline{R_B}\left(D_j\right). \tag{5}$$

With the definition of the fuzzy positive region, the fuzzy dependency function can be computed as follows:

$$\gamma_B(D)=\frac{|POS_B(D)|}{|U|}, \tag{6}$$

where $|\cdot|$ represents the cardinality of a set. Clearly, $0\leq\gamma_B(D)\leq 1$, which is defined as the ratio of the positive region over all objects in the feature space and used to evaluate the significance of a subset of features. The fuzzy dependency reflects the classification ability of a specific feature subset B.

3. FRS model based on a kernel-based similarity measure with a combination of distance metrics

The classical FRS estimates the fuzzy similarity relations of a specific feature subset through intersection operations. Fuzzy dependency membership discrimination may be tiny with multiple intersection operations in a high-dimensional data space. Considering such limitations, we assess fuzzy similarity relations through a kernel-based similarity measure to derive discriminable results. Furthermore, the classical FRS is usually utilized to address real-valued features, whereas a wide variety of real-world problems appear with mixed symbolic and real-valued features. Rather than transforming various features into a single type, we apply two different distance metrics to measure the distance of symbolic and real-valued features. Later, a combination of the two distance metrics is fed into the RBF kernel to calculate the fuzzy similarity relations of objects.

Suppose $DT=\langle U, A\cup D\rangle$ is the fuzzy decision system. In this case, $A\cap D=\phi$ and $A=A_S\cup A_R$, where A_S and A_R denote the families of symbolic and real-valued conditional features, respectively. The features are grouped into two subsets according to the feature type, such as $A_S=\left\{a_1,a_2,\ldots,a_{m_S}\right\}$ and $A_R=\left\{a_{m_S+1},\ldots,a_{m_R}\right\}$ and $A=A_S\cup A_R$ and $A_S\cap A_R=\phi$.

For symbolic features A_S, a common distance metric is that the distance value is 1 if two values are different and 0 otherwise. This metric has shortcomings in some specific conditions. All feature values are assumed to be of equal distance and thus cannot depict value pairs with differing degrees of similarities. VDM, a prevalent distance metric for symbolic features [40], can address this limitation because it considers the classification ability of specific features' values. VDM defines the distance of

$$d_A(x,y)=\frac{\sum_{i=1}^{m_S}\sum_{c_j\in V_d}\left|s_{a_i(x)\wedge c_j}/s_{a_i(x)}-s_{a_i(y)\wedge c_j}/s_{a_i(y)}\right|}{\max\left\{\sum_{i=1}^{m_S}\sum_{c_j\in V_d}\left|s_{a_i(x)\wedge c_j}/s_{a_i(x)}-s_{a_i(y)\wedge c_j}/s_{a_i(y)}\right|:x,y\in U\right\}}+\frac{\sqrt{\sum_i^{m_R}\left(a_i(x)-a_i(y)\right)^2}}{\max\left\{\sqrt{\sum_i^{m_R}\left(a_i(x)-a_i(y)\right)^2}:x,y\in U\right\}}. \tag{9}$$

Box I.

paired objects (x,y) under the feature set A_S as follows:

$$d_{A_S}(x,y)=\sum_{i=1}^{m_S}\sum_{c_j\in V_d}\left|s_{a_i(x)\wedge c_j}/s_{a_i(x)}-s_{a_i(y)\wedge c_j}/s_{a_i(y)}\right|,\ a_i(x),a_i(y)\in A_S, \tag{7}$$

where $V_d=\{c_1,\ldots,c_j,\ldots,c_r\}$ is the domain of the decision attribute set $D=\{d\}$ with nominal values and $a(x)$ denotes the value of the condition attribute a. m_S represents the number of symbolic family features A_S. $s_{a_i(x)\wedge c_j}$ represents the number of objects with condition attribute value $a_i(x)$ and arbitrary decision attribute value c_j. $s_{a_i(x)}$ represents the number of objects with the condition feature value $a_i(x)$. $s_{a_i(y)\wedge c_j}$ and $s_{a_i(y)}$ have similar annotations.

For real-valued features A_R, we use the Euclidean metric to measure the distance as follows:

$$d_{A_R}(x,y)=\sqrt{\sum_i^{m_R}\left(a_i(x)-a_i(y)\right)^2},\ a_i(x),a_i(y)\in A_R. \tag{8}$$

For the mixed features $A=A_S\cup A_R$, the distance metric is computed by a combination of Eqs. (7) and (8) as follows (see Box I)

Let $B\subseteq A$ and R_B be a fuzzy similarity induced by B. Then, a fuzzy similarity matrix can be generally represented as $R_B=(r_B(x,y))_{n\times n}$, where $0\le r_B(x,y)\le 1$ and $x,y\in U$. Data in the input space can actually be mapped implicitly into a high-dimensional space (feature space) with the help of kernel functions, where data are linearly separable in the new space (or linear separability is increased) [37]. Thus, we introduce that the fuzzy similarity degree $r_B(x,y)$ is computed by the RBF kernel as follows:

$$r_B(x,y)=\exp\left(-\left(d_B(x,y)\right)^2/2\delta^2\right),x,y\in U, \tag{10}$$

where $d_B(x,y)$ is computed by Eq. (9), and δ is a predefined parameter. B is a subset with mixed symbolic and real-valued features. The RBF kernel is beneficial for similarity measures involving nonlinear relationships.

Property 1. *Let $B\subseteq A$, then $R_A\subseteq R_B$.*

Definition 2. Let $DT=\langle U,A\cup D\rangle$ be the fuzzy decision system, $U/D=\{D_1,D_2,\ldots,D_r\}$ and $B\subseteq A$. R_B is a fuzzy similarity on U induced by Eq. (10). The lower and upper approximations of a decision class $D_j, j=1,2,\ldots,r$ with respect to B are defined as follows:

$$\underline{R_B}\left(D_j\right)(x)=\inf_{y\notin D_j}\left\{1-\exp\left(-\left(d_B(x,y)\right)^2/2\delta^2\right)\right\},x\in U; \tag{11}$$

$$\overline{R_B}\left(D_j\right)(x)=\sup_{y\in D_j}\left\{\exp\left(-\left(d_B(x,y)\right)^2/2\delta^2\right)\right\},x\in U. \tag{12}$$

Based on the fuzzy lower approximation, the fuzzy positive domain of decision D regarding feature subset B is computed by Eq. (5) such as $POS_B(D)=\cup_{j=1}^{r}\underline{R_B}\left(D_j\right)$.

A large positive domain is indicative of the classification ability of B. Thus, for a classification learning model, one typically finds a feature subset in which the classification model has a tremendously positive domain. The fuzzy dependency function is also computed as

$$\gamma_B(D)=\sum_{x\in U}POS_B(D)(x)/|U|. \tag{13}$$

Evidently, $0\le\gamma_B(D)\le 1$. The dependency function can describe the significance of one specific feature subset, thereby reflecting the classification capability of the feature subset B.

Property 2. *Let $DT=\langle U,A\cup D\rangle$ be the fuzzy decision system and $B_1,B_2\in A$ with mixed features. Then, $POS_{B_1}(D)\subseteq POS_{B_2}(D)$ if $B_1\subseteq B_2$.*

Property 3. *Let $DT=\langle U,A\cup D\rangle$ be the fuzzy decision system and $B_1,B_2\in A$ with mixed features. Then, $\gamma_{B_1}(D)\subseteq\gamma_{B_2}(D)$ if $B_1\subseteq B_2$.*

This property indicates that the fuzzy dependency function monotonically increases with regard to the size of the feature subset. Based on this monotonicity, we propose definitions considering the value of the feature number to determine the optimal feature subset.

Definition 3. Let $DT=\langle U,A\cup D\rangle$ be the fuzzy decision system and $B\subseteq A$. For any $a\in B$, we say feature a is indispensable in B if it satisfies $\gamma_{B-a}(D)-\theta\cdot\left(\frac{|B|-1}{N\cdot|U|}\right)<\gamma_B(D)-\theta\cdot\left(\frac{|B|}{N\cdot|U|}\right)$. Otherwise, a is said to be redundant in B.

$|B|$ is the cardinality of the feature subset B. N is the number of features in A. $|U|$ is the number of objects in U. $|B|/N$ represents the ratio of the number of selected features over the full features. θ $(0\le\theta\le 1)$ is a predefined threshold, which represents the marginal quantity utility factor regarding the number of selected features. This definition takes into account the significance of the selected feature subset and the selected number. Considering $\gamma_B=POS_B(D)/|U|$, multiplying $|B|/N$ by $1/|U|$ is to maintain the same value range as γ_B. In light of gaining the relatively largest γ_B and the relatively smallest $|B|/N$, the new significance evaluation of B is defined as $\gamma_B-\theta\left(\frac{|B|}{N\cdot|U|}\right)$. Referring to [34] in Definition 2, a is said to be indispensable if $\gamma_{B-a}<\gamma_B$. In our definition, $-\theta\left(\frac{|B|-1}{N\cdot|U|}\right)>-\theta\left(\frac{|B|}{N\cdot|U|}\right)$; if $\gamma_{B-a}(D)-\theta\cdot\left(\frac{|B|-1}{N\cdot|U|}\right)<\gamma_B(D)-\theta\cdot\left(\frac{|B|}{N\cdot|U|}\right)$, it further indicates that $\gamma_{B-a}<\gamma_B$ and proves that a is indispensable in B.

Corollary 1. *Let $DT=\langle U,A\cup D\rangle$ be the fuzzy decision system and $B\subseteq A$. For any $a\in B$, we say that a is redundant in B if it satisfies $\gamma_B(D)-\gamma_{B-a}(D)\le\frac{\theta}{N\cdot|U|}$. Otherwise, a is indispensable.*

Proof. According to Definition 3, if a is redundant, we have the following: $\gamma_{B-a}(D)-\theta\cdot\frac{|B|-1}{N\cdot|U|}\ge\gamma_B(D)-\theta\cdot\frac{|B|}{N\cdot|U|}\rightarrow\gamma_B(D)-\gamma_{B-a}(D)\le\theta\cdot\frac{|B|}{N\cdot|U|}-\theta\cdot\frac{|B|-1}{N\cdot|U|}\rightarrow\gamma_B(D)-\gamma_{B-a}(D)\le\frac{\theta}{N\cdot|U|}$. Thus, Corollary 1 holds.

Definition 4. Let $DT = \langle U, A \cup D \rangle$ be the fuzzy decision system and $B \subseteq A$. We say that B is a reduct of A if it satisfies the following:

(1) $\forall a \in B, \gamma_B(D) - \gamma_{B-a}(D) > \frac{\theta}{N \cdot |U|}$.

(2) $\gamma_A(D) - \gamma_B(D) \leq \theta \cdot \left(\frac{|A-B|}{N \cdot |U|}\right)$.

The definition shows that a reduct has a relative maximum dependency value and a minimal number of conditional features compared with the whole set of features A. According to Corollary 1, the first condition means that no single redundant feature can be split from the feature subset B. For the second condition, the loss of the dependency value of the full feature set A caused by dropping some features to obtain the feature subset B is compensated by the revenue of the decreasing number of features. Combining those two conditions implies that the feature subset of A is optimal. This definition is utilized for feature selection later, and θ is seen as the stop condition in the feature searching algorithm.

4. Robust FRS model using RCU measures

4.1. RCU measure for the object

According to the definition of the fuzzy lower approximation, $\underline{R_B}(D_j)(x) = \inf_{y \notin D_j} \{1 - R_B(x, y)\}$, $x \in U$, y is the nearest neighbour of the target object x and should not be in the same class as the target class, such as $y \notin D_j$. If the inherent class of y is the same as D_j but is misclassified to other classes in $U - D_j$, which means that y is the class noise, certain gaps will exist when using y to compute $\underline{R_B}(D_j)(x)$. In particular, this nearest neighbour y is regarded as a noisy object relative to the target object x. To lessen the negative influence of the noisy information, an RCU measure with a KNN–Bayes algorithm is proposed to detect the noisy neighbour. First, we introduce how to measure the RCU of a specific object.

Suppose $DT = \langle U, A \cup D \rangle$ is the fuzzy decision system. $D = \{d\}$ is the decision attribute set, and $V_d = \{c_1, c_2, \ldots, c_r\}$ is the domain of the decision attribute d with nominal values, which can also be seen as the class labels in the classification task. $D_j = U/c_j, j = 1, \ldots, r$ is the crisp equivalence class. When computing $\underline{R_B}(D_j)(x)$ ($B \subseteq A$, $x \in U$, suppose y is the nearest neighbour of the target object x), inspired by the uncertainty measure of the margin strategy in active learning [56], the RCU measure of one specific object y proposed in this study is computed as follows:

$$RCU(y) = \left|P(c_i|y) - P\left(c_j|y\right)\right|, y \in U, c_i, c_j \in V_d, c_i \neq c_j, i, j = 1, \ldots, r, \tag{14}$$

where c_i and c_j are nominal values or class labels of the decision attribute d. c_i is the current class label of y, and c_j is the class label used to divide samples into the target equivalence class D_j. $P(c_i|y)$ represents the posterior probability of the object y classified to class label c_i, and $P\left(c_j|y\right)$ represents the posterior probability of the object classified to class label c_j, $c_i \neq c_j$. Evidently, since $P(c_i|y), P\left(c_j|y\right) \in [0, 1]$, $RCU(y) \in [0, 1]$. If the value of $RCU(y)$ is small, then the difference between these two probabilities is slight. In other words, when the classification uncertainty is high and y is easily misclassified to c_j, then y is seen as class noise when computing $\underline{R_B}(D_j)(x)$. Such a noisy object can likely result in mistakes since the presupposition for the fuzzy lower approximation computation is that $y \notin D_j$. This classification uncertainty measure is computed by comparing the classification possibilities of its current class label and the target class label. We thus call it a relative uncertainty measure.

If $RCU(y) < \lambda$, then y is considered a relatively noisy object and should be ignored when computing the lower approximation of D_j. λ is a predefined threshold. Conversely, y is normal and acceptable for the lower approximation computation of D_j.

Next, with the basis of the mckNN algorithm utilized to compute a posterior probability in [57], an improved mckNN combining the KNN algorithm and Bayes theory, named the KNN–Bayes algorithm, is introduced to calculate the above posterior probability.

In the KNN algorithm, the class label of one specific object y is determined by its k nearest objects, denoted as a sorted set $\{y_1, \ldots, y_k\}$. The class labels of these neighbours are represented as a set $z = \left\{c^1, \ldots, c^k\right\}$ $\left(c^l \in V_d, l = 1, \ldots, k\right)$. This class set is then regarded as a new attribute set for the object y. In other words, y is re-represented by z. Thus, the posterior probability of y classified to class c $\left(c = c_i, c_j \in V_d\right)$ can be expressed as $P(c|y) = P(c|z) = P\left(c|c^1, \ldots, c^k\right)$. Later, Bayes decision theory is utilized to estimate the probability of classifying y to c:

$$P(c|y) = P(c|z) = \frac{P(c)P(z|c)}{P(z)}, y \in U, c = c_i, c_j \in V_d. \tag{15}$$

Assume that the KNNs of y are conditional dependent; in other words, the ith nearest neighbour is independent of the previous $(i-1)$ nearest neighbours. Then, the prior probability $P(z|c)$ can be computed as follows:

$$P(z|c) = P\left(c^1, \ldots, c^k|c\right) = P\left(c^1|c\right) \ldots P\left(c^k|c\right). \tag{16}$$

$P(c^l|c)(l = 1, \ldots, k)$ represents the probability that the object y_l with the class label c^l is the lth nearest neighbour of the object with the class label c. Suppose that the number of objects in U with class label c is s, and these objects are denoted as $x_1^c, \ldots, x_s^c$. In finding the k nearest objects of each x_t^c $(t = 1, \ldots, s)$ by the KNN algorithm, these neighbours constitute an $s \times k$ neighbour matrix $M_{s \times k}$:

$$\begin{matrix} x_1^c knn \longrightarrow \\ x_t^c knn \longrightarrow \\ x_s^c knn \longrightarrow \end{matrix} \begin{bmatrix} x'_{11} & x'_{12} & \cdots & x'_{1k} \\ \ldots & \ldots & \ldots & \ldots \\ x'_{s1} & x'_{s2} & \cdots & x'_{sk} \end{bmatrix}.$$

The number of objects with the class label c^l in the lth column in $M_{s \times k}$ is counted, which is denoted as $s_l = \left|\left\{x'_{tl} \in U : c\left[x'_{tl}\right] = c^l, t = 1, \ldots, s\right\}\right|, l = 1, \ldots, k$. Then, $P\left(c^l|c\right)$ is computed as $P\left(c^l|c\right) = s_l/s$. Thus, $P(z|c)$ can be expressed as follows:

$$P(z|c) = \frac{s_1}{s} \times \frac{s_2}{s} \times \cdots \times \frac{s_k}{s} = \prod_{l=1}^{k} \frac{s_l}{s}. \tag{17}$$

The prior probability $P(c)$ can also be estimated from the proportion of the objects labelled with class c among the universe of objects as follows:

$$P(c) = \frac{s}{|U|}. \tag{18}$$

Thus, the posterior probability $P(c|y)$ is estimated as follows:

$$P(c|y) = P(c|z) = \frac{\prod_{l=1}^{k} s_l\left(c^l\right)}{|U| \times s^{k-1} \times P(z)}. \tag{19}$$

$P(z)$ can be regarded as the evidence factor used to normalize $P(c|z)$ as $\sum_{c \in V_d} P(c|z) = 1$.

Example 1. Let $DT = \langle U, A \cup D \rangle$ be the fuzzy decision system, where the domain of the decision attribute set $D = \{d\}$ is $V_d = \{c_1, c_2\}$, and the objects are divided into equivalence classes as $D_1, D_2 = U/c_1, U/c_2$. $B \subseteq A$. $x \in U$ is the target object with class label c_2. To compute $\underline{R_B}(D_2)(x)$, we find the nearest object

of x, denoted as y with class label c_1, which means $y \in D_1$. The next step judges whether y is noisy by the RCU measure. To measure the RCU level of y, $P(c_1|y)$ and $P(c_2|y)$ are computed. First, finding $k = 6$ neighbours of y, the neighbours and their class labels are displayed as $\langle y_1, c_1\rangle$, $\langle y_2, c_1\rangle$, $\langle y_3, c_1\rangle$, $\langle y_4, c_2\rangle$, $\langle y_5, c_1\rangle$, $\langle y_6, c_2\rangle$, $y_1, \ldots, y_6 \in U$, and $c_1, c_2 \in V_d$ from near to far. Therefore, $z = \left\{c_1^1, c_1^2, c_1^3, c_2^4, c_1^5, c_2^6\right\}$ makes up the new attribute set to describe the characteristics of y, where the superscript represents the rank of neighbours. Suppose that the cardinality of U is $|U| = 100$ and the number of objects with class label c_1 in U is $s = 40$. From Eq. (18), we have $P(c_1) = \frac{40}{100} = 0.4$. To compute $P(z|c_1)$, the neighbour matrix $M_{s\times k}$ is constructed by the KNN algorithm. Next, we count the class label distribution in each column of $M_{s\times k}$: in the 1^{th} column, the number of objects with class label c_1 is 36, denoted as $s_1(c_1) = 36$ (which is the variant s_l in Eq. (17)). Similarly, $s_2(c_1) = 35$, $s_3(c_1) = 35$, $s_4(c_2) = 7$, $s_5(c_1) = 30$ and $s_6(c_2) = 9$. Thus, from Eq. (17), we derive $P(z|c_1) = \frac{36\times35\times35\times7\times30\times9}{(40)^6}$. Running the same process, we find the k nearest neighbours of the objects with class label c_2 (the number of these objects is $s = 60$) and construct the corresponding neighbour matrix $M_{s\times k}$. To compute $P(z|c_2)$ in $M_{s\times k}$, we have $s_1(c_1) = 5$, $s_2(c_1) = 6$, $s_3(c_1) = 9$, $s_4(c_2) = 48$, $s_5(c_1) = 12$ and $s_6(c_2) = 40$. Thus, we derive $P(c_2) = \frac{60}{100}$, $P(z|c_2) = \frac{5\times6\times9\times48\times12\times40}{(60)^6}$. In addition, $P(z)$ is used to obtain $P(c_1|y) + P(c_2|y) = 1$, and thus, the value of $P(z)$ is $P(z) = P(c_1)P(z|c_1) + P(c_2)P(z|c_2)$. From Eqs. (14) and (19), we have

$$RCU(y) = |P(c_1|y) - P(c_2|y)| = \left|\frac{P(c_1)P(z|c_1)}{P(c_1)P(z|c_1) + P(c_2)P(z|c_2)} - \frac{P(c_2)P(z|c_2)}{P(c_1)P(z|c_1) + P(c_2)P(z|c_2)}\right| = 0.989 - 0.011 = 0.978.$$

Suppose that the noise detection threshold $\lambda = 0.15$, as $RCU(y) = 0.98 > 0.15$; then, y is thought to be a normal object to compute $\underline{R_B}(D_2)(x)$.

4.2. Robust FRS model

Based on Eq. (14), the RCU value can distinguish the noisy object for estimating the equivalence class. Therefore, we present an RCU measure FRS (RCU FRS) model to enhance the robustness.

Definition 5. Let $DT = \langle U, A \cup D\rangle$ be the fuzzy decision system, $D = \{d\}$, and its domain is $V_d = \{c_1, c_2, \ldots, c_r\}$, $U/D = \{D_1, D_2, \ldots, D_r\}$ and $B \subseteq A$. R_B is the fuzzy similarity relation on U. The fuzzy lower and upper approximations of the RCU FRS model are defined as follows:

$$\underline{R_B^{RCU}}\left(D_j\right)(x) = \inf_{RCU(y)\geq\lambda}\{1 - R_B(x,y)\}, x \in U, j = 1, \ldots, r, \quad (20)$$

$$\overline{R_B^{RCU}}\left(D_j\right)(x) = \sup_{RCU'(y)\geq\lambda}\{R_B(x,y)\}, x \in U, j = 1, \ldots, r \quad (21)$$

where $RCU(y)$ in the lower approximation is calculated as $RCU(y) = \left|P(c_i|y) - P\left(c_j|y\right)\right|$, $c_i \neq c_j$. In the upper approximation, we have $RCU'(y) = \left|P\left(c_j|y\right) - P\left(V_d - c_j|y\right)\right|$, where c_i is the original class label of y, and c_j is the class label used to divide samples into the target equivalence class D_j. λ is a threshold given by users.

In the above definition, y is the nearest object with a class label $c_i\left(c_i \neq c_j\right)$ regarding the target object x when computing the lower approximation of equivalence class D_j with feature subset B. To improve the robustness of the FRS model, we should ensure that y is not a noisy object, which means that y definitely cannot be divided into D_j. To deduce whether y is a noisy object for D_j, we calculate the RCU value of y by $RCU(y) = \left|P(c_i|y) - P\left(c_j|y\right)\right|$, where c_j is the decision class corresponding to dividing the equivalence class D_j from U. If $RCU(y) < \lambda$, then y is considered relatively noisy and will be abandoned for the lower approximations of D_j. Next, the second nearest object of the target object x is the candidate to conduct a lower approximation. Conversely, y is a normal object. This notion means that the RCU FRS model ignores the objects that have high RCU levels when calculating the fuzzy approximations.

Next, we present how to construct the above definition. First, according to Eq. (2), we define the membership function of a specific object y to the equivalence class D_j under the RCU measure in the lower and upper approximations:

For the lower approximation,

$$D_j^{RCU}(y) = \begin{cases}1, RCU(y) < \lambda\\ 0, RCU(y) \geq \lambda\end{cases}, RCU(y) = \left|P(c_i|y) - P\left(c_j|y\right)\right|, \quad j = 1, \ldots, r. \quad (22)$$

For the upper approximation,

$$D_j^{RCU}(y) = \begin{cases}1, RCU'(y) \geq \lambda\\ 0, RCU'(y) < \lambda\end{cases}, RCU'(y) = \left|P\left(c_j|y\right) - P\left(V_d - c_j|y\right)\right| j = 1, \ldots, r. \quad (23)$$

For Eq. (22), the RCU level measured for the fuzzy lower approximation is used to detect whether object y has a large possibility of being classified into class D_j. The original class label for y is $c_i(c_i \neq c_j)$, which means $y \notin D_j$ currently. If $RCU(y) \geq \lambda$, it is thought to be a normal object and will not be misclassified to the equivalence class D_j, which holds $y \notin D_j$; therefore, $D_j^{RCU}(y) = 0$. Otherwise, y is noisy and easily misclassified into D_j, which means $y \in D_j$, and we derive $D_j^{RCU}(y) = 1$. For Eq. (23), $P(V_d - c_j|y)$ represents the possibility of y being classified to the arbitrary class in $V_d - c_j$. The RCU level measure for fuzzy upper approximation is used to detect whether y has a large possibility of not being classified to the equivalence class D_j. The original class label for y is c_j. If $RCU'(y) \geq \lambda$, we consider that it is a normal object and is actually in the equivalence class D_j, which is $y \in D_j$, and we derive $D_j^{RCU}(y) = 1$. Otherwise, y is noisy and has a large possibility of not being classified into D_j, which is $y \notin D_j$; therefore, $D_j^{RCU}(y) = 0$.

Based on the initial definition form of the lower and upper approximations, the definition considering the noise influence is presented as follows:

$$\begin{aligned}\underline{R_B^{RCU}}\left(D_j\right)(x) &= \inf_{y\in U}\max\left\{1 - R_B(x,y), D_j^{RCU}(y)\right\}\\ &= \inf_{RCU(y)<\lambda}\max\left\{1 - R_B(x,y), D_j^{RCU}(y)\right\} \wedge \inf_{RCU(y)\geq\lambda}\max\{1 - R_B(x,y), D_j^{RCU}(y)\}\\ &= \inf_{RCU(y)<\lambda}\max\{1 - R_B(x,y), 1\} \wedge \inf_{RCU(y)\geq\lambda}\max\{1 - R_B(x,y), 0\}'\\ &= 1 \wedge \inf_{RCU(y)\geq\lambda}\{1 - R_B(x,y)\}\\ &= \inf_{RCU(y)\geq\lambda}\{1 - R_B(x,y)\}\end{aligned} \quad (24)$$

$$\begin{aligned}\overline{R_B^{RCU}}\left(D_j\right)(x) &= \sup_{y\in U}\min\left\{R_B(x,y), D_j^{RCU}(y)\right\}\\ &= \sup_{RCU'(y)\geq\lambda}\min\left\{R_B(x,y), D_j^{RCU}(y)\right\} \vee \sup_{RCU'(y)<\lambda}\min\{R_B(x,y), D_j^{RCU}(y)\}\end{aligned}$$

$$= \sup_{RCU'(y)\geq\lambda} \min\{R_B(x,y), 1\} \vee \sup_{RCU'(y)<\lambda} \min\{R_B(x,y), 0\}$$
$$= \sup_{RCU'(y)\geq\lambda} \{R_B(x,y)\} \vee 0 \quad . \tag{25}$$
$$= \sup_{RCU'(y)\geq\lambda} \{R_B(x,y)\}$$

Based on the above discussion, each target object x has its own noisy objects. However, noisy objects may not be detected as noisy data when computing the fuzzy approximations of other equivalence classes for the same target object. Therefore, these noisy objects detected by the proposed RCU FRS model are called relatively noisy objects. In addition, a common data preprocessing step is also used to detect the noisy data, which are the absolute noisy data detected from the whole set. Finally, the RCU FRS model recognizes the relatively noisy objects to reduce the negative influence of estimating the lower and upper approximations.

Moreover, one related proposition for the definition of RCU FRS is as follows:

Proposition 1. *Let $DT = \langle U, A \cup D\rangle$ be the fuzzy decision system and $B \subseteq A$, $x \in U$, $U/D_j = \{D_1, \ldots, D_r\}$. Thus, the following statement holds.*

$\forall x \in U, D_j \in U/D$, $\underline{R_B^{RCU}}(D_j)(x) \supseteq \underline{R_B}(D_j)(x)$, $\overline{R_B^{RCU}}(D_j)(x) \subseteq \overline{R_B}(D_j)(x)$.

Proof. Combining the standard definition of FRS, we have the following:

$$\underline{R_B^{RCU}}(D_j)(x) = \inf_{RCU(y|D_j)\geq\lambda} \{1 - R_B(x,y)\} =$$
$$\inf_{|P(c(y)|y)-P(c_j|y)|\geq\lambda} \{1 - R_B(x,y)\} \geq \inf_{y\notin D_j} \{1 - R_B(x,y)\} = \underline{R_B}(D_j)(x),$$

$$\overline{R_B^{RCU}}(D_j)(x) = \sup_{RCU'(y|U-D_j)\geq\lambda} \{R_B(x,y)\}$$
$$= \sup_{|P(c(y)|y)-P(c_i|y)|\geq\lambda, c_i\neq c_j} \{R_B(x,y)\} \leq \sup_{y\in D_j} \{R_B(x,y)\} = \overline{R_B}(D_j)(x).$$

Thus, $\underline{R_B^{RCU}}(D_j)(x) \supseteq \underline{R_B}(D_j)(x)$ and $\overline{R_B^{RCU}}(D_j)(x) \subseteq \overline{R_B}(D_j)(x)$ hold.

Based on the fuzzy lower approximation in the RCU FRS model, the relative fuzzy positive domain of the concerned feature subset B can be computed as $POS'_B(D) = \bigcup_{j=1}^{r} \underline{R_B^{RCU}}(D_j)$. Hereafter, the relative fuzzy dependency function is defined as $\gamma'_B(D) = \sum_{x\in U} POS'_B(D)(x) / |U|$. Clearly, the relative fuzzy dependency function also follows the monotonically increasing property of the fuzzy dependency function mentioned in Properties 1–3.

5. Heuristic algorithm for feature selection

Feature selection aims to find a minimal subset with the same or superior classification performance as the extensive feature set. Although multiple reducts can be searched from the given data set, finding one of them in the classification learning tasks with the best classification performance is sufficient. Based on the discussions of the monotonically increasing properties of the fuzzy dependency function, we construct an iterative heuristic algorithm for feature selection. Considering the possibility of falling into a local optimum situation, we thus investigate heuristic forward and backward algorithms for feature searching. On the one hand, the forward heuristic algorithm selects one feature each time with the greatest significance into a reduct pool. According to Definitions 3 and 4, this algorithm ends when the marginal increase of the fuzzy dependency value is lower than the marginal loss quantity utility caused by adding one unit feature. However, the heuristic backward algorithm starts to drop one feature each time with the slightest significance feature. The algorithm ends when the marginal decrease of the fuzzy dependency value is larger than the marginal growth quantity utility caused by decreasing one unit feature. We compare the different results derived by these two algorithms and select the better performance as the selection result. Two algorithms are presented as follows:

Suppose that there are m objects in the universe U, n dimensions of the conditional features and the r decision equivalence classes are divided by the decision features. For the forward algorithm, the number of loops for Steps 3–11 is n, the number of loops for Steps 4–8 is r and the number of loops for Steps 5–7 is m. Therefore, the total number of loops for the forward algorithm is $n \times r \times m$. In the worst case, the time complexity of the forward algorithm is $O(mnr)$. In summary, the backward algorithm has the same running mechanism and the time complexity is the same. We run these two algorithms together, and the final time complexity for the proposed feature searching algorithm is $O(mnr)$.

The framework diagram is depicted in Fig. 1.

6. Experiments

We evaluate the operational utility of our proposed model in four parts. First, we compare the proposed RCU FRS model against other models grounded in feature selections with normal data sets. Our second experimental evaluation randomly adds class noise to the data sets to demonstrate that the proposed model outperforms existing models in the robustness level. Third, the effectiveness of the RBF-based similarity measure with mixed distance metrics is likewise evaluated by comparing it with three other kernel functions. Finally, the statistical analysis shows that the proposed model significantly improves the other competing methods in terms of classification accuracies.

6.1. Data sets and experimental procedures

To evaluate the proposed model, we use 15 data sets collected from the UCI machine learning repository, encompassing single-type and mixed-type features as well as binary and multiple classes. Table 1 depicts basic descriptive traits for each data set.

In our experiments, we compare the RCU model against some existing FRS models: combinational FRSs (CFRS) [27], algorithm based on variable distance parameter (AVDP) FRS [34] and DC_ratio FRS [50]. Meanwhile, to evaluate the impact of the RBF kernel-based similarity measure, we run experiments where the RBF kernel is replaced by the linear [35], wavelet [36] and Hermite kernels [37]. Two popular classifiers are utilized to assess the features selected from each model, namely, support vector machine (SVM) and KNN rule ($K = 3$). Moreover, we set $\delta = 1$ and $k = 6$ in the experiments. To make the evaluation more robust, a 10-fold cross-validation process is invoked where the data set is randomly divided into 10 parts and the training–testing procedure is performed 10 times in turn. The mean classification accuracies and corresponding standard deviations, number of optimal selected features and mean ranks are used as the baseline metrics to evaluate the models.

6.2. The impact of parameters θ and λ

θ is the stop condition of the feature selection algorithm and can be seen as the marginal utility of adding or dropping one

Forward algorithm: A heuristic forward algorithm based on the RCU FRS model

Input: $DT=\langle U, A\cup D\rangle$, threshold λ and marginal feature quantitative value factor θ, parameter value δ in the RBF kernel.

Output: A feature reduct *reduct*

1. $reduct \leftarrow \varnothing$, $B \leftarrow A - reduct$, $start \leftarrow 1$;
2. **while** *start* **do**
3. **for** $t \leftarrow 1$ to $|B|$ **do**
4. **for** $j \leftarrow 1$ to $|U/D|$ **do**
5. **for** $i \leftarrow 1$ to $|U|$ **do**
6. Compute the fuzzy lower approximation $\underline{R^{RCU}_{reduct\cup a_t}}(D_j)(x_i)$ by Eq. (20);
7. **end for**
8. **end for**
9. Compute the fuzzy dependency membership as follows:
10. $\gamma'_{reduct\cup a_t}(D)=\left(\max\limits_{D_j\in U/D}\left\{\sum_{i=1}^{|U|}\underline{R^{RCU}_{reduct\cup a_t}}(D_j)(x_i)\right\}\right)\Big/|U|$;
11. **end for**
11. Select the feature a_t so that $\gamma'_{reduct\cup a_t}(D)$ has the maximum value;
12. **if** $\gamma'_{reduct\cup a_t}(D)-\left(\theta\cdot(|reduct|+1)\right)/\left(N\cdot|U|\right)\geq\gamma'_{reduct}(D)-\left(\theta\cdot(|reduct|)\right)/\left(N\cdot|U|\right)$ and $B\neq\varnothing$ **do**
13. $reduct \leftarrow reduct\cup a_t$, $B \leftarrow A - reduct$;
14. **else do**
15. $start \leftarrow 0$;
16. **end if**
17 **end while**
18. Return *reduct*.

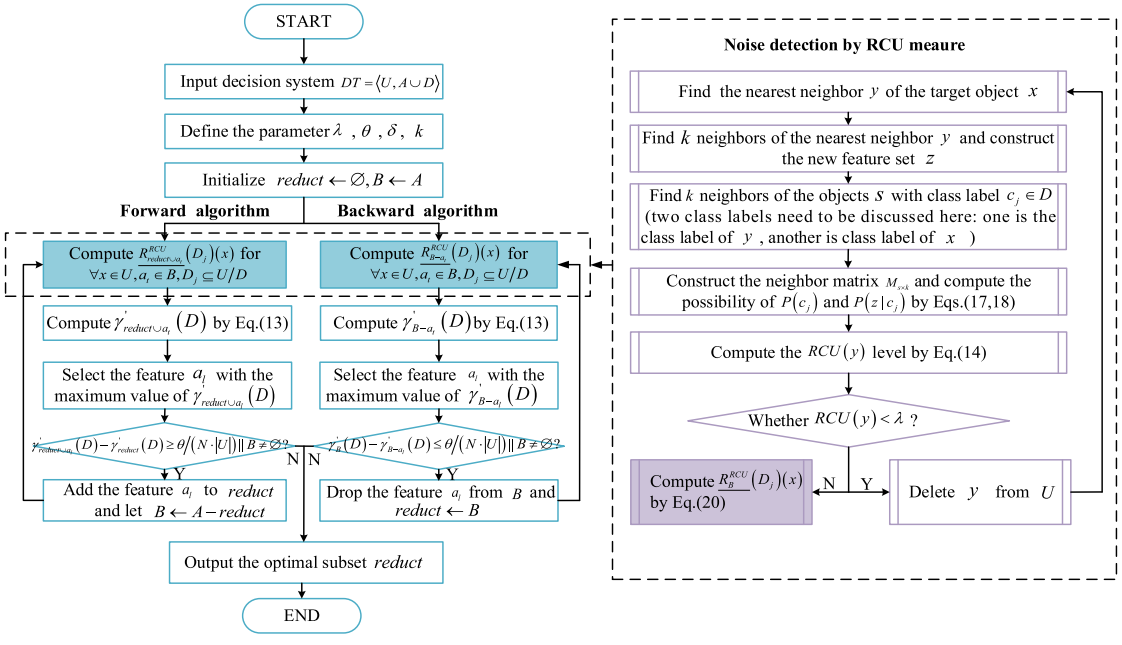

Fig. 1. Framework of the proposed robust FRS-based feature selection model.

Backward algorithm: A heuristic backward algorithm based on the RCU FRS model

Input: $DT=\langle U, A\cup D\rangle$, threshold λ and marginal feature quantitative value factor θ, parameter value δ in the RBF kernel.

Output: A feature reduct *reduct*

1. $reduct \leftarrow \varnothing$, $B \leftarrow A$, $start \leftarrow 1$;
2. **while** *start* **do**
3. **for** $t \leftarrow 1$ to $|B|$ **do**
4. **for** $j \leftarrow 1$ to $|U/D|$ **do**
5. **for** $i \leftarrow 1$ to $|U|$ **do**
6. Compute the fuzzy lower approximation $\underline{R^{RCU}_{B-a_t}}(D_j)(x_i)$ by Eq. (20);
7. **end for**
8. **end for**
9. Compute the fuzzy dependency membership as follows:
10. $\gamma'_{B-a_t}(D)=\left(\max\limits_{D_j\in U/D}\left\{\sum_{i=1}^{|U|}\underline{R^{RCU}_{B-a_t}}(D_j)(x_i)\right\}\right)\Big/|U|$;
11. **end for**
11. Select the feature a_l so that $\gamma'_{B-a_l}(D)$ has the maximum value;
12. **if** $\gamma'_{B-a_l}(D)-\left(\theta\cdot\left(|B|-1\right)/\left(N\cdot|U|\right)\right)\geq\gamma'_{B}(D)-\left(\theta\cdot\left(|B|\right)/\left(N\cdot|U|\right)\right)$ and $B\neq\varnothing$ **do**
13. $B\leftarrow B-a_l$, $reduct\leftarrow B$;
14. **else do**
15. $start\leftarrow 0$;
16. **end if**
17 **end while**
18. Return *reduct*.

Table 1
Basic descriptive statistics of the 15 adopted data sets.

No.	Data sets	Sample	Features (Symbolic and real-valued)	Classes
1	Wine	178	13 (0 & 13)	3
2	Heart	270	13 (8 & 5)	2
3	Forestfire	244	10 (0 & 10)	2
4	Hepatitis	155	18 (13 & 5)	2
5	Ionos	351	34 (2 & 32)	2
6	Gamma	19020	10 (0 & 10)	2
7	Credit	690	15 (9 & 6)	2
8	German	1000	24 (21 & 3)	2
9	Sonar	208	60 (0 & 60)	2
10	Wdbc	569	30 (0 & 30)	2
11	Wpbc	198	33 (0 & 33)	2
12	Parkinson	240	46 (2 & 42)	2
13	Movement	360	90 (0 & 90)	16
14	Urban land cover	675	147 (0 & 147)	9
15	Obesity	2111	16 (3 &13)	7

feature. λ is the threshold to detect noise objects. The original data set is randomly divided into 10 subsets, where 9 serve as the training set and the remaining 1 is utilized as the test set. In the training phase, inspired by the grid searching algorithm, the parameters λ and θ are determined by exhaustively examining the following values: λ: [0, 0.5] with a step of 0.05 and θ: [0, $|A|$] with a step of 5, where $|A|$ represents the cardinality of the full features in one specific data set. The corresponding optimal feature subset is selected. In the testing phase, the reduced test set is sent to the classifier, and the feature subset with the best average classification accuracy (Acc) is chosen as the final result. The aforementioned three traits are computed at the same time.

To visually show the impact of λ and θ, the performance of the proposed RCU model with multigranularity parameters is displayed in Fig. 2. To vividly exhibit the varied results, the examples use three data sets encompassing Hepatitis, Ionos and Sonar with 20% class noise joined randomly, where the classifier is SVM. From Fig. 2, we can easily find that Acc varies with different λ and θ. For certain data sets, changes in λ and θ retain small variations in the Acc, such as Hepatitis. In contrast, the performance of some data sets is affected largely by the parameters λ and θ, such as Ionos and Sonar. The likely reason is that the performance of parameter values is related to the distribution and diversity of data. The lack of data diversity results in a smooth fluctuation for various parameters. Diversity-rich data behave entirely the opposite, where distinct results can be found in a range of parameter values. Furthermore, Figs. 3 and 4 are depicted to analyse λ and θ separately.

From Fig. 3, the parameter of λ is taken from 0 to 0.5 by a step of 0.05, and the result is shown. Despite the increase and decrease

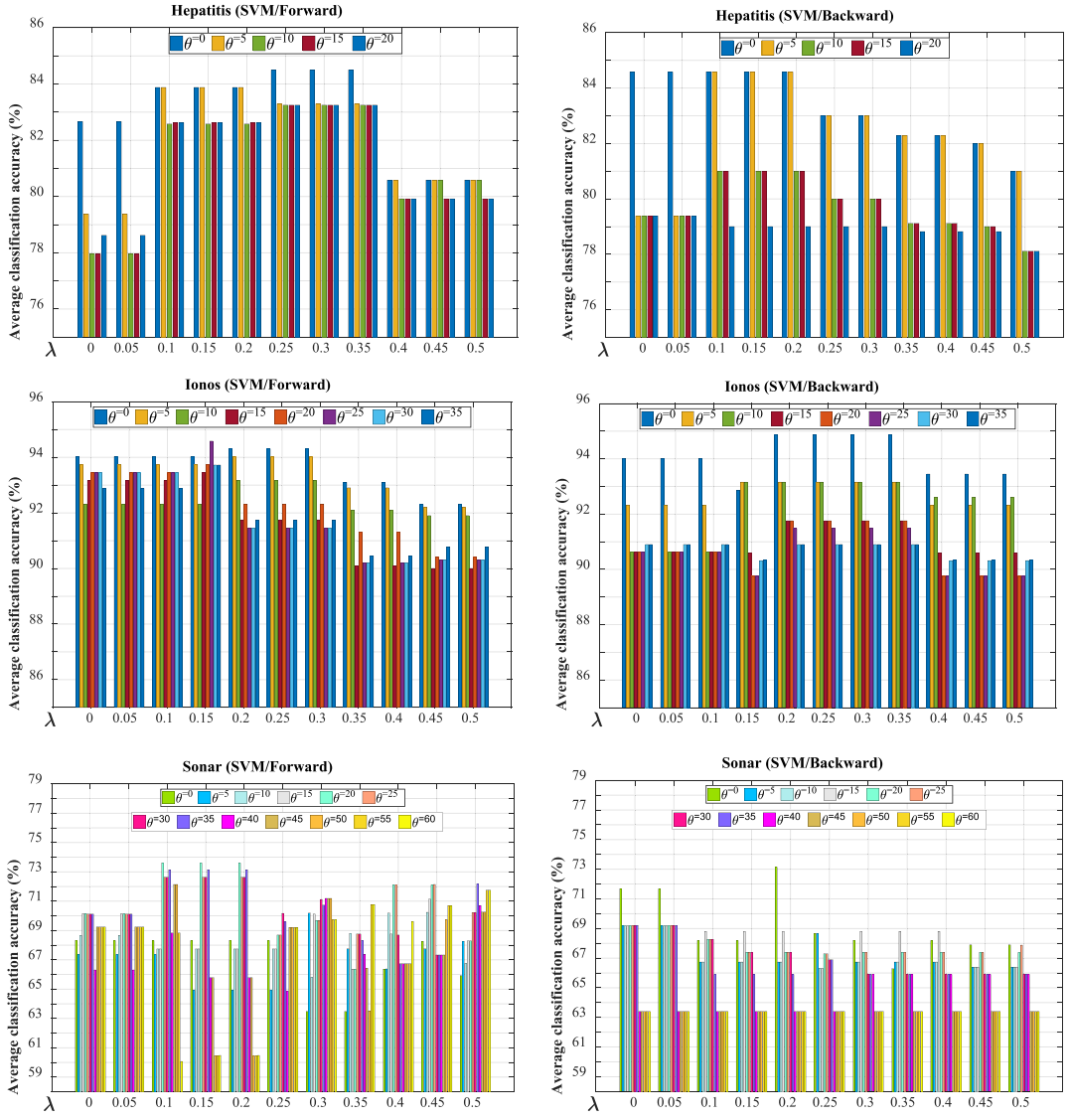

Fig. 2. Average classification accuracies varying with parameters λ and θ.

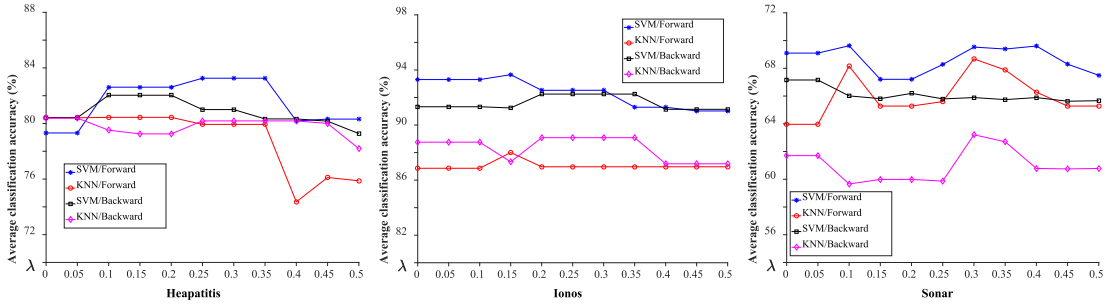

Fig. 3. Average classification accuracies varying with parameter λ and classifiers.

in classification accuracies in local intervals, the global trend of the curve is downwards. Three phases can explain this situation. When λ is too small, more noise objects are overlooked, and the RCU model does not work. When λ increases, noisy objects are detected in the fuzzy approximation computation process. Thus, the selected features improve the classification result. When λ increases to be too large, normal objects may be incorrectly filtered out, which leads to a lower performance in the selected feature subsets. In particular, different data sets have different traits, and the optimal value of λ is distinctive. From Fig. 4, the parameter θ is taken from 0 to the number of its full features with a step of 5. The increasing and decreasing trends of classification

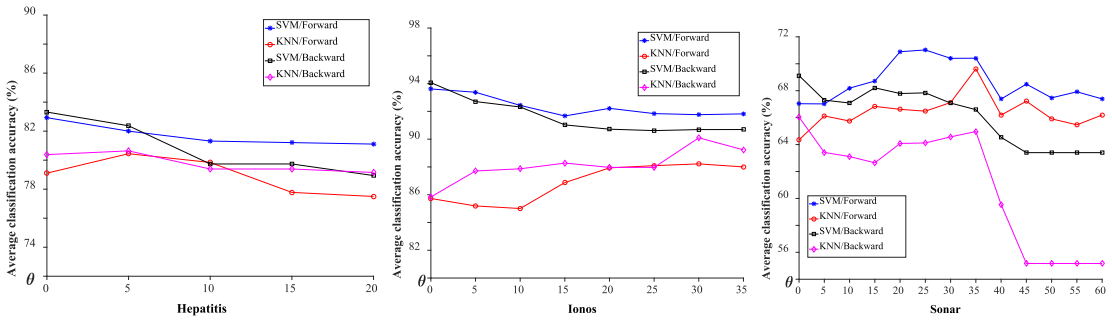

Fig. 4. Average classification accuracies varying with parameter θ and classifiers.

Table 2
Number of optimal selected features with normal data sets.

	Raw data	CFRS		DC_Ratio		AVDP		RCU	
		SVM	KNN	SVM	KNN	SVM	KNN	SVM	KNN
Wine	13	12	11	**6**	**6**	12	12	8	12
Heart	13	9	8	8	**7**	9	9	**7**	9
Forestfire	10	4	**5**	**2**	**2**	**2**	**2**	**2**	**2**
Hepatitis	18	12	**5**	12	11	**5**	**5**	6	11
Ionos	34	26	19	19	9	8	8	25	**5**
Gamma	10	9	7	7	7	8	**6**	9	9
Credit	15	6	9	**5**	**5**	13	13	**5**	**5**
German	24	**10**	14	13	14	22	22	11	14
Sonar	60	39	48	46	46	16	**8**	10	10
Wdbc	30	18	18	18	23	7	7	8	**6**
Wpbc	33	15	18	27	27	12	12	**7**	**7**
Parkinson	46	19	19	18	23	11	13	**3**	6
Movement	90	**37**	**37**	74	68	41	82	73	59
Urban land cover	147	86	86	105	96	119	142	**27**	87
Obesity	16	15	15	14	14	15	15	**7**	**7**
Average length	32.27	21.33	21.26	24.93	23.87)	20	23.73	**13.87**	**16.60**
(Mean Rank)	(-)	(5.43)	(4.93)	(4.76)	(4.67)	(4.70)	(4.70)	**(3.2)**	**(3.6)**

Note. The strategy for assigning rankings: items that compare equally receive the same ranking number, which is the mean of what they would have under ordinal rankings. The mean rank is the average value of the rankings in each column. The best result of the row is in bold and underscored.

accuracies caused by θ are similar to λ, and the reason for the changes is similar. The result shows that the RCU models can maintain high accuracies over a wide value range of θ.

6.3. The impact of searching algorithms and classifiers

Reflected from Figs. 3 and 4, the forward and backward searching algorithms show different performances when setting the same λ or θ. Since two searching algorithms start in different positions, one starts to add features one by one with the largest significance and the other starts to drop features one by one with the least significance, running with two initialization searching positions can avoid getting stuck in a local optimum situation. The better result is used as the final result. Moreover, from Figs. 3 and 4, the RCU FRS model always achieves a better result in the SVM than in the KNN implementation. Two reasons can be summarized: (1) The RCU FRS model can effectively recognize the noise objects where they are located around the classification boundary. Removing the detected noise objects by the proposed model relatively enlarges the margin between classes, which is consistent with the maximum classification margin principle in SVMs. (2) KNN is a lazy clustering process. Thus, it is normal to have a relatively poor performance when the classification boundary is linearly inseparable.

6.4. Analysis of results

Three comparison experiments are conducted. The first compares the FRS models in the normal data sets, and the second adds 20% class noise randomly in the original data sets to verify the robustness of the FRS models. The last verifies the effectiveness of the RBF-based similarity measure by altering the kernel functions in the RCU FRS model.

6.4.1. Classification performance comparison

Table 2 shows the number of optimal selected features in the four FRS models with two classifiers. Tables 3 and 4 denote the comparative classification performance for each model. All of them are conducted in normal data sets. From Table 2, the proposed RCU FRS model derives both the optimal and suboptimal average length of the selected features and the mean ranks. From Tables 3 and 4, we see that the RCU FRS model achieves higher classification accuracies than other models with SVM and KNN classifiers. With SVM, we can observe that the classification accuracy of the RCU FRS model is superior on 10 out of 15 data sets. With KNN, the performance of the proposed model is not in an obvious dominance position. Although only 5 out of 15 data sets are superior, the average classification accuracy and its mean rank are still the best.

In summary, the AVDP FRS model obtains the number of selected features, and the average rank is slightly inferior to that of the proposed model. However, it obtains the worst Acc. The reason is that its attribute reduction granularity is too fine, which harms the classification accuracy. The CFRS and DC_ratio FRS models yield similar results since they both employ the sample pair selection strategy for feature selection, where multiple intersection operations are used. Instead, we use the RBF-based similarity measure with mixed normalized distance metrics to

Table 3
Comparison of classification accuracies of reduced data with SVM.

	Raw data	CFRS	DC_ratio	AVDP	RCU
Wine	98.33 ± 2.55	98.33 ± 2.55	98.33 ± 2.55	97.19 ± 2.81	97.22 ± 3.73
Heart	77.04 ± 4.91	77.41 ± 4.52	77.41 ± 4.52	77.41 ± 4.52	77.41 ± 5.09
Forestfire	90.22 ± 5.11	90.62 ± 5.43	92.67 ± 2.94	91.03 ± 5.11	92.67 ± 4.30
Hepatitis	83.17 ± 6.79	82.5 ± 7.95	82.5 ± 7.95	83.17 ± 5.32	84.58 ± 4.04
Ionos	93.44 ± 5.12	93.44 ± 5.12	94.02 ± 4.12	94.02 ± 5.02	94.30 ± 5.27
Gamma	85.91 ± 0.86	85.69 ± 0.99	86.11 ± 0.80	85.69 ± 0.99	85.55 ± 0.96
Credit	84.49 ± 16.96	85.35 ± 17.40	85.22 ± 17.32	84.35 ± 16.70	85.51 ± 17.49
German	75.40 ± 3.29	75.10 ± 2.12	76.10 ± 3.11	75.00 ± 3.22	75.80 ± 3.12
Sonar	66.90 ± 10.64	65.83 ± 15.52	68.29 ± 11.44	65.43 ± 14.30	73.48 ± 17.94
Wdbc	97.71 ± 2.23	97.89 ± 2.05	97.89 ± 2.05	97.18 ± 1.97	97.18 ± 1.97
Wpbc	80.32 ± 4.67	80.87 ± 4.74	80.34 ± 5.14	79.82 ± 4.52	82.87 ± 5.00
Parkinson	80.83 ± 10.74	80.00 ± 10.67	80.00 ± 9.82	79.17 ± 12.64	82.08 ± 9.87
Movement	78.89 ± 10.48	79.72 ± 10.25	79.44 ± 10.48	78.61 ± 10.02	80.83 ± 10.20
Urban land cover	82.97 ± 5.10	84.30 ± 4.95	83.42 ± 4.60	83.56 ± 4.59	82.97 ± 3.17
Obesity	68.98 ± 9.59	68.37 ± 9.41	70.44 ± 2.93	68.37 ± 9.41	73.67 ± 9.39
Acc	82.97 ± 6.60	83.03 ± 6.91	83.47 ± 5.98	82.67 ± 6.74	84.41 ± 6.77
(Mean rank)	(3.46)	(2.93)	(2.43)	(3.96)	**(2.21)**

Note: $\alpha \pm \beta$ represents that α is the average classification accuracy and β is the relative standard deviation produced in the 10-fold cross-validation.

Table 4
Comparison of classification accuracies of reduced data with KNN.

	Raw data	CFRS	DC_ratio	AVDP	RCU
Wine	94.90 ± 4.00	96.60 ± 3.73	96.60 ± 3.73	96.05 ± 4.45	96.05 ± 4.45
Heart	78.52 ± 7.73	79.63 ± 7.27	79.63 ± 7.27	78.89 ± 10.08	80.00 ± 9.10
Forestfire	86.95 ± 7.82	89.43 ± 7.40	93.08 ± 3.55	88.58 ± 6.16	91.90 ± 5.59
Hepatitis	82.38 ± 11.22	79.92 ± 6.97	80.38 ± 13.13	81.17 ± 10.18	81.21 ± 7.01
Ionos	84.33 ± 7.58	86.03 ± 5.79	90.03 ± 5.88	92.31 ± 5.72	91.75 ± 4.81
Gamma	83.23 ± 1.01	82.91 ± 1.01	83.60 ± 0.91	82.97 ± 1.11	83.11 ± 1.18
Credit	84.78 ± 10.48	84.49 ± 11.00	82.75 ± 12.74	84.93 ± 11.08	82.32 ± 15.78
German	72.50 ± 4.83	73.80 ± 3.22	73.8 ± 3.22	71.80 ± 3.68	72.60 ± 2.69
Sonar	63.83 ± 14.22	62.48 ± 17.28	62.40 ± 15.63	63.93 ± 13.73	72.55 ± 15.10
Wdbc	97.01 ± 2.23	97.01 ± 2.51	97.36 ± 1.80	96.30 ± 2.01	96.31 ± 3.09
Wpbc	73.71 ± 8.15	78.76 ± 5.58	74.16 ± 7.66	76.76 ± 2.47	77.32 ± 3.11
Parkinson	78.33 ± 13.02	78.75 ± 9.94	78.33 ± 10.67	74.17 ± 10.00	80.42 ± 10.04
Movement	78.06 ± 11.55	78.06 ± 13.00	78.89 ± 11.47	77.22 ± 13.02	79.44 ± 12.86
Urban land cover	78.06 ± 5.99	77.92 ± 4.35	79.10 ± 4.36	77.91 ± 6.08	76.86 ± 6.51
Obesity	76.28 ± 8.14	75.66 ± 8.89	75.99 ± 8.39	75.99 ± 8.39	88.93 ± 11.87
Acc	80.86 ± 7.86	81.43 ± 7.20	81.74 ± 7.36	81.27 ± 7.21	83.38 ± 7.55
(Mean rank)	(3.36)	(3.10)	(2.57)	(3.53)	**(2.43)**

compute the fuzzy relations in fuzzy approximation. The higher classification accuracies demonstrate that the discrimination ability of the fuzzy relations produced by the RBF-kernel-based similarity measure is superior to the intersection operations in the CFRS and DC_ratio models. From Tables 3 and 4, we find that many data sets derive the same or similar classification accuracies among the five listed FRS models. Since the data sets collected from UCI are defaulted to be noise-free, the five models have the same nearest object and the same lower approximation with regard to the target object. Therefore, we later add noise randomly to display the relatively large differences among these models and demonstrate the robustness of the proposed model.

6.4.2. Robustness analysis

In this section, experiments are executed to evaluate the robustness against noise based on different FRS models for feature selection. We assume that the data collected from UCI are noise-free. A total of 20% class noise is randomly added in the feature selection process to each data set. (Noisy objects exist only in the feature selection process, and the data used in the classification process are normal.) To make the objectivity of the evaluation promising, we add random class noise 10 times and run the experiments 10 times accordingly. Finally, the average accuracies of 10 experiments are derived for evaluation.

Table 5 represents the number of optimal feature sets selected from different FRS models on noisy data sets with SVM and KNN. Tables 6 and 7 show the corresponding classification accuracies. Fig. 5 depicts the comparison of performance with and without class noise. Table 5 shows that the proposed model produces the optimal average length as well as the optimal and suboptimal mean ranks. The suboptimal average length is yielded by the CFRS model, but its mean rank value is in the fifth position out of eight in total. From Tables 6 and 7, the RCU FRS model obtains higher Acc than other models with SVM and KNN, and its mean rank value is also the best. With SVM in Table 6, the classification accuracies of 11 out of 15 noisy data sets perform better. Similarly, 12 out of 15 noisy data sets achieve better classification accuracies with KNN in Table 7. From Fig. 5, since raw data are classified with no feature selection and noisy objects are only added in the feature selection process, no change occurs in the two cases. Furthermore, we can observe that the other four FRS models are affected by noisy data sets: the average classification accuracies of SVM and KNN implementations decrease when noisy objects are randomly added. However, the classification performance of the RCU FRS model involves the mildest change after adding the class noise. In contrast, the CFRS model undergoes the largest change, the DC_ratio FRS model experiences the second lowest change and the AVDP model has the middle performance change.

Table 5
Number of optimal selected features with noisy data sets.

	Raw data	CFRS		DC_Ratio		AVDP		RCU	
		SVM	KNN	SVM	KNN	SVM	KNN	SVM	KNN
Wine	13	**3**	**3**	6	6	10	8	10	10
Heart	13	9	8	8	7	9	**5**	7	9
Forestfire	10	4	5	**2**	**2**	5	**2**	**2**	**2**
Hepatitis	18	12	**5**	12	11	**5**	**5**	6	6
Ionos	34	16	16	19	9	8	8	21	**5**
Gamma	10	**6**	9	7	7	**6**	8	**6**	**6**
Credit	15	6	9	**5**	**5**	10	13	**5**	**5**
German	24	**11**	14	14	13	22	22	15	15
Sonar	60	40	39	46	46	16	8	**5**	13
Wdbc	30	17	17	15	15	**7**	**7**	12	12
Wpbc	33	17	15	27	27	12	12	**9**	**9**
Parkinson	46	17	18	18	23	**11**	13	12	12
Movement	90	40	41	80	80	67	79	80	**11**
Urban land cover	147	67	**41**	113	113	118	87	107	107
Obesity	16	15	15	14	14	15	15	**8**	**8**
Average length	32.27	18.67	**17.00**	25.73	25.20	21.40	19.47	20.33	**15.33**
(Mean Rank)	(-)	(4.73)	(4.87)	(5.4)	(4.97)	(4.77)	(4.13)	**(3.8)**	**(3.37)**

Table 6
Comparison of classification accuracies of reduced data with SVM (20% noisy objects randomly added).

	Raw data	CFRS	DC_ratio	AVDP	RCU
Wine	98.33 ± 2.55	91.67 ± 9.70	96.67 ± 5.09	96.69 ± 3.69	98.33 ± 2.55
Heart	77.04 ± 4.91	77.04 ± 4.91	77.04 ± 4.91	76.67 ± 4.40	77.41 ± 4.52
Forestfire	90.22 ± 5.11	90.20 ± 5.52	91.85 ± 4.72	91.03 ± 5.00	92.67 ± 4.30
Hepatitis	83.17 ± 6.79	81.83 ± 9.40	82.50 ± 7.95	81.21 ± 7.62	84.58 ± 4.04
Ionos	93.44 ± 5.12	92.88 ± 5.45	92.02 ± 5.96	92.89 ± 5.28	94.87 ± 3.07
Gamma	85.91 ± 0.86	85.97 ± 0.83	86.11 ± 0.80	85.62 ± 0.85	85.67 ± 0.89
Credit	84.49 ± 16.96	65.77 ± 7.46	85.55 ± 17.67	85.55 ± 17.67	85.55 ± 17.67
German	75.40 ± 3.29	75.00 ± 2.53	76.10 ± 3.11	75.40 ± 3.29	75.80 ± 3.12
Sonar	66.90 ± 10.64	63.00 ± 17.25	68.33 ± 9.17	66.76 ± 18.32	72.62 ± 10.59
Wdbc	97.71 ± 2.23	97.71 ± 1.93	94.73 ± 3.23	96.13 ± 2.34	97.54 ± 1.96
Wpbc	80.32 ± 4.67	78.32 ± 5.49	80.34 ± 5.14	80.34 ± 4.06	81.37 ± 4.34
Parkinson	80.83 ± 10.74	79.58 ± 11.25	79.58 ± 9.94	77.50 ± 12.1	80.00 ± 11.46
Movement	78.89 ± 10.48	79.17 ± 10.34	79.17 ± 10.56	79.44 ± 9.15	80.56 ± 4.03
Urban land cover	82.97 ± 5.10	84.15 ± 4.37	83.86 ± 4.36	83.71 ± 4.12	83.26 ± 3.64
Obesity	68.98 ± 9.59	68.37 ± 9.41	68.75 ± 9.24	68.37 ± 9.41	72.21 ± 9.16
Acc	82.97 ± 6.60	80.71 ± 7.06	82.84 ± 6.79	82.48 ± 7.15	84.16 ± 5.69
(Mean rank)	(2.97)	(3.84)	(2.78)	(3.69)	**(1.72)**

Table 7
Comparison of classification accuracies of reduced data with KNN (20% noisy objects randomly added).

	Raw data	CFRS	DC_ratio	AVDP	RCU
Wine	94.90 ± 4.00	90.56 ± 10.26	94.67 ± 6.31	96.08 ± 3.57	96.60 ± 3.83
Heart	78.52 ± 7.73	79.63 ± 7.27	79.63 ± 7.27	78.52 ± 8.57	80.00 ± 9.21
Forestfire	86.95 ± 7.82	88.20 ± 6.29	92.27 ± 3.72	88.58 ± 6.16	91.90 ± 5.59
Hepatitis	82.38 ± 11.22	79.08 ± 12.00	80.38 ± 13.13	79.87 ± 7.24	83.88 ± 9.38
Ionos	84.33 ± 7.58	83.46 ± 5.72	90.03 ± 5.88	88.03 ± 5.24	91.90 ± 5.59
Gamma	83.23 ± 1.01	83.56 ± 0.85	83.60 ± 0.91	82.75 ± 0.90	82.97 ± 1.11
Credit	84.78 ± 10.48	64.17 ± 8.35	79.15 ± 15.36	81.48 ± 14.07	81.77 ± 13.74
German	72.50 ± 4.83	72.70 ± 2.83	70.90 ± 1.92	72.20 ± 4.83	72.70 ± 2.69
Sonar	63.83 ± 14.22	59.02 ± 15.39	59.52 ± 15.93	56.14 ± 19.72	71.60 ± 10.89
Wdbc	97.01 ± 2.23	96.48 ± 2.72	94.73 ± 3.51	95.78 ± 2.64	96.49 ± 1.75
Wpbc	73.71 ± 8.15	74.68 ± 7.12	74.16 ± 7.66	73.26 ± 3.70	75.79 ± 5.70
Parkinson	78.33 ± 13.02	75.42 ± 12.14	73.33 ± 10.74	73.75 ± 13.31	77.50 ± 15.28
Movement	78.06 ± 11.55	76.94 ± 13.09	78.06 ± 13.00	76.94 ± 12.17	80.83 ± 7.11
Urban land cover	78.06 ± 5.99	77.93 ± 3.90	77.77 ± 4.83	77.47 ± 4.17	79.91 ± 4.09
Obesity	76.28 ± 8.14	74.19 ± 8.69	75.99 ± 8.39	75.99 ± 8.39	84.84 ± 12.27
Acc	80.86 ± 7.86	78.40 ± 7.78	80.28 ± 7.90	79.79 ± 7.65	83.24 ± 7.22
(Mean rank)	(2.97)	(3.63)	(3.23)	(3.97)	**(1.5)**

In normal data sets, CFRS, DC_ratio and RCU FRS models have improved the Acc compared to SVM classification without feature selection. All four FRSs achieve progress in the Acc compared to KNN classification without feature selection. In noisy data sets, only the RCU FRS model improves the Acc compared with SVM classification without feature selection. The case with KNN also has a similar result.

The reasons for these changes are as follows. The CFRS, DC_ratio, AVDP and RCU FRS models are capable of improving the classification accuracies in normal data sets. The CFRS model is advantageous in addressing mixed-feature types but has no ability to contend with noisy data sets. Thus, its classification performance faces a sharp shift when noisy objects are added. The AVDP model uses the idea of the k-mean of neighbours when computing a lower approximation to decrease the effect of noisy objects. This approach sometimes works, but it is too unrefined when addressing complex noisy situations. The DC_ratio FRS model improves the robustness by detecting the noisy object

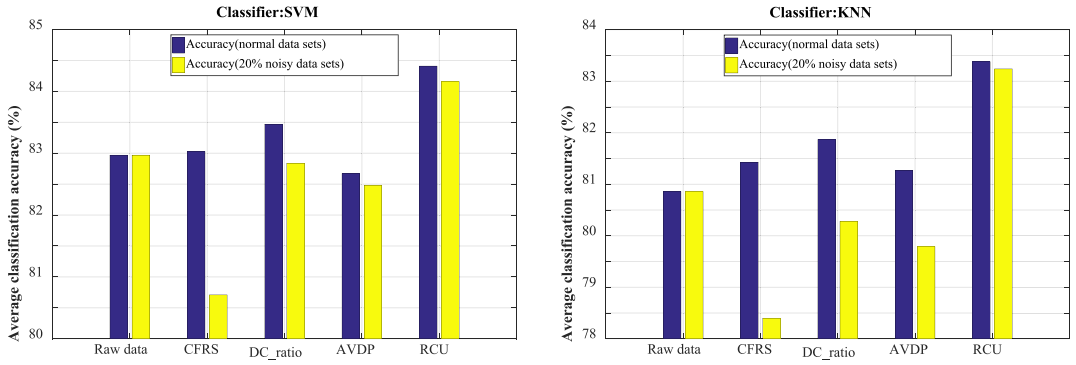

Fig. 5. Robustness comparison of different FRS models with SVM and KNN.

Table 8
Classification accuracy and selected number of features of the RCU FRS model using different kernels with SVM.

	Linear		Hermite		Wavelet		RBF	
	Accuracy	Num	Accuracy	Num	Accuracy	Num	Accuracy	Num
Wine	97.75 ± 2.76	**6**	95.49 ± 3.36	7	97.22 ± 3.73	9	97.22 ± 3.73	8
Heart	77.41 ± 4.52	9	76.30 ± 4.12	**4**	77.78 ± 4.97	7	77.41 ± 5.09	7
Forestfire	92.27 ± 3.26	3	92.67 ± 4.30	**2**	90.22 ± 5.17	7	92.67 ± 4.30	**2**
Hepatitis	84.50 ± 8.84	10	81.17 ± 12.79	**6**	80.54 ± 9.56	9	84.58 ± 4.04	**6**
Ionos	94.02 ± 5.49	16	93.15 ± 5.46	**11**	95.16 ± 3.14	19	94.30 ± 5.27	25
Gamma	71.57 ± 0.45	**2**	72.63 ± 0.51	4	86.03 ± 0.82	8	85.55 ± 0.96	9
Credit	85.22 ± 17.44	5	74.35 ± 6.05	4	85.51 ± 17.01	**2**	85.51 ± 17.49	5
German	74.40 ± 3.85	13	71.00 ± 2.10	**6**	74.8 ± 2.86	12	75.80 ± 3.12	11
Sonar	65.38 ± 18.60	37	72.52 ± 12.32	13	69.79 ± 9.97	39	73.48 ± 17.94	**10**
Wdbc	97.18 ± 1.97	14	94.91 ± 2.53	**6**	97.71 ± 2.09	21	97.18 ± 1.97	8
Wpbc	76.29 ± 1.99	**4**	76.29 ± 1.99	7	80.79 ± 3.99	17	82.87 ± 5.00	7
Parkinson	80.42 ± 10.71	22	80.03 ± 10.24	20	80.42 ± 10.71	40	82.08 ± 9.87	**3**
Movement	26.67 ± 7.68	**2**	57.50 ± 11.66	17	78.89 ± 9.72	77	80.83 ± 10.20	73
Urban land cover	82.66 ± 4.20	51	82.53 ± 3.04	29	82.68 ± 3.73	41	82.97 ± 3.17	**27**
Obesity	31.98 ± 3.56	**2**	24.40 ± 0.93	**2**	72.21 ± 9.16	8	73.67 ± 9.39	7
Acc & Average length	75.85 ± 6.35	13.07	76.33 ± 5.43	**9.20**	83.32 ± 6.44	21.07	84.41 ± 6.77	13.87
(Mean Rank)	(2.93)	(2.6)	(3.47)	**(1.67)**	(2.10)	(3.37)	**(1.50)**	(2.37)

in the nearest neighbour of the target object using the ratio of different classes. The effectiveness may be influenced when the initial noisy objects exist in the ratio computation. In general, our proposed model shows the strongest robustness from the comparison results. The model considers not only noisy detection of the nearest neighbour to the target object but also the class distribution. It has been demonstrated that this idea can enhance the robustness of the feature selection process.

6.4.3. Impact analysis of the RBF-based similarity measure

To evaluate the additive impact of the RBF-based similarity measure on the feature selection, we run experiments where the RBF kernel in the RCU FRS model is replaced with three other baseline kernels: linear, wavelet and Hermite. More setting information of these three methods can be obtained in Appendix A. The performance results of normal data sets and data sets with 20% class noise randomly added with the SVM classifier are shown in Tables 8 and 9, respectively. The corresponding results with KNN in Tables 12 and 13 can also be found in Appendix A. Accuracy in the following tables represents the average accuracy and standard deviation derived from the 10-fold cross-validation process, and Num represents the number of optimal feature subsets.

From Tables 8, 9, 12 and 13, whether with SVM or KNN, the Acc of the RBF-based FRS model outperforms other kernel-based FRS models on most data sets, followed by the wavelet kernel-based FRS model. The linear and Hermite kernel-based FRS models have similar performances: their Accs are at a normal level on most data sets but extremely poor on a few data sets, such as the Movement and Obesity data sets. However, with regard to the average length and mean rank in the number of selected features, the linear and Hermite kernel-based FRS models can always achieve the optimal and suboptimal positions. The RBF-based model obtains the third position, and the wavelet-based model obtains the last position. Based on observations in the experiments, at the same stop condition of the RCU FRS model, we find that the linear and Hermite kernel-based models can derive smaller numbers of feature subsets, but the classification accuracy is also affected by the sharp and excessive reduction. The wavelet kernel-based model derives a relatively large number of feature subsets. The Acc does not clearly improve because redundant features still exist. However, the RBF-based model can balance both factors, producing an acceptable number of feature subsets and obviously boosting the classification accuracy. Therefore, from the comparison results, we conclude that the RBF is superior to other kernels in the similarity measure of the RCU FRS model.

6.5. Statistical analysis

To further explore whether the average classification accuracies of these four models are significantly different, referring to [11,50], the Friedman test [58] and Bonferroni–Dunn test [59] are employed. The Friedman statistic is defined as follows:

$$\chi_F^2 = \frac{12N}{k(k+1)}\left(\sum_{i=1}^{k} R_i^2 - \frac{k(k+1)}{4}\right), F_F = \frac{(N-1)\chi_F^2}{N(k-1)-\chi_F^2}, \tag{26}$$

Table 9
Classification accuracy and selected number of features of the RCU FRS model using different kernels with SVM (20% noisy objects randomly added).

	Linear		Hermite		Wavelet		RBF	
	Accuracy	Num	Accuracy	Num	Accuracy	Num	Accuracy	Num
Wine	93.73 ± 5.39	**3**	91.11 ± 8.31	4	97.22 ± 3.73	6	98.33 ± 2.55	10
Heart	77.41 ± 5.35	**4**	76.30 ± 4.12	**4**	77.04 ± 5.19	7	77.41 ± 4.52	7
Forestfire	92.27 ± 3.26	3	91.85 ± 3.04	**2**	90.62 ± 4.75	6	92.67 ± 4.30	**2**
Hepatitis	80.54 ± 5.49	6	80.58 ± 9.96	**2**	79.38 ± 2.25	3	84.58 ± 4.04	6
Ionos	93.73 ± 5.39	13	93.15 ± 4.99	**4**	94.30 ± 7.38	24	94.87 ± 3.07	21
Gamma	65.85 ± 0.42	**2**	72.63 ± 0.51	4	86.03 ± 0.82	8	85.67 ± 0.89	6
Credit	84.49 ± 16.85	**2**	72.61 ± 5.16	3	85.51 ± 17.49	**2**	85.55 ± 17.67	5
German	69.80 ± 2.18	**2**	71.20 ± 1.89	6	74.40 ± 2.62	17	75.80 ± 3.12	15
Sonar	67.83 ± 16.43	12	70.60 ± 12.72	11	70.26 ± 12.84	30	72.62 ± 10.59	**5**
Wdbc	96.48 ± 2.24	10	94.03 ± 2.24	**4**	97.89 ± 2.05	26	97.54 ± 1.96	12
Wpbc	76.09 ± 1.99	**2**	76.29 ± 1.99	**2**	79.84 ± 5.29	25	81.37 ± 4.34	9
Parkinson	79.58 ± 10.11	19	76.67 ± 0.99	**5**	79.58 ± 10.28	40	80.00 ± 11.46	12
Movement	15.83 ± 3.52	**4**	47.22 ± 14.38	14	78.61 ± 9.94	27	80.56 ± 4.03	80
Urban land cover	81.93 ± 5.21	48	80.45 ± 4.36	**28**	82.38 ± 3.62	72	83.26 ± 3.64	107
Obesity	31.98 ± 3.56	**2**	24.4 ± 0.93	**2**	71.26 ± 8.72	8	72.21 ± 9.16	8
Acc &Average length (Mean Rank)	73.84 ± 5.83 (3.13)	8.80 (1.90)	74.61 ± 5.04 (3.40)	**6.33 (1.6)**	82.95 ± 6.46 (2.30)	20.07 (3.43)	84.16 ± 5.69 **(1.17)**	20.33 (3.07)

Table 10
Value of F_F for different classification algorithms.

	SVM	KNN
F_F	7.85	6.26

Table 11
Value of F_F for different classification algorithms (comparison for different kernels).

	SVM	KNN
F_F	31.58	17.77

where k is the number of models, N is the number of data sets and R_i is the average rank of model i among all data sets. To validate the RCU FRS model, we compare the data sets with and without noisy information. We set $k = 5$ and $N = 30$; therefore, the degrees of freedom are $(k-1)(N-1) = 117$. From the standard table, we have $F(4, 117) = 2.00$ for $\alpha = 0.10$. Table 10 shows the value of F_F for different classification algorithms through the calculation.

As 7.85 and 6.26 are larger than 2.18, the null hypothesis is rejected. Thus, we conclude that the five FRS models are significantly different. In addition, to validate that the RBF-based RCU model is significantly different from other kernel-based models, we have $k = 4$, $N = 30$ and $F(3, 87) = 2.15$ for $\alpha = 0.10$. Table 10 shows the Friedman statistic.

Since 31.58 and 17.77 are larger than 2.15, we can also derive that these kernel-based FRS models are significantly different (see Table 11).

To determine whether the RCU FRS model achieves competitive performance against the other FRS models, the Bonferroni–Dunn test is employed. The proposed model is considered the control model. The difference between the two models is compared with the following critical difference (CD):

$$CD = q_\alpha \sqrt{\frac{k(k+1)}{6N}}. \tag{27}$$

We have $q_\alpha = 2.241$ at the significance level $\alpha = 0.1$, and thus CD=0.915 ($k = 5$, $N = 30$) to compare the five FRS models. We also have $q_\alpha = 2.128$ at the significance level $\alpha = 0.1$, and thus CD=0.709 ($k = 4$, $N = 30$) to compare the four kernel-based FRS models. The performances of the RCU FRS model and the comparison models are considered to be significantly different if their average ranks on all data sets differ by at least one CD.

Figs. 6 and 7 vividly illustrate the CD diagrams with SVM and KNN. The axis locates the average rank of each comparison model (lower ranks to the right). In each figure, any comparison model that is interconnected with the RCU model in a thick line is deemed to be not obviously different from the RCU model or the RBF-based model. Otherwise, no connecting line indicates a significant difference.

From the above figures, we can conclude the following: (1) With SVM, the RCU FRS model obtains statistically better performance than the raw data (no feature selection process), CFRS and AVDP FRS models. With regard to different kernel-based models, the RBF-based RCU FRS model performs significantly better than the linear and Hermite-based RCU FRS models. However, there is no consistent evidence to demonstrate significant differences between the RCU and DC_ratio FRS models or the RBF and wavelet kernels. (2) With KNN, the RCU FRS model is significantly better than the raw data, CFRS and AVDP FRS models. The RBF-based RCU FRS model achieves significantly better performance than the linear, Hermite and wavelet kernel-based RCU FRS models. No consistent evidence indicates a significant difference between the RCU and DC_ratio FRS models.

7. Conclusions

FRS theory contributes significantly to feature selection. In FRS models, numerous existing studies apply multiple intersections to compute the fuzzy rough approximations. This approach may cause the feature subsets to be less discriminable in certain situations. Simultaneously, noisy information usually corrupts the FRS model in practice. Enhancing the robustness of the FRS model is a desirable task. To solve these problems, we propose the RBF kernel-based similarity measure to compute the upper and lower approximations, where VDM and Euclidean metrics are employed to handle mixed symbolic and real-valued features. Moreover, we propose a novel robust FRS model, named the RCU FRS model, which can detect and remove noisy information using the relative classification measure by KNN and Bayes rules. In this model, the influence of the nearest object to the target object is considered, and the insight class distribution of the neighbours of the nearest object is also of concern. Finally, a series of experiments is conducted to assess the proposed model. Comparison analysis shows that the proposed model is superior to other FRS models, and the RBF kernel is an applicable channel for the computation of fuzzy relations in the FRS. Statistical analysis by the Friedman test and Bonferroni–Dunn test demonstrates that the RCU FRS model is significantly different from or at least as good as existing works

In the proposed model, we are concerned with the influence of class noise. The feature noise is not considered. In addition, it will

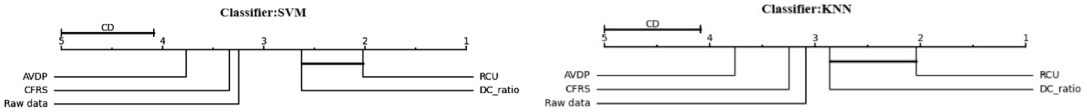

Fig. 6. Comparison of the RCU FRS model against other FRS models with SVM and KNN.

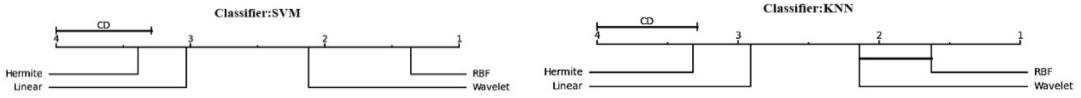

Fig. 7. Comparison of the RBF-based model against other kernel-based FRS models with SVM and KNN.

be possible to advance the feature selection process when taking the correlation of features into consideration. Thus, in future works, we plan to generalize the robust FRS model in the feature noise environment and measure the correlation relationships between features when conducting feature selection tasks.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

The authors would like to express the sincere appreciation to the editor and anonymous reviewers for their insightful comments, which greatly improve the quality of this paper. This work is supported partially by the National Natural Science Foundation of China under grant number 71871229 and the Hunan Provincial Natural Science Foundation of China under grant number 2021JJ30031.

Appendix A

The computation formulas of the linear, Hermite and wavelet kernels are as follows:

(1) Linear

$$k(x,y)=\sum_{j=1}^{n}x_jy_j \tag{28}$$

(2) Hermite ($n=2$)

$$k(x,y)=1+\sum_{j=1}^{n}x_jy_j+\sum_{j=1}^{n}\left(x_j^2-1\right)\left(y_j^2-1\right)$$
$$=\sum_{j=1}^{n}\left(1+x_jy_j+\left(x_j^2-1\right)\left(y_j^2-1\right)\right) \tag{29}$$

(3) Wavelet

$$k(x,y)=\prod_{j=1}^{n}\left(\cos\left(1.75\frac{x_j-y_j}{\delta}\right)\exp\frac{\left\|x_j-y_j\right\|^2}{2\delta^2}\right) \tag{30}$$

All data sets are normalized to the interval [0,1] by the min–max standardization before experiments. In addition, to limit the value of the similarity measure to the range of [0,1] in the experiment, all similarity values are computed and stored in the list and then normalized to the interval [0,1] using the min–max standardization approach. To control the variables, the procedure of addressing mixed features in the wavelet kernel is also the same as the proposed method: for x_i-y_i, the real-valued features use the Euclidean distance metric, and the symbolic features use the VDM metric. We set $\delta=1$ in the wavelet kernel. Since the feature values in the linear and Hermite kernels are computed in the form of a product, the abovementioned metrics are not applicable in Eqs. (28) and (29).

Appendix B. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.knosys.2022.109795.

Table 12
Classification accuracy and the number of selected features of the RCU FRS model using different kernels with KNN.

	Linear		Hermite		Wavelet		RBF	
	Accuracy	Num	Accuracy	Num	Accuracy	Num	Accuracy	Num
Wine	95.46 ± 3.48	**6**	93.20 ± 6.15	7	97.22 ± 5.12	**6**	96.05 ± 4.45	12
Heart	79.63 ± 7.08	**4**	69.63 ± 7.73	**4**	79.63 ± 10.51	7	80.00 ± 9.10	9
Forestfire	91.45 ± 3.71	4	91.90 ± 5.59	**2**	88.58 ± 7.20	6	91.90 ± 5.59	**2**
Hepatitis	81.21 ± 8.77	8	81.83 ± 7.11	**6**	83.75 ± 6.30	9	81.21 ± 7.01	11
Ionos	91.17 ± 4.31	6	92.29 ± 4.26	**4**	94.33 ± 7.38	60	91.75 ± 4.81	5
Gamma	65.39 ± 0.78	**2**	73.61 ± 0.90	4	83.33 ± 0.92	8	83.11 ± 1.18	9
Credit	85.51 ± 17.49	**2**	72.32 ± 4.02	4	85.51 ± 17.49	**2**	82.32 ± 15.78	5
German	71.30 ± 3.77	13	67.90 ± 5.24	**6**	72.60 ± 3.69	16	72.60 ± 2.69	14
Sonar	70.60 ± 13.45	34	69.21 ± 8.41	13	64.33 ± 12.91	54	72.55 ± 15.10	**10**
Wdbc	96.13 ± 2.04	10	94.73 ± 2.35	**6**	96.84 ± 2.05	26	96.31 ± 3.09	**6**
Wpbc	73.84 ± 7.82	20	72.71 ± 7.53	**7**	73.66 ± 8.50	25	77.32 ± 3.11	**7**
Parkinson	77.08 ± 10.91	22	78.33 ± 11.46	18	79.17 ± 10.87	15	80.42 ± 10.04	**6**
Movement	33.89 ± 7.43	**4**	64.40 ± 10.74	17	76.94 ± 13.32	27	79.44 ± 12.86	59
Urban land cover	80.13 ± 4.91	48	74.36 ± 6.08	**28**	76.58 ± 5.48	72	76.86 ± 6.51	87
Obesity	42.73 ± 4.25	**2**	13.97 ± 2.80	**2**	88.84 ± 12.27	8	88.93 ± 11.87	7
Acc & Average length	75.70 ± 6.68	12.33	74.03 ± 6.02	**8.53**	83.38 ± 1.67	22.73	83.38 ± 7.55	16.60
(Mean Rank)	(2.90)	(2.2)	(3.30)	(**1.77**)	(2.07)	(3.2)	(**1.73**)	(2.83)

Table 13
Classification accuracy and the number of selected features of the RCU FRS model using different kernels with KNN (20% noisy objects randomly added).

	Linear		Hermite		Wavelet		RBF	
	Accuracy	Num	Accuracy	Num	Accuracy	Num	Accuracy	Num
Wine	92.71 ± 5.63	**3**	91.08 ± 7.92	4	96.63 ± 4.45	9	96.60 ± 3.83	10
Heart	78.52 ± 8.41	9	69.63 ± 7.73	**4**	78.52 ± 7.73	7	80.00 ± 9.21	9
Forestfire	91.03 ± 4.31	3	89.83 ± 5.10	**2**	88.17 ± 7.14	7	91.90 ± 5.59	**2**
Hepatitis	81.21 ± 8.77	8	79.92 ± 9.29	**2**	81.75 ± 7.93	10	83.88 ± 9.38	6
Ionos	86.62 ± 7.10	16	88.60 ± 4.18	**5**	86.61 ± 5.27	19	91.90 ± 5.59	**5**
Gamma	65.39 ± 6.78	**2**	73.61 ± 0.90	4	83.30 ± 0.92	8	82.97 ± 1.11	6
Credit	82.03 ± 14.63	5	69.71 ± 5.92	3	85.51 ± 17.49	**2**	81.77 ± 13.74	5
German	70.70 ± 3.41	7	65.70 ± 4.90	**6**	71.90 ± 3.30	17	72.70 ± 2.69	15
Sonar	67.83 ± 11.80	16	68.67 ± 9.03	**11**	64.29 ± 14.11	30	71.60 ± 10.89	13
Wdbc	95.77 ± 2.65	14	93.15 ± 2.53	**4**	96.83 ± 2.21	21	96.49 ± 1.75	12
Wpbc	71.71 ± 6.50	17	72.18 ± 9.12	**2**	72.18 ± 5.36	21	75.79 ± 5.70	9
Parkinson	77.50 ± 11.06	21	70.83 ± 8.74	**5**	78.75 ± 14.61	31	77.50 ± 15.28	12
Movement	33.89 ± 7.43	**4**	56.11 ± 10.82	14	76.94 ± 13.32	27	80.83 ± 7.11	11
Urban land cover	79.24 ± 4.91	67	74.36 ± 6.08	**28**	75.39 ± 6.24	41	79.91 ± 4.09	107
Obesity	42.73 ± 4.25	**2**	13.97 ± 2.80	**2**	76.28 ± 8.14	8	84.84 ± 12.27	8
Acc & Average length	74.46 ± 7.18	12.93	71.82 ± 6.34	**6.4**	80.87 ± 2.13	17.2	83.24 ± 7.22	15.33
(Mean Rank)	(2.93)	(2.5)	(3.43)	**(1.43)**	(2.20)	(3.43)	**(1.43)**	(2.63)

References

[1] D. Lei, P. Liang, J. Hu, Y. Yuan, New online streaming feature selection based on neighborhood rough set for medical data, Symmetry 12 (10) (2020) 1635.

[2] M.Q. Li, L. Zhang, Multinomial mixture model with feature selection for text clustering, Knowl.-Based Syst. 21 (7) (2008) 704–708.

[3] M. Toğaçar, Z. Cömert, B. Ergen, Classification of brain MRI using hyper column technique with convolutional neural network and feature selection method, Expert Syst. Appl. 149 (2020) 113274.

[4] T.S. Andrews, M. Hemberg, M3Drop: Dropout-based feature selection for scRNASeq, Bioinformatics 35 (16) (2018) 2865–2867.

[5] H. Liu, H. Motoda, Feature Selection for Knowledge Discovery and Data Mining, Springer Science & Business Media, 2012.

[6] R.W. Swiniarski, A. Skowron, Rough set methods in feature selection and recognition, Pattern Recognit. Lett. 24 (6) (2003) 833–849.

[7] Q. Hu, D. Yu, J. Liu, C. Wu, Neighborhood rough set based heterogeneous feature subset selection, Inform. Sci. 178 (18) (2008) 3577–3594.

[8] K. Zhang, J. Zhan, W.Z. Wu, On multi-criteria decision-making method based on a fuzzy rough set model with fuzzy alpha-neighborhoods, IEEE Trans. Fuzzy Syst. 29 (9) (2021) 2491–2505.

[9] D. Dubois, H. Prade, Rough fuzzy sets and fuzzy rough sets*, Int. J. Gen. Syst. 17 (2–3) (1990) 191–209.

[10] P.A. Estevez, M. Tesmer, C.A. Perez, J.M. Zurada, Normalized mutual information feature selection, IEEE Trans. Neural Netw. 20 (2) (2009) 189–201.

[11] Z. Yuan, H. Chen, P. Zhang, J. Wan, T. Li, A novel unsupervised approach to heterogeneous feature selection based on fuzzy mutual information, IEEE Trans. Fuzzy Syst. (2021) 1.

[12] Y. Li, B.L. Lu, Feature selection based on loss-margin of nearest neighbor classification, Pattern Recognit. 42 (9) (2009) 1914–1921.

[13] M.R. Gauthama Raman, N. Somu, K. Kirthivasan, R. Liscano, V.S. Shankar Sriram, An efficient intrusion detection system based on hypergraph - genetic algorithm for parameter optimization and feature selection in support vector machine, Knowl.-Based Syst. 134 (2017) 1–12.

[14] M. Paniri, M.B. Dowlatshahi, H. Nezamabadi-pour, MLACO: A multi-label feature selection algorithm based on ant colony optimization, Knowl.-Based Syst. 192 (2020) 105285.

[15] J. Hamidzadeh, E. Rezaeenik, M. Moradi, Predicting users' preferences by fuzzy rough set quarter-sphere support vector machine, Appl. Soft Comput. 112 (2021) 107740.

[16] S. Moslemnejad, J. Hamidzadeh, Weighted support vector machine using fuzzy rough set theory, Soft Comput. 25 (13) (2021) 8461–8481.

[17] Z.J. Viharos, K.B. Kis, Á. Fodor, M.I. Büki, Adaptive, hybrid feature selection (AHFS), Pattern Recognit. 116 (2021) 107932.

[18] G. Kou, Y. Xu, Y. Peng, F. Shen, Y. Chen, K. Chang, S. Kou, Bankruptcy prediction for SMEs using transactional data and two-stage multiobjective feature selection, Decis. Support Syst. 140 (2021) 113429.

[19] Y. Xue, H. Zhu, J. Liang, A. Słowik, Adaptive crossover operator based multi-objective binary genetic algorithm for feature selection in classification, Knowl.-Based Syst. 227 (2021) 107218.

[20] Y. Hu, Y. Zhang, D. Gong, X. Sun, Multi-participant federated feature selection algorithm with particle swarm optimizaiton for imbalanced data under privacy protection, IEEE Trans. Artif. Intell. (2022) 1.

[21] M. kelidari, J. Hamidzadeh, Feature selection by using chaotic cuckoo optimization algorithm with levy flight, opposition-based learning and disruption operator, Soft Comput. 25 (4) (2021) 2911–2933.

[22] P. Zhang, T. Li, G. Wang, C. Luo, H. Chen, J. Zhang, D. Wang, Z. Yu, Multi-source information fusion based on rough set theory: A review, Inf. Fusion 68 (2021) 85–117.

[23] B. Sun, W. Ma, Y. Qian, Multigranulation fuzzy rough set over two universes and its application to decision making, Knowl.-Based Syst. 123 (2017) 61–74.

[24] N.N. Morsi, M.M. Yakout, Axiomatics for fuzzy rough sets, Fuzzy Sets and Systems 100 (1) (1998) 327–342.

[25] B. Moser, On the T-transitivity of kernels, Fuzzy Sets and Systems 157 (13) (2006) 1787–1796.

[26] Z. Yuan, H. Chen, X. Yang, T. Li, K. Liu, Fuzzy complementary entropy using hybrid-kernel function and its unsupervised attribute reduction, Knowl.-Based Syst. 231 (2021) 107398.

[27] D. Chen, Y. Yang, Attribute reduction for heterogeneous data based on the combination of classical and fuzzy rough set models, IEEE Trans. Fuzzy Syst. 22 (5) (2014) 1325–1334.

[28] D. Chen, L. Zhang, S. Zhao, Q. Hu, P. Zhu, A novel algorithm for finding reducts with fuzzy rough sets, IEEE Trans. Fuzzy Syst. 20 (2) (2012) 385–389.

[29] R. Jensen, Q. Shen, New approaches to fuzzy-rough feature selection, IEEE Trans. Fuzzy Syst. 17 (4) (2009) 824–838.

[30] C. Wang, M. Shao, Q. He, Y. Qian, Y. Qi, Feature subset selection based on fuzzy neighborhood rough sets, Knowl.-Based Syst. 111 (2016) 173–179.

[31] Y. Yang, D. Chen, H. Wang, E.C.C. Tsang, D. Zhang, Fuzzy rough set based incremental attribute reduction from dynamic data with sample arriving, Fuzzy Sets and Systems 312 (2017) 66–86.

[32] P. Maji, S. Paul, Rough-fuzzy clustering for grouping functionally similar genes from microarray data, IEEE/ACM Trans. Comput. Biol. Bioinform. (2012) 1–14.

[33] Y. Qian, J. Liang, W.Z.Z. Wu, C. Dang, Information granularity in fuzzy binary GrC model, IEEE Trans. Fuzzy Syst. 19 (2) (2011) 253–264.

[34] C. Wang, Y. Huang, M. Shao, X. Fan, Fuzzy rough set-based attribute reduction using distance measures, Knowl.-Based Syst. 164 (2019) 205–212.

[35] Y. Zhang, S. Wang, K. Xia, Y. Jiang, P. Qian, Alzheimer's disease multiclass diagnosis via multimodal neuroimaging embedding feature selection and fusion, Inf. Fusion 66 (2021) 170–183.

[36] M. Wang, C. Wu, L. Wang, D. Xiang, X. Huang, A feature selection approach for hyperspectral image based on modified ant lion optimizer, Knowl.-Based Syst. 168 (2019) 39–48.

[37] V. Hooshm, J. Hamidzadeh, New Hermite orthogonal polynomial kernel and combined kernels in support vector machine classifier, Pattern Recognit. 60 (2016) 921–935.

[38] A. Zeng, T. Li, D. Liu, J. Zhang, H. Chen, A fuzzy rough set approach for incremental feature selection on hybrid information systems, Fuzzy Sets and Systems 258 (2015) 39–60.

[39] U. Ravale, N. Marathe, P. Padiya, Feature selection based hybrid anomaly intrusion detection system using K means and RBF kernel function, Procedia Comput. Sci. 45 (2015) 428–435.

[40] C. Stanfill, D. Waltz, Toward memory-based reasoning, Commun. ACM 29 (12) (1986) 1213–1228.

[41] P.E. Danielsson, Euclidean distance mapping, Comput. Graph. Image Process. 14 (3) (1980) 227–248.
[42] X. Jia, Y. Rao, L. Shang, T. Li, Similarity-based attribute reduction in rough set theory: A clustering perspective, Int. J. Mach. Learn. Cybern. 11 (5) (2020) 1047–1060.
[43] S. Luo, D. Miao, Z. Zhang, Y. Zhang, S. Hu, A neighborhood rough set model with nominal metric embedding, Inform. Sci. 520 (2020) 373–388.
[44] A. Hamed, M. Tahoun, H. Nassar, KNNHI: Resilient KNN algorithm for heterogeneous incomplete data classification and K identification using rough set theory, J. Inf. Sci. (2022) 01655515211069539.
[45] X. Zhu, X. Wu, Class noise vs. Attribute noise: A quantitative study, Artif. Intell. Rev. 22 (3) (2004) 177–210.
[46] J.M. Fernández Salido, S. Murakami, Rough set analysis of a general type of fuzzy data using transitive aggregations of fuzzy similarity relations, Fuzzy Sets and Systems 139 (3) (2003) 635–660.
[47] Q. Hu, S. An, D. Yu, Soft fuzzy rough sets for robust feature evaluation and selection, Inform. Sci. 180 (22) (2010) 4384–4400.
[48] Q. Hu, L. Zhang, S. An, D. Zhang, D. Yu, On robust fuzzy rough set models, IEEE Trans. Fuzzy Syst. 20 (4) (2012) 636–651.
[49] S. Zhao, H. Chen, C. Li, X. Du, H. Sun, A novel approach to building a robust fuzzy rough classifier, IEEE Trans. Fuzzy Syst. 23 (4) (2015) 769–786.
[50] Y. Li, S. Wu, Y. Lin, J. Liu, Different classes' ratio fuzzy rough set based robust feature selection, Knowl.-Based Syst. 120 (2017) 74–86.
[51] N. Verbiest, C. Cornelis, F. Herrera, OWA-FRPS: A prototype selection method based on ordered weighted average fuzzy rough set theory, in: D. Ciucci, M. Inuiguchi, Y. Yao, D. Ślęzak, G. Wang (Eds.), Rough Sets, Fuzzy Sets, Data Mining, and Granular Computing, Springer Berlin Heidelberg, Berlin, Heidelberg, 2013, pp. 180–190.
[52] J. Zhan, H. Jiang, Y. Yao, Three-way multi-attribute decision-making based on outranking relations, IEEE Trans. Fuzzy Syst. 29 (10) (2021) 2844–2858.
[53] J. Wang, X. Ma, Z. Xu, J. Zhan, Three-way multi-attribute decision making under hesitant fuzzy environments, Inform. Sci. 552 (2021) 328–351.
[54] W. Wang, J. Zhan, J. Mi, A three-way decision approach with probabilistic dominance relations under intuitionistic fuzzy information, Inform. Sci. 582 (2022) 114–145.
[55] J. Ye, J. Zhan, W. Ding, H. Fujita, A novel three-way decision approach in decision information systems, Inform. Sci. 584 (2022) 1–30.
[56] O. Reyes, C. Morell, S. Ventura, Effective active learning strategy for multi-label learning, Neurocomputing 273 (2018) 494–508.
[57] Y. Zhang, Z. Zhou, Cost-sensitive face recognition, IEEE Trans. Pattern Anal. Mach. Intell. 32 (10) (2010) 1758–1769.
[58] M. Friedman, A comparison of alternative tests of significance for the problem of m rankings, Ann. Math. Stat. 11 (1) (1940) 86–92.
[59] O.J. Dunn, Multiple comparisons among means, J. Amer. Statist. Assoc. 56 (293) (1961) 52–64.