On Robust Fuzzy Rough Set Models

Qinghua Hu, *Member, IEEE*, Lei Zhang, *Member, IEEE*, Shuang An, David Zhang, *Fellow, IEEE*, and Daren Yu

***Abstract*—Rough sets, especially fuzzy rough sets, are supposedly a powerful mathematical tool to deal with uncertainty in data analysis. This theory has been applied to feature selection, dimensionality reduction, and rule learning. However, it is pointed out that the classical model of fuzzy rough sets is sensitive to noisy information, which is considered as a main source of uncertainty in applications. This disadvantage limits the applicability of fuzzy rough sets. In this paper, we reveal why the classical fuzzy rough set model is sensitive to noise and how noisy samples impose influence on fuzzy rough computation. Based on this discussion, we study the properties of some current fuzzy rough models in dealing with noisy data and introduce several new robust models. The properties of the proposed models are also discussed. Finally, a robust classification algorithm is designed based on fuzzy lower approximations. Some numerical experiments are given to illustrate the effectiveness of the models. The classifiers that are developed with the proposed models achieve good generalization performance.**

***Index Terms*—Fuzzy rough sets, model, robustness, rough sets.**

I. Introduction

ROUGH set theory [1], especially fuzzy rough set theory [2], which encapsulates two kinds of uncertainty of fuzziness and roughness into a single model, has attracted much attention from the domains of granular computing, machine learning, and uncertainty reasoning over the past decade [3]–[10]. Fuzzy information granulation and approximate reasoning are two elemental modules of human cognition and reasoning [11], [59]. We form fuzzy concepts of the universe according to their attributes and utilize these concepts to approximately describe other objects. The fuzzy rough set theory imitates the idea hidden in human reasoning. This theory granulates the universe of discourse into a set of fuzzy concepts based on fuzzy relations and then approximates arbitrary fuzzy sets with these fuzzy concepts. Thus, this theory is considered to be an important mathematical tool for granular computing [12], [48], [49], [55], [58].

In the framework of Pawlak's rough sets, the universe is divided into a set of equivalence classes (which are also called elemental concepts) according to the attribute values of the objects. Usually, the knowledge about the universe is limited, such that we can just divide the universe into a set of granules of limited granularity. In this case, if we utilize these elemental concepts to describe new sets, we are usually not able to obtain perfect description as there is a region of approximation boundary. Rough sets are introduced to characterize the boundary in approximation. In fuzzy rough sets, operators of fuzzy lower and upper approximations were defined. The difference between upper approximation and lower approximation is called the boundary of the approximated subset. If the elemental concepts that are used in approximation are of large granularity, the boundary region would be large as well. However, if the elemental concepts become finer by introducing new knowledge, the boundary region would correspondingly become smaller. The size of boundary region reflects the approximation capability of the elemental concepts. Assume that the elemental concepts are generated with some attributes and the subsets to be approximated are the classification of the objects. Then, the size of the boundary reflects the capability of the attributes to describe the classification.

Fuzzy rough set theory has been successfully used in gene clustering [13], feature selection [3], [5], [9], attribute reduction [7], [14], [15], case generation, and rule extraction [16]–[18], [56]. The dependence function, which is defined as the ratio of the sizes of the lower approximation of classification over the universe, plays a key role in these applications. This function underlies a number of learning algorithms, including feature selection, attribute reduction, rule extraction, and decision trees [19], [52], [56]. In these algorithms, a function is required to evaluate attribute quality. The learned model is expected to work well on unseen samples that are generated with the same but unknown probability distribution as the training set.

Since the real distribution is not available, dependence between features and decision is estimated with a finite set of training samples. We assume that the test samples satisfy the same distribution as the training set. However, usually, this assumption is not exactly true due to uncertainty of randomness. In this case, the robustness of learning algorithms is very important; otherwise, small deviation in training samples may lead to completely different models. Robustness becomes more and more important as data quality cannot be guaranteed in real-world applications. Noisy information may be introduced in data

Manuscript received June 13, 2011; revised September 17, 2011; accepted December 5, 2011. Date of publication December 22, 2012; date of current version August 1, 2012. This work was supported by the National Natural Science Foundation of China under Grant 60703013 and Grant 10978011, the Key Program of National Natural Science Foundation of China under Grant 60932008, the National Science Fund for Distinguished Young Scholars under Grant 50925625, the China Postdoctoral Science Foundation, and Key Laboratory of Condition Monitoring and Control for Power Plant Equipment of North China Electric Power University under Grant 2011CM001. The work of Q. H. Hu was supported by The Hong Kong Polytechnic University under G-YX3B.

Q. H. Hu was with the Harbin Institute of Technology, Harbin 150001, China. He is now with Tianjin University, Tianjin 300072, China (e-mail: huqinghua@hit.edu.cn).

L. Zhang and D. Zhang are with the Hong Kong Polytechnic University, Kowloon, Hong Kong (e-mail: cslzhang@comp.polyu.edu.hk; csdzhang@comp.polyu.edu.hk).

S. An was with the Harbin Institute of Technology, Harbin 150001, China. She is now with Northeastern University at Qinhuangdao, Qinhuangdao 066004, China (e-mail: anshuang_001@163.com).

D. R. Yu is with the Harbin Institute of Technology, Harbin 150001, China (e-mail: Yudaren@hit.edu.cn).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TFUZZ.2011.2181180

acquisition, storage, and transmission [20], [21]. Especially in the context of data mining, data are not well prepared for a single task, and they may be gathered from multiple sources and for multiple tasks [22]. It is inevitable to deal with corrupted information in this case [23]–[26], [51], [57], [60].

Claimed as a powerful mathematical tool to deal with uncertainty, it is shown that the rough set models are heavily sensitive to noisy samples. Experimental analysis reveals that fuzzy rough sets are sensitive to mislabeled samples [40]. One mislabeled sample may result in significantly different fuzzy approximations of classification. The classical model of fuzzy rough sets should be improved to deal with noisy tasks.

In fact, several works have been conducted to combat with noise in the domain of rough sets. In 1990, Yao *et al.* gave the model of decision-theoretic rough sets [27]. In 1993, Ziarko introduced the model of variable precision rough sets (VPRS) [28], and in 2005, Slezak and Ziarko proposed the Bayesian rough set model. Among these models, VPRS was widely discussed and used in dealing with noisy tasks [29]–[31]. Meanwhile, the classical fuzzy rough set model was also improved to analyze noisy data. In 2003, Salido and Murakami presented a β-precision aggregation fuzzy rough set model based on β-precision aggregation triangular operators [32]. In 2004, the model of variable precision fuzzy rough sets (VPFRS) was introduced in [33], where the fuzzy memberships of a sample to the lower and upper approximations are computed with fuzzy inclusion. In 2007, Hu *et al.* introduced another fuzzy rough set model based on fuzzy granulation and fuzzy inclusion [14]. In addition, in 2007, Cornelis *et al.* presented a model called vaguely quantified rough sets (VQRS) [34], which was used in constructing a robust feature selection algorithm in 2008 [35]. In 2009, Zhao *et al.* constructed a new model, which is called fuzzy variable precision rough sets (FVPRS), to handle noise of misclassification and perturbation [47]. In 2010, Cornelis *et al.* constructed a model of fuzzy rough sets based on ordered weighted average operators [54].

To the best of our knowledge, no extensive work has been devoted to discussing the influence of noise on rough approximation and on the statistics defined in rough sets so far. To this end, we try to give answers to the following questions in this paper: Why are the current models of rough sets sensitive to noise? What effect does the noisy information have on rough computation? How are robust models of rough sets developed for handling noisy tasks? The answers will help us to understand and utilize fuzzy rough models and construct effective algorithms based on the models.

Roughly speaking, there are two types of noise: attribute noise and class noise [36]. Attribute noise is usually introduced by sensors in data acquisition [37], while class noise is generated by sample mislabeling. Attribute noise leads to the variations of samples' locations in feature spaces, while class noise changes the class labels of samples. These two kinds of noise have different impact on learning algorithms and have different effect on dependence estimation when the rough set theory is considered. Unfortunately, no work has been devoted to studying and comparing the performances of these models in dealing with noise so far. In this paper, we focus on these problems and give systematic analysis on robust fuzzy rough set models. We expect to reveal why the classical model is not robust to noisy samples and analyze the advantages and disadvantages of the current models and indicate how to improve them.

The remainder of this paper is organized as follows. First, we give preliminary knowledge of rough sets and fuzzy rough sets in Section II; then, we discuss the existing models of robust fuzzy rough sets in Section III. A collection of robust models of fuzzy rough sets are introduced in Section IV. Experimental analysis is given in Section V. Finally, conclusions are drawn in Section VI.

II. Preliminaries and Problem Description

We review the definitions of rough sets and fuzzy rough sets in this section and discuss their disadvantages.

A. Basic Definitions

Given a finite set of objects $U = \{x_1, x_2, \ldots, x_n\}$ described with a set of attributes $A = \{a_1, a_2, \ldots, a_m\}$, each object $x_i \in U$ can be formulated as a vector $x_i = \langle x_{i1}, x_{i2}, \ldots, x_{im}\rangle$, where x_{ij} is the *j*th feature value of sample x_i. We call $IS = \langle U, A\rangle$ an information system. As to Pawlak's rough set model, the value domains of features are discrete. Under this condition, i.e., $\forall B \subseteq A$, an equivalence relation can be generated over the universe: $IND(B) = \{(x, y) \in U \times U | a(x) = a(y), \forall a \in B\}$. With $IND(B)$, U is partitioned into a family of equivalence classes. The equivalence class induced by attribute *B* and sample x_i is denoted by $[x_i]_B$, which is a subset of samples having the same feature values as x_i. Given a classification task, the class labels of the objects are known in advance. Let X be a subset of objects belonging to the same class. The lower and upper approximations of X with respect to B are defined as $\underline{B}X = \{x_i \in U | [x_i]_B \subseteq X\}$ and $\overline{B}X = \{x_i \in U | [x_i]_B \cap X \neq \emptyset\}$, respectively. The lower approximation of X consists of the samples whose equivalence classes consistently belong to X, while its upper approximation is the subset of samples whose equivalence classes have objects in X. If $\underline{B}X \neq \overline{B}X$, the approximation boundary of X is computed as $BND_B(X) = \overline{B}X - \underline{B}X$, which is the subset of objects whose equivalence classes are inconsistent. They have the same feature values but belong to different classes. As per classification modeling, we want to know the consistent patterns hidden in datasets.

The aforementioned model is constructed under the assumption that only discrete features exist in the information system. In practice, most of classification tasks are described with numerical features or fuzzy information. In this case, neighborhood relations or fuzzy similarity relations are used and neighborhood or fuzzy granules are generated. Then, we use these granules to approximate decision classes. Let U be a finite set of objects and R be a fuzzy similarity relation on U generated with features B (note that we here assume that $R(x, y)$ monotonously decreases with the distance between x and y). We have $R(x, x) = 1$, $R(x, y) = R(y, x)$, and $R(x, z) \geq T(R(x, y), R(y, z))$. $[x_i]_B = r_{i1}/x_1 + r_{i2}/x_2 + \cdots + r_{\mathrm{in}}/x_n$ is the fuzzy granule induced by x_i and B. For any fuzzy subset

X in U, two pairs of fuzzy approximation operators are defined as

$$\begin{aligned}
\underline{R_S}X(x) &= \inf_{y\in U} S(N(R(x,y)), X(y)) \\
\overline{R_T}X(x) &= \sup_{y\in U} T(R(x,y), X(y)) \\
\underline{R_\vartheta}X(x) &= \inf_{y\in U} \vartheta(R(x,y), X(y)) \\
\overline{R_\sigma}X(x) &= \sup_{y\in U} \sigma(N(R(x,y)), X(y)) \qquad (1)
\end{aligned}$$

where T, S, ϑ, and σ stand for fuzzy triangular norm (T-norm), fuzzy triangular conorm (T-conorm), T-residuated implication and its dual, respectively, and N is a negator. The standard negator is defined as $N(x) = 1 - x$. Some typical fuzzy operators are listed as follows:

$$\begin{aligned}
&T_M(a,b) = \min\{a,b\}, S_M(a,b) = \max\{a,b\} \\
&\vartheta_M(a,b) = \begin{cases} 1, & a \le b \\ b, & a > b \end{cases} \sigma_M(a,b) = \begin{cases} 0, & a \ge b \\ b, & a < b. \end{cases} \qquad (2)
\end{aligned}$$

See [4], [38], [39], and [46] for more information on fuzzy operators and their properties.

The aforementioned models of fuzzy rough sets can be used in classification and regression analysis. In this paper, we focus on classification tasks. The derived results can be easily extended to regression analysis. Given a classification learning task $DT = \langle U, A, D\rangle$, where D is the decision attribute dividing the objects into classes $d_1, d_2, \ldots, d_N$, R is a fuzzy relation computed with $B \subseteq A$ and a kernel function. Then, triangular norm $T_{\cos}$ should be introduced because the derived fuzzy relations are $T_{\cos}$-fuzzy equivalence relations [40]. As $x \in U$, we have

$$\begin{aligned}
\underline{R_S}d_i(x) &= \inf_{y\in U-d_i} (1 - R(x,y)) \\
\overline{R_T}d_i(x) &= \sup_{y\in d_i} R(x,y) \\
\underline{R_\theta}d_i(x) &= \inf_{y\in U-d_i} (\sqrt{1-R^2(x,y)}) \\
\overline{R_\sigma}d_i(x) &= \sup_{y\in d_i} (1 - \sqrt{1-R^2(x,y)}). \qquad (3)
\end{aligned}$$

Given $x \in d_i$, it is easy to show that $1 \ge \underline{R_S}d_i(x) \ge 0$ ($1 \ge \underline{R_\theta}d_i(x) \ge 0$). $\underline{R_S}d_i(x)$ ($\underline{R_\theta}d_i(x)$) is considered as the level at which x certainly belongs to d_i, and $\overline{R_T}d_i(x)$ ($\overline{R_\sigma}d_i(x)$) is the degree at which x possibly belongs to d_i. Obviously, $\underline{R_S}d_i$ ($\underline{R_\theta}d_i(x)$) and $\overline{R_T}d_i$ ($\overline{R_\sigma}d_i(x)$) are two fuzzy subsets. We have $\underline{R_S}d_i \subseteq d_i \subseteq \overline{R_T}d_i$ ($\underline{R_\theta}d_i \subseteq d_i \subseteq \overline{R_\sigma}d_i$). $\underline{R_S}d_i$ ($\underline{R_\theta}d_i$) is also called fuzzy positive region of d_i, and $BND_R(d_i) = \overline{R_S}d_i - \underline{R_T}d_i$ ($BND_R(d_i) = \overline{R_\theta}d_i - \underline{R_\sigma}d_i$) is called fuzzy boundary of d_i.

In classification learning, it is natural to desire that the membership of each sample belonging to its decision is as large as possible. A function, which is called dependence of D on B, is defined as

$$\gamma_B(D) = \frac{|\cup_{i=1}^N \underline{R}d_i|}{|U|} \qquad (4)$$

where R is the fuzzy similarity relation induced with B, and $|\bullet|$ is the fuzzy cardinality of fuzzy sets. Dependence reflects the capability of B in approximating D for it is used to measure quality of features in feature selection, attribute reduction, and rule learning. Dependence plays the key role in rough-set-based learning techniques [1], [3], [5], [9], [18], [19], [40]. The robustness of these learning algorithms significantly depends on the properties of the dependence function.

B. Problem Description

Given $DT = \langle U, A, D\rangle$, d_i is the subset of samples with decision label i, and R is a fuzzy similarity function, such as Gaussian function: $R(x,y) = \exp(-\|x-y\|^2/\sigma)$, where $\|x-y\|$ is the distance between x and y.

If Gaussian kernel function is used to compute the similarity and $\phi(x)$ is the corresponding nonlinear mapping, we have $\|\phi(x) - \phi(y)\|^2 = \phi(x)\phi(x) + \phi(y)\phi(y) - \phi(x)\phi(y) - \phi(y)\phi(x)$. According to the properties of kernel functions, we have $\|\phi(x) - \phi(y)\|^2 = 2 - 2R(x,y)$ [40]. Therefore, $1 - R(x,y)$ is the squared distance between x and y in the kernel feature space. Thus, $\underline{R_S}d_i(x)$ is the minimal distance between x and the samples out of d_i. A similar result can also be derived from $\underline{R_\theta}d_i(x)$.

In Relief-based feature evaluation [41], the algorithm searches two nearest neighbors of each sample x: one from the same class, called nearest hit, which is denoted by $NH(x)$, and the other from the different classes, called nearest miss, which is denoted by $NM(x)$. Then, features are evaluated by

$$\theta = \frac{1}{n}\sum_x \|x - NM(x)\| - \|x - NH(x)\| \qquad (5)$$

where n is the number of samples. In fuzzy rough sets, the membership of x to fuzzy lower approximation can be written as $\underline{R_S}d_i(x) = 1 - R(x, NM(x))$ or $\underline{R_\theta}d_i(x) = \sqrt{1 - R^2(x, NM(x))}$.

As to the Gaussian function

$$\underline{R_S}d_i(x) = 1 - \exp(-\|x - NM(x)\|^2/\sigma) \qquad (6)$$

and

$$\underline{R_\theta}d_i(x) = \sqrt{1 - \exp^2(-\|x - NM(x)\|^2/\sigma)}. \qquad (7)$$

As $\gamma_B(D) = |\cup_i^N \underline{R}d_i|/|U| = |\cup_i^N \underline{R}d_i|/n = \frac{1}{n}\sum_{x\in U} \underline{R}d_i(x)$, dependence is the mean of the fuzzy lower approximation memberships, which are determined by the distance between objects and their nearest miss.

Now, we discuss the influence of attribute noise. We consider Gaussian noise as it widely exists in real-world applications. The noisy lower approximation is computed as $\underline{R_S^n}d_j(x) = \underline{R_S}d_j(x) + \beta \cdot N(0,1)$, where $N(0,1)$ is a standard normal distribution. Then, the noisy dependence is

$$\begin{aligned}
\gamma_B^N(D) &= \sum_x \underline{R_S^n}d_j(x)/n \\
&= \sum_x \left[\underline{R_S}d_j(x) + \beta \cdot N(0,1)\right]/n \\
&= \sum_x \underline{R_S}d_j(x)/n + \beta \cdot N(0,1). \qquad (8)
\end{aligned}$$

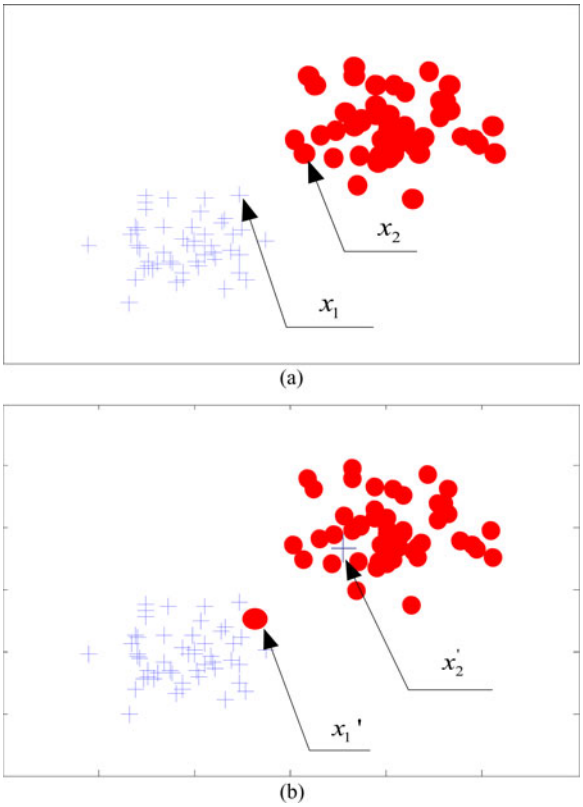


Fig. 1. Two toy classification tasks. (a) Clean dataset. (b) Dataset with two noisy samples.

$\gamma_B^N(D) = \gamma_B(D) + \beta \cdot N(0,1)$. The conclusion shows that the attribution noise might not cause significant variation of dependence if noise satisfies normal distribution.

Then, we consider the influence of class noise on dependence. Fig. 1 shows a binary classification task, where "+" stands for samples from Class 1, and "•" means samples from Class 2. Fig. 1(a) presents a clean dataset, where two classes of samples are well separated. The near miss of all the samples from Class 1 is x_2, whereas the nearest miss of samples from Class 2 is x_1. In this case, dependence is computed as

$$r = \frac{1}{n}\left\{\sum_{x_i \in C_1}(1 - R(x_i, x_2)) + \sum_{x_j \in C_2}(1 - R(x_j, x_1))\right\}. \tag{9}$$

However, in Fig. 1(b), there are two noisy samples: x_1' and x_2'. In the new dataset, the near miss of samples from Class 1 is x_1' except x_2', while the nearest miss of samples from Class 2 is x_2' except x_1'. Now, dependence is

$$r' = \frac{1}{n}\left\{\sum_{x_i \in C_1, x_i \neq x_2'}(1 - R(x_i, x_1')) + [1 - R(x_2', NM(x_2'))] + \sum_{x_j \in C_2, x_j \neq x_1'}(1 - R(x_j, x_2')) + [1 - R(x_1' NM(x_1'))]\right\}. \tag{10}$$

As the distances between the samples and their nearest miss decrease greatly in the second case, dependence in the new case will significantly reduce if the noisy samples exist. The previous analysis shows that dependence defined in fuzzy rough sets is not sensitive to attribute noise, but sensitive to class noise. Several mislabeled samples would alter dependence greatly as they compute lower approximation with a sensitive statistic. A robust rough set model should overcome this problem.

III. Robust Models of Fuzzy Rough Sets

In this section, we introduce some improved models of fuzzy rough sets. These models are claimed to be robust in dealing with noisy tasks.

A. β-Precision Fuzzy Rough Sets

The β-precision fuzzy rough set (β-PFRS) model was introduced to overcome the problem that VPRS cannot handle numerical features [32]. We first present the definitions of β-precision quasi-T-norm and β-precision quasi-T-conorm [32].

Definition 1: Let T be a T-norm operator, i.e., $T : I \times I \to I$, which can be extended to the N-dimensional case with the associative property, i.e., $T : I^N \to I$. Its corresponding β-precision quasi-T-norm, i.e., T_β $(\beta \in [0, 0.5])$, should be a mapping, i.e., $T_\beta : I^N \to I$, such that $\forall X = (x_1, x_2, \ldots, x_N) \in I^N$ expressed in descending order, i.e., $T_\beta(X) = T(x_1, x_2, \ldots, x_n)$ with $n = \max_k\{k \in [0, 1, 2, \ldots, N] : k \leq \sum_1^N x_i(1-\beta)\}$.

Definition 2: Let S be a T-conorm operator, i.e., $S : I \times I \to I$, which can be extended to the N-dimensional case with the associative property, i.e., $S : I^N \to I$. Its corresponding β-precision quasi-T-conorm, i.e., S_β $(\beta \in [0, 1])$, should be a mapping, i.e., $S_\beta : I^N \to I$, such that $\forall X = \{x_1, x_2, \ldots, x_N\} \in I^N$ expressed in ascending order, i.e., $S_\beta(X) = S(x_1, x_2, \ldots, x_n)$ with $n = \max_k\{k \in [0, 1, 2, \ldots, N] : k \leq \sum_1^N (1 - x_i)(1-\beta)\}$.

Definition 3: With the aforementioned triangular operators, we can obtain the β-precision version of fuzzy rough set model:

$$\begin{aligned}
\underline{R_S}X(x) &= T_{\beta_{y \in U}} S(N(R(x,y)), X(y)) \\
\overline{R_T}X(x) &= S_{\beta_{y \in U}} T(R(x,y), X(y)) \\
\underline{R_\vartheta}X(x) &= T_{\beta_{y \in U}} \vartheta(R(x,y), X(y)) \\
\overline{R_\sigma}X(x) &= S_{\beta_{y \in U}} \sigma(N(R(x,y)), X(y)).
\end{aligned} \tag{11}$$

In fact, this model will degrade to the one proposed in [32] if we use operators min and max:

$$\begin{aligned}
\underline{R_S}X(x) &= \min_{\beta_{y \in U}} \max(1 - R(x,y), X(y)) \\
\overline{R_T}X(x) &= \max_{\beta_{y \in U}} \min(R(x,y), X(y)).
\end{aligned} \tag{12}$$

The authors extended Ziarko's VPRS to the Dubois and Prade fuzzy rough set framework by replacing the maximum and minimum operators used to calculate the inclusion indexes with their β-precision counterparts. Here, we show that β-PFRS cannot degenerate to VPRS, although this model is robust to noise.

Given a classification task $DT = \langle U, A, D\rangle$, d_i is one class of samples. Then, $\forall x \in d_i$, we have

$$\underline{R}d_i(x) = \min_{\beta_{y\in U}} \max(1 - R(x,y), d_i(y)). \tag{13}$$

If $y \in d_i$, $d_i(y) = 1$, $\max(1 - R(x,y), d_i(y)) = 1$. If $y \notin d_i$, $\max(1 - R(x,y), d_i(y)) = 1 - R(x,y)$. As a whole, $\underline{R}d_i(x) = \min_{\beta_{y\notin d_i}}(1 - R(x,y))$. Suppose there are v samples $\{x_i^1, x_i^2, \ldots, x_i^v\}$ out of d_i and $g_i^j = 1 - R(x, x_i^j)$ is ranked in descending order. That is to say, $j_1 \le j_2 \Longrightarrow g_i^{j_1} \ge g_i^{j_2}$. $\underline{R}d_i(x) = \min_\beta(g_i^1, g_i^2, \ldots, g_i^v) = \min(g_i^1, g_i^2, \ldots, g_i^u)$, where $u = \max_k\{k \in [0,1,2,\ldots,v] : k \le \sum_1^v g_i^j(1-\beta)\}$.

β-precision rough sets do not compute the lower approximation based on the minimal g_i^j because it may be computed by the mislabeled sample. It computes low approximations based on g_i^u which is the minimal one after some little g_i are removed. Essentially, g_i^u can be considered as the k-trimmed minimum, where the value of k depends on parameter β.

If x is a normal sample, $g_i^u, g_i^{u+1}, \ldots, g_i^v$ are computed with class noisy samples, and $g_i^1, g_i^2, \ldots, g_i^v$ are calculated with normal samples; then, $\underline{R}d_i(x) = 1 - R(x, g_i^u)$ can correctly reflect the membership of x to the lower approximation of its decision. This way, the mislabeled samples are omitted in computing the approximation of a normal sample.

If x is a mislabeled sample, there might be a lot of samples belonging to different classes around x. In this case, even some of them are omitted in computing lower approximation, $\underline{R}d_i(x) = 1 - R(x, g_i^u)$ is still small enough. Therefore, the noisy sample still gets a small membership. If there are a lot of mislabeled samples, the dependence function will return a small value.

The previous analysis shows that β-precision rough sets are robust to class noise.

B. Variable Precision Fuzzy Rough Sets

The first model considering class noise was developed in [27] and extended in [42]. However, the model of VPRS gets popular in real-world applications and is widely used in feature selection and rule extraction [28].

Definition 4: Given a classification task $DT = \langle U, A, D\rangle$, let X be a subset of samples. The lower and upper approximations of X with respect to B are defined as

$$\begin{aligned} \underline{B_\beta}X &= \{[x_i]_B | IN([x_i]_B, X) \ge 1 - \beta\} \\ \overline{B_\beta}X &= \{[x_i]_B | IN([x_i]_B, X) \ge \beta\} \end{aligned} \tag{14}$$

where $IN(A,B) = |A \cap B|/|B|$, and $0 \le \beta \le 0.5$.

In this model, if majority of samples in $[x_i]_B$ belong to X, we group $[x_i]_B$ into the lower approximation of X, regardless of the class label of x_i. However, $[x_i]_B$ is grouped into the lower approximation of X if and only if all the samples in $[x_i]_B$ are elements of X according to Pawlak rough sets. As per the original model, $[x_i]_B$ is computed as a boundary set if there is a mislabeled sample in $[x_i]_B$; however, the mislabeled sample will be ignored as to VPRS. The advantage of VPRS is the ability of being robust to class noise, and the disadvantage is that the mislabeled samples are not reflected in dependence. Two datasets, i.e., one with mislabeled samples and the other without noise, would produce the same value of dependence. Therefore, the dependence function in VPRS cannot reflect any noise information if the noise level in each equivalence is lower than β. Obviously, it is not reasonable in practice if we select the noisy dataset, instead of the clean one if we cannot distinguish them using dependence. In addition, the model of VPRS is constructed with equivalence classes induced by nominal features; therefore, it cannot directly be used in numerical and fuzzy information.

Given $X = \{x_1, x_2, \ldots, x_n\}$, the lower approximation of VPFRS was defined as

$$\mu_{\underline{R_u}}F(X_i) = \begin{cases} \inf_{x\in S_{i_u}} \vartheta(\mu_{X_i}(x), \mu_F(x)) \\ \text{if } \exists \alpha_u = \sup\{\alpha \in (0,1] : e_\alpha(X_i, F) \le 1 - u\} \\ 0, \text{otherwise} \end{cases} \tag{15}$$

where $S_{i_u} = supp(X_i \cap X^F_{i_{\alpha_u}})$ and $e_\alpha(X_i, F) = 1 - \frac{|X_i \cap X^F_{i_\alpha}|}{|X_i|} = 1 - \frac{|X_i \cap (X_i \cap F)_\alpha|}{|X_i|}$.

We see that VPFRS is robust to mislabeled samples as $\mu_{\underline{R_u}}F(X_i)$ is computed with the samples which satisfy $\mu_F(x) > \alpha_u$.

C. Vaguely Quantified Rough Sets

In 2007, Cornelis *et al.* introduced vague quantifiers to soften the definitions for upper and lower approximations in VPRS and β-PFRS [34]. VPRS uses the rough membership function

$$R_X(x_i) = \frac{|[x_i]_R \cap X|}{|[x_i]_R|}. \tag{16}$$

Given $0 \le u \le l \le 1$, the lower and upper approximations in VPRS were defined as

$$x_i \in \underline{R}X, \quad \text{if} \quad R_X(x_i) > l, \quad x_i \in \overline{R}X, \quad \text{if} \quad R_X(x_i) \ge u. \tag{17}$$

Obviously, this definition leads to crisp boundary of approximations. Based on such observation, vague quantifiers were introduced. An example of a fuzzy quantifier is the following parametrized formula. For $0 \le \alpha \le \beta \le 1$, and $x \in [0,1]$

$$Q_{(\alpha,\beta)}(x) = \begin{cases} 0, & x \le \alpha \\ \dfrac{2(x-\alpha)^2}{(\beta-\alpha)^2}, & \alpha \le x \le \dfrac{\alpha+\beta}{2} \\ 1 - \dfrac{2(x-\alpha)^2}{(\beta-\alpha)^2}, & \dfrac{\alpha+\beta}{2} \le x \le \beta \\ 1, & x \ge \beta. \end{cases} \tag{18}$$

This function can be used to reflect the vague quantifiers, such as *some* or *most* from nature language, when different parameter values are used, such as $Q_{(0.1,0.6)}(x)$ and $Q_{(0.2,1)}(x)$. If x is the fuzzy inclusion of two fuzzy sets, this function can be used to extend the fuzzy rough sets to vague quantified fuzzy rough

sets:

$$\underline{R_{Q_l}}X(y) = Q_l\left(\frac{|R_y \cap X|}{|R_y|}\right)$$
$$\overline{R_{Q_u}}X(y) = Q_u\left(\frac{|R_y \cap X|}{|R_y|}\right) \quad (19)$$

where R_y is the fuzzy similarity class of y with respect to fuzzy relation R, and Q_l and Q_u are value quantifiers for fuzzy lower and upper approximations, respectively. The performance of this model depends on the selection of value quantifiers and the setting of parameters.

In VQRS, the lower and upper approximations are the functions of inclusion degree of samples' fuzzy class in the set to be approximated. The functions are given by the vague quantifiers. In this context, they do not agree with the basic requirements of rough sets as the model does not have the basic properties of rough sets [4]. VQRS computes the lower and upper approximations based on inclusions, which are robust statistics; therefore, this model should be robust to noise because a few noisy samples have little influence on the inclusion degrees of samples.

D. Fuzzy Variable Precision Rough Sets

In 2009, Zhao *et al.* proposed a model of FVPRS [18]. Some delicate mathematical properties of the model was proved in [18]. We here discuss the robustness of this model.

Definition 5: Given $DT = \langle U, A, D\rangle$, $B \subseteq A$ generates a fuzzy similarity relation over U and X is a fuzzy set in U. The fuzzy lower and upper approximation operators with a variable precision parameter $\alpha \in [0,1]$ are defined as $\forall x \in U$

$$\underline{R_S}_\alpha X(x) = \inf_{X(y)\le\alpha} S(N(R(x,y)),\alpha) \wedge \inf_{X(y)>\alpha} S(N(R(x,y)),X(y))$$
$$\overline{R_T}_\alpha X(x) = \sup_{X(y)\ge N(\alpha)} T(R(x,y),N(\alpha)) \vee \sup_{X(y)<N(\alpha)} T(R(x,y),X(y)$$
$$\underline{R_\vartheta}_\alpha X(x) = \inf_{X(y)\le\alpha} \vartheta(R(x,y),\alpha) \wedge \inf_{X(y)>\alpha} \vartheta(R(x,y),X(y))$$
$$\overline{R_\sigma}_\alpha X(x) = \sup_{X(y)\ge N(\alpha)} \sigma(N(R(x,y)),N(\alpha)) \vee \sup_{X(y)<N(\alpha)} \sigma(N(R(x,y)),X(y)).$$

Definition 6: Let X be a class of samples. The aforementioned operators can be rewritten as

$$\underline{R_S}_\alpha X(x) = \inf_{X(y)=0} S(N(R(x,y)),\alpha)$$
$$\overline{R_T}_\alpha X(x) = \sup_{X(y)=1} T(R(x,y),N(\alpha))$$
$$\underline{R_\vartheta}_\alpha X(x) = \inf_{X(y)=0} \vartheta(R(x,y),\alpha)$$
$$\overline{R_\sigma}_\alpha X(x) = \sup_{X(y)=1} \sigma(N(R(x,y)),N(\alpha)). \quad (20)$$

Now, we take $\underline{R_S}_\alpha X(x)$ as an example to discuss the robustness of these operators. The same conclusion can be derived for other operators. We consider the triangular conorm "max" and $N(\alpha) = 1-\alpha$ here. $\underline{R_S}_\alpha X(x) = \inf_{y\notin X} \max\{1 - R(x,y), \alpha\}$.

Case 1: suppose there is a sample $y \notin X$ such that $1 - R(x,y)$ is the minimal one among the samples coming from $U - X$. That is to say, y is the nearest miss of x: $NM(x)$. If $1 - R(x,y) \ge \alpha$, $\underline{R_S}_\alpha X(x) = 1 - R(x,y)$; otherwise, $\underline{R_S}_\alpha X(x) = \alpha$. If x is a mislabeled sample, the nearest miss y of x is a normal sample, and y is close to x, such that $1 - R(x,y) < \alpha$, then x should get a small membership as it is a mislabeled sample. However, FVPRS assigns α to it.

Case 2: Suppose x is a normal sample. However, the nearest miss of x is a mislabeled sample. That is to say there exists a mislabeled sample $y \in U - X$ close to x, which leads to $1 - R(x,y) < \alpha$. Then, FVPRS sets $\underline{R_S}_\alpha X(x)$ as α.

If users give a large value to α, the first case is not rational because class-noisy samples should take small memberships. However, the second case is not reasonable should α take a small value for normal samples, producing large memberships. With the previous analysis, we can see that there exists a contradiction for the model of FVPRS in dealing with class-noisy tasks. FVPRS may be effective if we just consider attribute noise, which just perturbs the membership values in a small arrangement.

E. Soft Fuzzy Rough Sets

In 2010, Hu *et al.* introduced a new robust model of fuzzy rough sets, which are called soft fuzzy rough sets, where soft threshold was used to compute fuzzy lower and upper approximations [53].

Assume $X \subseteq U$ and $x \in X$. The fuzzy lower approximation of x to X can be considered as the distance from sample point x and the subset of samples $U - X$. The distance between x and X is computed as the minimal distance between the point and points in X. That is

$$\|x - X\| = \min_{y\in X} \|x - y\|.$$

As we know, the statistic min is sensitive to noise. If there is a noisy sample $x' \in X$ close to x, the distance between x and X is determined by the noisy sample, which leads to the sensitivity of fuzzy rough sets. Therefore, soft distance was introduced in [53].

Definition 7: Given an object x and a set of objects X, the soft distance between x and X is defined as

$$SD(x,X) = \arg\sup_{\|x-y\|\ y\in X}\{\Delta(x,y) - \beta m_X\} \quad (21)$$

where β is a penalty factor, and $m_X = |\{y_i | \|x - y_i\| < \|x - y\|\}|$.

Then, with the soft distance, we can give the definition of soft fuzzy rough sets.

Definition 8: Let U be a nonempty universe, and let R be a fuzzy similarity relation on U and $X \subseteq U$. The soft fuzzy lower and upper approximations of X are defined as

$$\begin{cases} \underline{R^S}(X)(x) = SD(x, U - X) \\ \overline{R^S}(X)(x) = SD(x, X) \end{cases} \tag{22}$$

where

$$\forall y \in X, \|x - y\| = 1 - R(x, y). \tag{23}$$

Soft fuzzy rough sets introduced the soft threshold technique in computing lower and upper approximations. Here, β is a key parameter. If β is large, fewer samples are omitted in computing soft distance. If β approaches to a very large number, the model of soft fuzzy rough sets degrades to classical fuzzy rough sets. If we specify a small value for β, then some samples in X close to x would be disregarded to obtain a larger soft distance. In essence, soft fuzzy rough sets compute the lower and upper approximations based on the *k*th nearest neighbor from $U - X$, where k is determined by parameter β.

With the previous analysis, we know that soft fuzzy rough sets share the common idea with β-PFRS. When they compute the fuzzy lower and upper approximations, they do not use the nearest neighbors except for the *k*th nearest neighbor. This way, the noisy samples may be disregarded when computing approximations. However, different models use different strategies to determine the value of k. This is the key difference of these models. In real-world application, the parameters used in the models have complex interaction ways with noise. Therefore, it is usually difficult to obtain an optimal value. In this case, a simple and understandable model is easy to be accepted. In the next section, we introduce some models of robust fuzzy rough sets based on a general definition of robust nearest neighbor.

IV. Fuzzy Rough Sets Based on Robust Nearest Neighbor

The previous discussion shows that both $\underline{R_S} d_i(x)$ or $\underline{R_\vartheta} d_i(x)$ depends on the nearest miss of x, i.e., the nearest sample from different classes of *x*. As we know, the statistics of minimum and maximum are very sensitive to noisy samples. Just one noisy sample would change the minimum or maximum of a random variable. The sensitiveness of these statistics leads to the poor performance of fuzzy rough sets in dealing with noisy datasets. Some robust statistics should be introduced to substitute the operators of minimum and maximum in the fuzzy rough set model.

Definition 9: Given a random variable X and its *n* samples $x_1, x_2, \ldots, x_n$ sorted in the ascending order, the k-trimmed minimum of X is x_{k+1}; the k-trimmed maximum of X is x_{n-k-1}; k-mean minimum of X is $\sum_{i=1}^{k} x_i/k$; k-mean maximum of X is $\sum_{i=n-k}^{n} x_i/k$, and k-median minimum of X is median$(x_1, \ldots, x_k)$; k-mean maximum of X is median$(x_{n-k}, \ldots, x_n)$, denoted by $\min_{k-\text{trimmed}}(X)$, $\max_{k-\text{trimmed}}(X)$, $\min_{k-\text{mean}}(X)$, $\max_{k-\text{mean}}(X)$, $\min_{k-\text{median}}(X)$, and $\max_{k-\text{median}}(X)$, respectively.

Definition 10: Given $DT = \langle U, A, D \rangle$, R is a fuzzy similarity relation induced by $B \subseteq C$ and $R(x, y)$ monotonously decreases with their distance $\|x - y\|$. If d_i is one class of samples labeled with i and $x \in d_i$, then the robust fuzzy rough operators are defined as

$$\begin{aligned} \underline{R_S}_{k-\text{trimmed}} d_i(x) &= \min_{y \notin d_{i\,k-\text{trimmed}}} 1 - R(x, y) \\ \overline{R_T}_{k-\text{trimmed}} d_i(x) &= \max_{y \in d_{i\,k-\text{trimmed}}} R(x, y) \\ \underline{R_\vartheta}_{k-\text{trimmed}} d_i(x) &= \min_{y \notin d_{i\,k-\text{trimmed}}} \sqrt{1 - R^2(x, y)} \\ \overline{R_\sigma}_{k-\text{trimmed}} d_i(x) &= \max_{y \in d_{i\,k-\text{trimmed}}} 1 - \sqrt{1 - R^2(x, y)} \end{aligned} \tag{24}$$

$$\begin{aligned} \underline{R_S}_{k-\text{mean}} d_i(x) &= \min_{y \notin d_{i\,k-\text{mean}}} 1 - R(x, y) \\ \overline{R_T}_{k-\text{mean}} d_i(x) &= \max_{y \in d_{i\,k-\text{mean}}} R(x, y) \\ \underline{R_\vartheta}_{k-\text{mean}} d_i(x) &= \min_{y \notin d_{i\,k-\text{mean}}} \sqrt{1 - R^2(x, y)} \\ \overline{R_\sigma}_{k-\text{mean}} d_i(x) &= \max_{y \in d_{i\,k-\text{mean}}} 1 - \sqrt{1 - R^2(x, y)} \end{aligned} \tag{25}$$

$$\begin{aligned} \underline{R_S}_{k-\text{median}} d_i(x) &= \min_{y \notin d_{i\,k-\text{median}}} 1 - R(x, y) \\ \overline{R_T}_{k-\text{median}} d_i(x) &= \max_{y \in d_{i\,k-\text{median}}} R(x, y) \\ \underline{R_\vartheta}_{k-\text{median}} d_i(x) &= \min_{y \notin d_{i\,k-\text{median}}} \sqrt{1 - R^2(x, y)} \\ \overline{R_\sigma}_{k-\text{median}} d_i(x) &= \max_{y \in d_{i\,k-\text{median}}} 1 - \sqrt{1 - R^2(x, y)}. \end{aligned} \tag{26}$$

The aforementioned models do not compute the lower and upper approximations with respect to the nearest samples as they might be outliers. These new models use k-trimmed or the mean or the median of k nearest samples to compute the membership of fuzzy approximations. This way, the variation of approximations caused by outliers is expected to be reduced; thus, the new models may be robust.

Given a binary classification task, $x \in d_1$ is a normal sample, and $y_1 \in d_2$ is an outlier close to x such that $R(x, y_1) = 0.9$. While as a normal sample, $y_2 \in d_2$ is the second nearest sample of x from d_2, and $R(x, y_2) = 0.2$. As per the classical fuzzy rough set model, $\underline{R_S} d_1(x) = 1 - R(x, y_1) = 1 - 0.9 = 0.1$. However, if we use the 1-trimmed model, $\underline{R_S}_{1-\text{trimmed}} d_1(x) = 1 - R(x, y_2) = 0.8$. This way, the noisy sample is ignored in the new model. At the same time, assume $x_1 \in d_1$ is the second nearest sample of y_1, and $R(x_1, y_1) = 0.88$. According to the classical model, $\underline{R_S} d_2(y_1) = 1 - R(x, y_1) = 1 - 0.9 = 0.1$ and as per the 1-trimmed model, $\underline{R_S}_{1-\text{trimmed}} d_2(x) = 1 - R(x_1, y_1) = 0.12$. We see that although the nearest sample x is ignored, y_1 still obtains a small value of membership. In fact, the membership should be small enough since y_1 is a noisy sample. This example shows that the proposed model can not only reduce the influence of noisy samples on computation of

approximations of normal samples but can recognize the noisy samples and give small memberships to them as well.

Now, we discuss the properties of these robust models. For simplification, we here just discuss the models defined with k-trimmed minimum and maximum operators. Some of the following properties are easily extended to other cases.

Proposition 1: Given $DT = \langle U, A, D\rangle$, R is a fuzzy similarity relation induced by $B \subseteq C$, and $R(x, y)$ monotonously decreases with their distance $\Delta(x, y)$. If d_i is a class of samples labeled with i and $x \in U$ and k is a positive integer, we have

$$\begin{aligned}
&\underline{R_S}_{k-\text{trimmed}} d_i(x) \geq \underline{R_S} d_i(x)\\
&\overline{R_T}_{k-\text{trimmed}} d_i(x) \leq \overline{R_T} d_i(x)\\
&\underline{R_\vartheta}_{k-\text{trimmed}} d_i(x) \geq \underline{R_\vartheta} d_i(x)\\
&\overline{R_\sigma}_{k-\text{trimmed}} d_i(x) \leq \overline{R_\sigma} d_i(x).
\end{aligned} \tag{27}$$

If $k = 0$, the equality holds.

Proof: Assume that $y_1, y_2, \ldots, y_k, \ldots, y_N$ are the samples from the different classes of x, and we have $1 - R(x, y_1) \leq 1 - R(x, y_2) \leq \cdots \leq 1 - R(x, y_{k+1}) \leq \cdots \leq 1 - R(x, y_N)$. $\underline{R_S} d_i(x) = 1 - R(x, y_1)$ and $\underline{R_S}_{k-\text{trimmed}} d_i(x) = 1 - R(x, y_{k+1})$. Therefore, it is easy to obtain $\underline{R_S}_{k-\text{trimmed}} d_i(x) \geq \underline{R_S} d_i(x)$.

Suppose $y_1, y_2, \ldots, y_k, \ldots, y_N$ are the samples from d_i, and $1 - R(x, y_1) \leq 1 - R(x, y_2) \leq \cdots \leq 1 - R(x, y_{k+1}) \leq \cdots \leq 1 - R(x, y_N)$. $\overline{R_T} d_i(x) = \max_{y \in d_i} R(x, y_i) = R(x, y_1)$, while $\overline{R_T}_{k-\text{trimmed}} d_i(x) = \max_{y \in d_{i\,k-\text{trimmed}}} R(x, y_{k+1})$. As we know $1 - R(x, y_1) \leq 1 - R(x, y_{k+1})$, $R(x, y_1) \geq R(x, y_{k+1})$; therefore, $\overline{R_T}_{k-\text{trimmed}} d_i(x) \leq \overline{R_T} d_i(x)$. Analogically, we can also obtain $\underline{R_\vartheta}_{k-\text{trimmed}} d_i(x) \geq \underline{R_\vartheta} d_i(x)$; $\overline{R_\sigma}_{k-\text{trimmed}} d_i(x) \leq \overline{R_\sigma} d_i(x)$.

Proposition 2: Given $DT = \langle U, A, D\rangle$, R is a fuzzy similarity relation induced by $B \subseteq C$, and $R(x, y)$ monotonously decreases with their distance $\Delta(x, y)$. d_i is a class of samples labeled with i and $x \in U$. Provided that k_1 and k_2 are two positive integers and $k_1 \leq k_2$, we have

$$\begin{aligned}
&\underline{R_S}_{k_1-\text{trimmed}} d_i(x) \leq \underline{R_S}_{k_2-\text{trimmed}} d_i(x)\\
&\overline{R_T}_{k_1-\text{trimmed}} d_i(x) \geq \overline{R_T}_{k_2-\text{trimmed}} d_i(x)\\
&\underline{R_\vartheta}_{k_1-\text{trimmed}} d_i(x) \leq \underline{R_\vartheta}_{k_2-\text{trimmed}} d_i(x)\\
&\overline{R_\sigma}_{k_1-\text{trimmed}} d_i(x) \geq \overline{R_\sigma}_{k_2-\text{trimmed}} d_i(x).
\end{aligned} \tag{28}$$

Proof: The proof is straightforward.

Proposition 3: Given $DT = \langle U, A, D\rangle$, R_1 and R_2 are two fuzzy similarity relations over U induced by B_1 and B_2, respectively, and $R_1 \subseteq R_2$. Since $x \in d_i$ and k is a positive integer, we have

$$\begin{aligned}
&\underline{R_{1S}}_{k-\text{trimmed}} d_i(x) \geq \underline{R_{2S}}_{k-\text{trimmed}} d_i(x)\\
&\overline{R_{1T}}_{k-\text{trimmed}} d_i(x) \leq \overline{R_{2T}}_{k-\text{trimmed}} d_i(x)\\
&\underline{R_{1\vartheta}}_{k-\text{trimmed}} d_i(x) \geq \underline{R_{2\vartheta}}_{k-\text{trimmed}} d_i(x)\\
&\overline{R_{1\sigma}}_{k-\text{trimmed}} d_i(x) \leq \overline{R_{2\sigma}}_{k-\text{trimmed}} d_i(x).
\end{aligned} \tag{29}$$

Proof: Assume r_j^1 and r_j^2 are the fuzzy similarity degrees between x and sample x_j out of d_i in terms of R_1 and R_2, respectively. Since $R_1 \subseteq R_2$, for $x_j \notin d_i$, we have $r_j^1 \leq r_j^2$. Suppose x_{k_1} is the sample such that $\underline{R_{1S}}_{k-\text{trimmed}} d_i(x) = 1 - R_1(x, x_{k_1})$, and x_{k_2} is the sample such that $\underline{R_{2S}}_{k-\text{trimmed}} d_i(x) = 1 - R_2(x, x_{k_2})$. It is easy to obtain that $R_1(x, x_{k_1}) \leq R_2(x, x_{k_2})$. Thus, $\underline{R_{1S}}_{k-\text{trimmed}} d_i(x) \geq \underline{R_{2S}}_{k-\text{trimmed}} d_i(x)$.

On the other side, we assume that r_j^1 and r_j^2 are the fuzzy similarity degrees between x and sample x_j in d_i in terms of R_1 and R_2. Since $R_1 \subseteq R_2$, for $x_j \in d_i$, we have $r_j^1 \leq r_j^2$. Suppose x_{k_1} is the sample such that $\overline{R_{1T}}_{k-\text{trimmed}} d_i(x) = R_1(x, x_{k_1})$, and x_{k_2} is the sample such that $\overline{R_{2T}}_{k-\text{trimmed}} d_i(x) = R_2(x, x_{k_2})$. We can also obtain that $R_1(x, x_{k_1}) \leq R_2(x, x_{k_2})$. Thus, $\overline{R_{1T}}_{k-\text{trimmed}} d_i(x) \leq \overline{R_{2T}}_{k-\text{trimmed}} d_i(x)$.

Analogically, we can derive that $\underline{R_{1\vartheta}}_{k-\text{trimmed}} d_i(x) \geq \underline{R_{2\vartheta}}_{k-\text{trimmed}} d_i(x)$ and $\overline{R_{1\sigma}}_{k-\text{trimmed}} d_i(x) \leq \overline{R_{2\sigma}}_{k-\text{trimmed}} d_i(x)$.

The fuzzy lower and upper operators are widely used in evaluating features or extracting rules from data. In this case, relations between objects are generated with features. Features are added one by one in forward algorithms, and correspondingly, the similarity degrees between objects get increasingly weaker. Then, the lower approximation of decision classes becomes larger, while the upper approximation becomes smaller. Therefore, the classification boundary, which is the inconsistent region of classification, is reduced. The average memberships of samples to fuzzy lower approximations of their decision classes is defined as fuzzy dependence of decision on the corresponding attributes:

$$\gamma_B(D) = \frac{1}{n} |\cup_{i=1}^{N} \underline{R} d_i|, \tag{30}$$

where $\underline{R} d_i$ is the fuzzy lower approximation of class d_i under fuzzy similarity relation R, and $|\cdot|$ is the fuzzy cardinality of a fuzzy set, which is computed with $\sum_{x \in d_i} \underline{R} d_i(x)$.

Besides the statistics used previously, some other robust statistics can also be introduced, such as quartile. In the aforementioned models, we should specify the value of k. The theory of exploratory data analysis provides some theoretical results [43].

Given $DT = \langle U, A, D\rangle$, $x \in d_i$, $r(x, x_j)$ is the similarity between x and x_j out of d_i. Then, we rank the similarity degrees in the ascending order, denoted by $r_1, r_2, \ldots, r_J$. The lower and upper fourths are r_L and r_U. We compute $d_F = r_U - r_L$. Then, the similarity degrees less than $r_L - 1.5 d_F$ or greater than $r_U + 1.5 d_F$ are viewed as outliers. If $r(x, x_j)$ satisfies Gaussian distribution, the probability of samples less than $r_L - 1.5 d_F$ or greater than $r_L + 1.5 d_F$ is 0.00698. Correspondingly, the number of outliers can be computed as $k = 0.4 + 0.007 \times J$. This way, we can automatically compute the value of k from datasets.

In order to show the effectiveness of the proposed models, we design a fuzzy lower approximation-based classifier. Given a set of training samples $DT = \langle U, A, D\rangle$, x is a test sample. We compute the fuzzy lower approximation of each candidate class with different fuzzy rough set models. The decision function is

$$d = \arg_{d_i} \max\{\underline{R^*} d_1(x), \underline{R^*} d_2(x), \ldots, \underline{R^*} d_k(x)\} \tag{31}$$

where $\underline{R^*}$ is a certain fuzzy lower approximation operator.

This algorithm assigns x the class label which achieves the largest fuzzy lower approximation. Suppose $x \in d_j$. We compute $\underline{R^*}d_j(x)$. If x really belongs to d_j, it will be far from the samples in the other classes. Thus, $\underline{R^*}d_j(x)$ should be large; otherwise, $\underline{R^*}d_j(x)$ is small. Therefore, the earlier algorithm is rational to classify x into the class label producing the largest fuzzy lower approximation if no class noise exists.

However, if there are class-noisy samples, the previous algorithm may not work when the classical fuzzy rough set model is employed, while the robust model is still effective in this case. Some numerical experiments are described to test the proposed models in the next section.

The aforementioned classification algorithm has the same time complexity as the nearest neighbor rule, and no training processing is required.

V. Numerical Experiments

In this section, we present some numerical experiments to illustrate the performance of different models. First, we generate a toy classification task to reveal the performance of different fuzzy rough set models in dealing with a noisy task. Then, we present the experimental results on real-world tasks.

A. Toy Example

We first discuss how to measure robustness in statistics theory. Some techniques were developed to measure the robustness of a statistic, such as the sensitivity curve, the influence function, and the breakdown point [44]. We introduce some of them that are used in the following discussion.

Definition 11: Let $\{T_n\}$ be some sequence of statistics; $T_n(x)$ denote the statistic from $\{T_n\}$ on the sample $X = \{x_1, x_2, \ldots, x_n\}$ and $\{T_{n+1}\}(X, x)$ denote the same statistic on the sample $X = \{x_1, x_2, \ldots, x_n, x\}$. Then, the function

$$SC_n(x; T_n, X) = (n+1)[T_{n+1}(x, X) - T_n(X)] \tag{32}$$

characterizes the sensitivity of T_n to the addition of one observation at x and is called the sensitivity curve.

Definition 12: Let d be such a distance; then, the breakdown point ε of the estimator $T_n = T(F_n)$ for the functional $T(F)$ at F is defined as

$$\varepsilon(T, F) = \sup\left\{\varepsilon \le 1 : \sup_{F: d(F, F_0) < \varepsilon} |T(F) - F(F_0) < \text{œ}|\right\}. \tag{33}$$

The influence curve computes the influence of different noise on the statistic T, while the breakdown point gives us how much noise the statistic T can tolerate. The first one shows us how the noise impacts on the statistic.

We generate a binary classification task with 100 samples, as shown in Fig. 2, where "•" stands for the samples from Class 1 and " + " for Class 2. We can see that these two classes of samples are well separable.

Now, we show the memberships of these samples to their decision classes in Fig. 3, where FRS, β-PFRS, FVPRS, k-trimmed, k-mean, and k-median stand for fuzzy lower approximations computed with the classical model, β-precision fuzzy rough model, fuzzy variable precision rough sets, k-trimmed, k-mean, and k-median fuzzy rough models, respectively. In the experiment, we set $k = 3$, $\alpha = 0.1$, and $\beta = 0.02$.

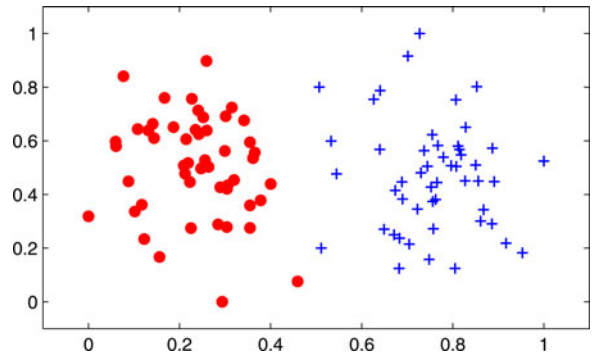


Fig. 2. Artificial dataset, where samples are divided into two classes, and each sample is described with two features.

From Fig. 3, we see that most of the samples produce large memberships. The smallest membership is still larger than 0.3. If we set $\alpha = 0.1$ for FVPRS, since all the memberships are larger than 0.1, FVPRS is equal to FRS.

Now, we consider the dependence functions that are defined in the different models. Dependences of decision on features computed with different models are shown in Fig. 4. FRS and FVPRS obtain the same output because the membership of each sample is larger than α. In this case, FVPRS is equivalent to FRS. β-precision, k-trimmed, k-mean, and k-median do not compute the memberships with the nearest sample from different classes; thus, these models return larger dependence values than FRS and FVPRS as FRS and FVPRS use the nearest sample from different classes. In addition, dependence computed with the k-trimmed model is larger than those with k-mean and k-median models.

We analyze the sensitivity curve of dependences computed with different models. We add one new sample into the dataset and change the location of this sample. We try the location of $(x, 0.5)$, where $x = 0, 0.01, \ldots, 0.99, 1$. Assume this sample comes from Class 2. As we know, the region where $x \le 0.5$ belongs to Class 1. If the new sample is located at this region, it can be considered as a class-noisy sample. Namely, the class of this sample is mislabeled in data gathering. It should belong to Class 1 but mislabeled as Class 2. However, if $x > 0.5$, the sample becomes normal. As a whole, if $x \le 0.5$, the new sample is viewed to be class noisy, while if $x > 0.5$, the sample should be considered as a normal one.

We compute the dependence between decision and features with different models when the location of the sample changes.

Fig. 5 presents the curves of dependence. These curves reflect the performance of different models in dealing with noisy tasks. First, observing FRS and FVPRS, we see that dependences computed with them vary drastically when the location of the new sample changes. The dependence decreases from 0.85 to 0.67 and then rises to 0.85. In most of the cases, FVPRS returns the same results as FRS. When $x \in [0.15, 0.30]$, FVPRS is a little larger than FRS, which shows that some memberships of samples to the lower approximations of their decision classes are less

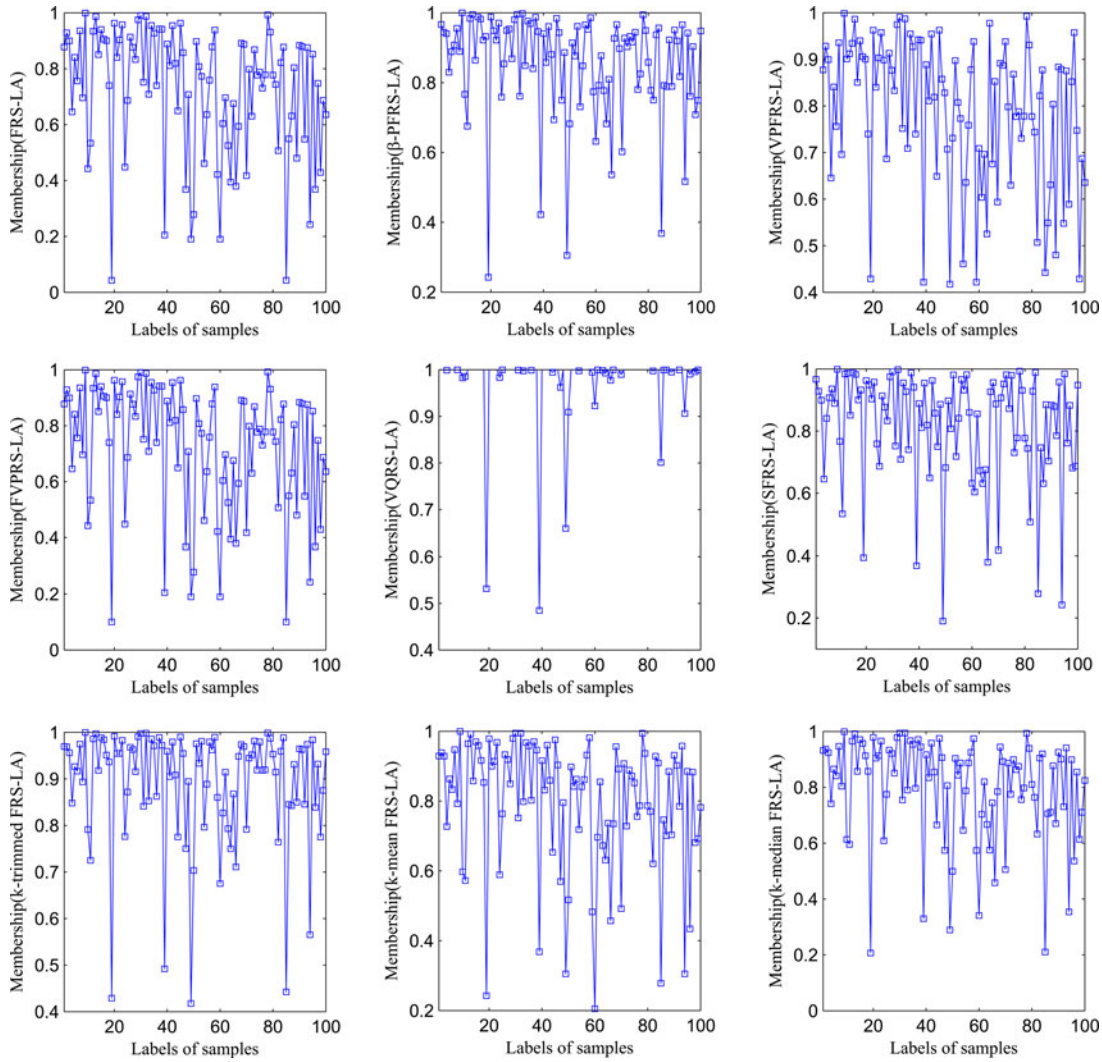


Fig. 3. Membership of samples computed with different models.

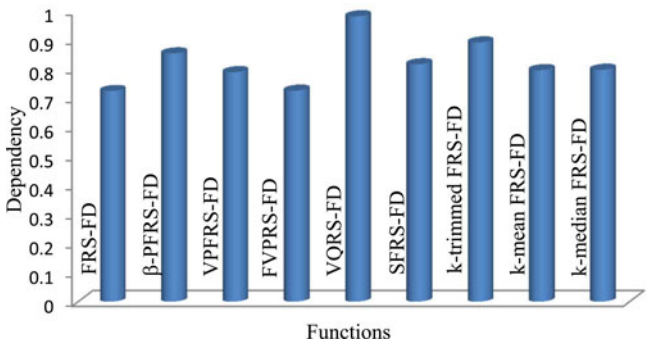


Fig. 4. Dependence of decision on the features.

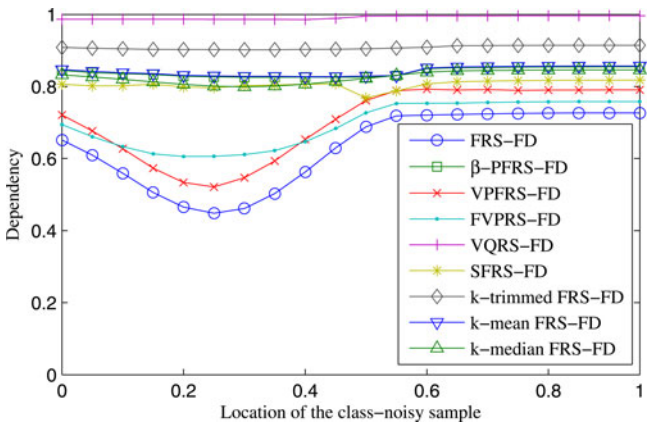


Fig. 5. Sensitivity curve of dependence with the location of noise.

than 0.1 and that FVPRS assigns 0.1. In most cases, the memberships are greater than 0.1. As per β-precision, k-trimmed, k-mean, and k-median models, the noisy sample does not have a large influence on dependence. Relatively speaking, the k-mean model is more sensitive as the statistic mean is sensitive to noise, and the k-median model is more robust than the k-mean model as statistic median is more robust than mean. In addition, k-trimmed and β-precision models are also robust. In fact, we set $k = 3$; the noisy sample has no influence on computing memberships of samples as this noisy sample would be disregarded. The β-PFRS model has the same mechanism as the k-trimmed

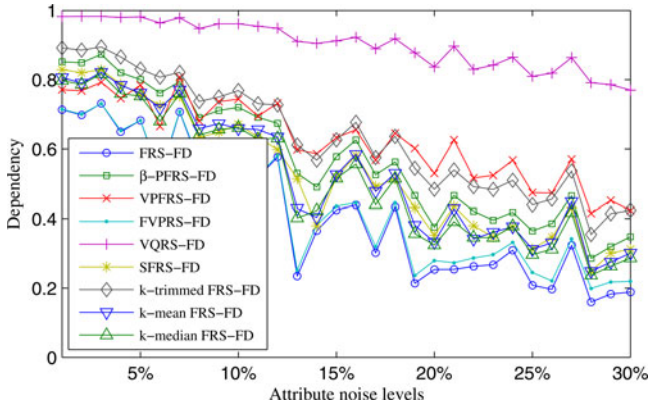


Fig. 6. Dependence variation with the noise level.

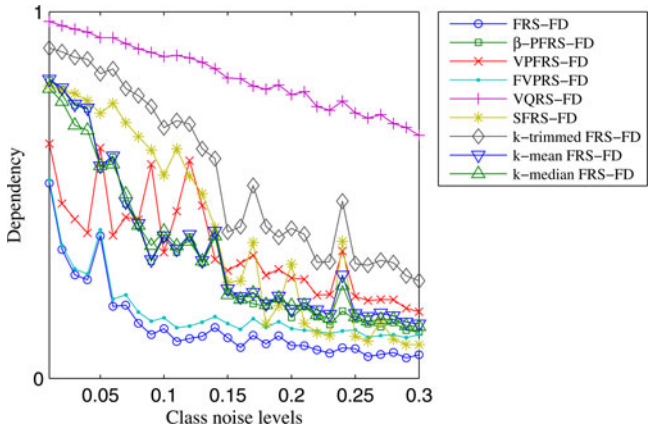


Fig. 7. Dependence variation with level of class noise.

model in dealing with class noise. Both of them disregard some outliers in computing fuzzy lower approximation. Their difference lies in how to set the number of outliers to be disregarded. In the k-trimmed model, users give the value of k in advance, while k is computed from datasets in the β-precision model. From Fig. 5, we can see that both fuzzy rough sets and FVPRS are sensitive to class noise. A single noisy sample might produce great influence on dependence. The sensitiveness would make the algorithms developed with these models not effective in analyzing real-world tasks.

Furthermore, we add some random numbers on each attribute value as attribute noise. The random numbers satisfy the normal distribution with mean zero and standard deviation $0.01 \times i$, where $i = 1, 2, \ldots, 30$. Then, we observe the variation of dependence calculated with different models. Fig. 6 presents the dependence curves. We see that all the dependence values decrease when the deviation of noise increases. However, no abrupt change is observed. If the scale of noise is less than 0.06, no variation can be seen from these curves. The curves show that these fuzzy rough models are not much sensitive to attribute noise as dependence is computed as the average of memberships of samples.

Now, we consider the breakpoint of dependence. We add some class-noisy samples into the original dataset by randomly drawing some samples from the dataset and revising their class

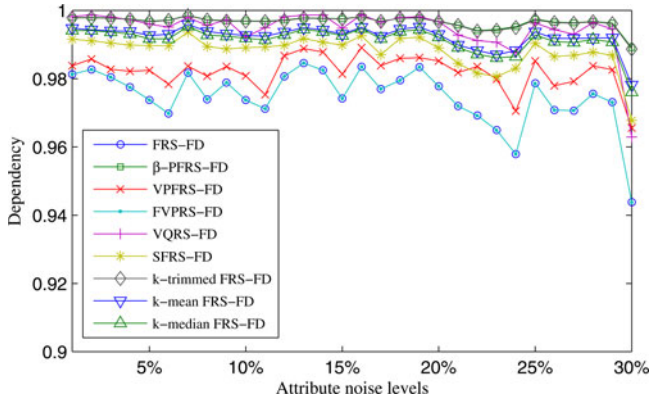


Fig. 8. Dependence variation with level of attribute noise (wine).

labels. The revised samples are considered as class-noisy points. The rate of mislabeled samples takes values from 0.01 to 0.30. Then, we observe the variation of dependences calculated with different models. The dependence curves are given in Fig. 7. Dependences decrease from 0.85 to 0.2 after about ten mislabeled samples are added when FRS and FVPRS are used. At the same time, we also see that other four models are not only robust to class noise, but also able to reflect the level of noise. Dependences nearly linearly decrease with increase of the noise level.

The aforementioned analysis shows that FRS and FVPRS are sensitive to attribute noise and class noise. β-PFRS and the proposed models are much more robust.

B. Experimental Results on Real-World Tasks

Now, we introduce a real-world classification task, named wine [45]. These data are the results of chemical analysis of wines grown in the same region in Italy but derived from three different cultivars. The analysis determined the quantities of 13 constituents found in each of the three types of wines. There are 169 samples described with 13 numerical features and one decision variable.

Fig. 8 gives the dependence curve when the level of attribute noise increases. We see that although attribute noise is added, dependence is still higher than 0.99. We know that the distances between samples in high-dimensional spaces are usually very large; the similarity between samples might be small in this case. Then, each sample produces a large membership to the lower approximation of its decision class. Therefore, the average membership is close to 1.

Then, we add some class noise into the dataset. Here, the class noise is added in the same way as described earlier. We compute dependences with noisy datasets based on different fuzzy rough models. The curves in Fig. 9 describe that dependence varies with the level of class noise. First, we see that dependences that are obtained with FRS and FVPRS are smaller than the other models, and FRS and FVPRS return the same values of dependence. This shows that even though some mislabeled samples are added, FVPRS still does not have effect on the computation of dependence. In order to explain it, we show the memberships of samples to their fuzzy lower approximations in Fig. 10, where 30% mislabeled samples are added. We see that the smallest

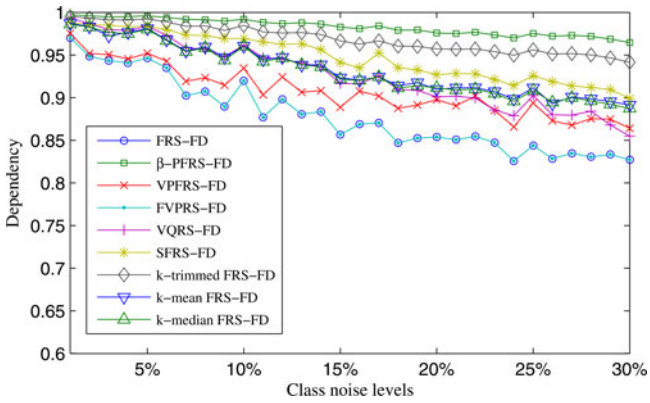


Fig. 9. Dependence variation with level of class noise (wine).

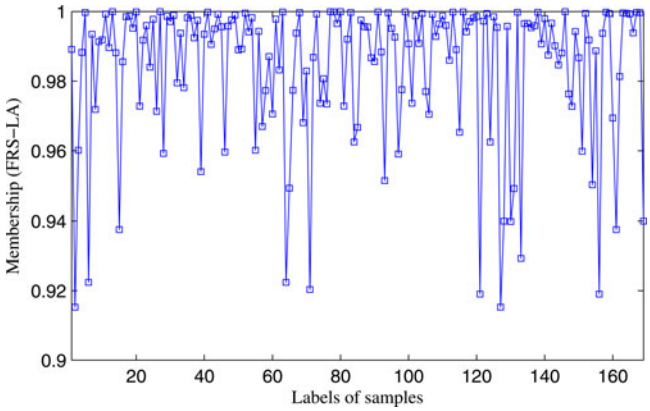


Fig. 10. Membership values of samples to the lower approximations of their decision classes when class noise is added (wine).

membership is still larger than 0.9. If $\alpha = 0.1$ in FVPRS, which is far smaller than 0.9, FVPRS does not work in this case. Therefore, FRS and FVPRS produce the same dependence. In the meanwhile, k-trimmed, k-mean, k-median, and β-precision models are more robust than FRS and FVPRS. k-mean and k-median models nearly get the same dependence values. This experiment shows that β-precision, k-trimmed, k-mean, and k-median are robust in dealing with real-world applications.

Stability is another metric to characterize the robustness of different models [50]. In this technique, we compute dependences between single features and decision at different noise level. As per a robust model, the difference between the dependences computed under different noise level should be small enough; otherwise, we think the model is not robust. Thus, the average correlation of dependences can be calculated as a metric of robustness.

In order to quantitatively characterize the robustness of different models, we compute the average correlation between dependence vectors calculated with the raw dataset and those calculated with the noisy datasets. Both attribute noise and class noise are considered here. The results are shown in Tables I and II, respectively. In the following experiments, we use Gaussian kernel to compute the fuzzy similarity relations between samples and the kernel parameter $\sigma = 0.15$; β-precision FRS: $\beta = 0.9$; VPFRS: $\beta = 0.9$, and FVPRS: $\alpha = [0.6, 0.9]$, In VQRS, parameter u takes values in [0.1,0.2]. As for the robust nearest neighbor models, we try k in $\{3, 5, 7, 9, 11\}$, and finally, the best results are reported.

From the aforementioned results, we see that attribute noise has less influence on dependence than class noise. β-precision and VQFRS are not sensitive to attribute noise and class noise, while FRS and FVPRS are very sensitive to class noise. Compared with β-precision and VQRS, k-trimmed, k-mean, and k-median fuzzy rough sets are competent. In real-world applications, the used techniques should be robust enough. At the same time, they should be able to reflect the noise level as well.

Now, we compute the performance of the fuzzy lower approximation-based classifiers on these classification tasks with tenfold cross validation. The classification accuracies are given in Table III.

Comparing FRS with other algorithms, we see that VQRS, k-trimmed, k-mean, and k-median are better than FRS in most tasks, while beta-precision, VPFRS, and FVPRS are equal to or worse than FRS. As for WPBC, the model of β-PFRS obtains the lowest accuracy among all the models. In addition, VQRS produces the worst performance on thyroid gland task, which leads to the relatively worse average performance than the proposed techniques. The models of k-mean and k-median fuzzy rough sets are more stable than VQRS and k-trimmed models.

Now, we analyze the performance of different models in dealing with noisy data tasks. 5%, 10%, and 15% class noise and attribute noise are added into the raw datasets. Then, we apply the classification algorithms on the noisy datasets and compute the classification accuracies with tenfold cross validation. Tables IV and V present the classification performances of different fuzzy rough set models.

First, we observe the variation of classification performance when class-noise level increases. It is easy to see that the performances of FRS, VPFRS, and FVPRS sharply drop as class-noise level increases. However, the classifiers based on β-precision, VQRS, k-trimmed, k-mean, and k-median models are more robust. The classification accuracies do not change much. Moreover, we can also see that k-trimmed, k-mean, and k-median models usually produce the best performance among the eight models. In addition, the β-precision model also produces good performance on most of the tasks except WPBC. Now, we perform *t*-test on the classification results to compare the proposed models and others. Comparing k-trimmed, k-mean, and k-median models with the classical fuzzy rough sets, the values of $t_{0.99}$ are 2.75, 4.54, and 4.25, respectively. However, β-PFRS and VPFRS just get $t_{0.75} = 0.82$ and $t_{0.5} = 0.07$, respectively. At the same time, there is no significant difference between FVPRS and FRS. Therefore, the proposed models outperform FRS and FVPRS. Among k-trimmed, k-mean, and k-median models, k-mean and k-median models are slightly better than k-trimmed one.

Now, we discuss the attribute noise in Table V. Although the proposed models are designed for class noise, we can see that k-trimmed, k-mean, and k-median models outperform FRS when attribute noise exists. As per k-mean and k-median models, the values of $t_{0.95}$ are 2.02 and 2.06, respectively. On the contrary, the β-precision model is worse than FRS in this case, and VPFRS and FVPRS achieve the almost same results as FRS.

TABLE I
CORRELATION BETWEEN RAW DEPENDENCE AND NOISY DEPENDENCE (CLASS NOISE)

Data	FRS -FD	β-PFRS -FD	VPFRS -FD	FVPRS -FD	VQRS -FD	SFRS -FD	k-trimmed FRS-FD	k-mean FRS-FD	k-median FRS-FD
wine	0.78	0.98	0.94	0.42	0.99	0.93	0.97	0.97	0.95
WDBC	0.70	0.99	0.97	0.39	0.99	0.97	0.98	0.97	0.96
WPBC	0.61	0.96	0.92	0.37	0.95	0.94	0.96	0.96	0.95
ionosphere	0.50	0.92	0.84	0.49	0.98	0.96	0.91	0.92	0.88
diabetes	0.39	0.97	0.97	0.96	0.99	0.96	0.93	0.97	0.95
iris	0.60	0.99	0.98	0.90	0.99	1.00	0.99	0.99	0.98
thyroid-gland	0.31	0.99	0.99	0.44	0.99	0.98	0.99	0.99	0.99

TABLE II
CORRELATION BETWEEN RAW DEPENDENCE AND NOISY DEPENDENCE (ATTRIBUTE NOISE)

Data	FRS -FD	β-PFRS -FD	VPFRS -FD	FVPRS -FD	VQRS -FD	SFRS -FD	k-trimmed FRS-FD	k-mean FRS-FD	k-median FRS-FD
wine	0.93	0.95	0.98	0.98	0.98	0.94	0.94	0.97	0.97
WDBC	0.96	0.98	0.98	0.93	0.99	0.97	0.98	0.98	0.97
WPBC	0.90	0.90	0.91	0.83	0.97	0.96	0.92	0.95	0.93
ionosphere	0.76	0.74	0.74	0.97	0.97	0.95	0.94	0.88	0.85
diabetes	0.78	0.94	0.95	0.59	0.99	0.96	0.91	0.96	0.96
iris	0.95	0.99	0.98	0.82	0.99	0.99	0.99	0.99	0.98
thyroid-gland	0.78	0.97	0.99	0.73	0.98	0.99	0.98	0.98	0.98

TABLE III
CLASSIFICATION ACCURACY (%) COMPARISON ON REAL-WORLD DATASETS

Data	FRS -LA	β-PFRS -LA	VPFRS -LA	FVPRS -LA	VQRS -LA	SFRS -LA	k-trimmed FRS-LA	k-mean FRS-LA	k-median FRS-LA
wine	94.9	95.9	93.8	94.9	95.4	95.5	97.1	97.7	96.0
WDBC	95.4	95.8	94.0	95.4	96.7	96.9	96.3	97.5	96.8
WPBC	69.2	52.4	69.2	69.2	70.7	76.8	71.8	74.3	73.2
ionosphere	71.5	68.4	71.2	71.5	71.9	86.4	70.4	71.5	71.5
diabetes	70.8	71.6	74.1	70.4	74.9	75.4	75.0	75.8	74.2
iris	95.3	96.7	95.3	94.0	96.0	96.7	95.3	95.3	96.0
thyroid-gland	95.3	94.9	90.2	93.0	88.4	95.3	94.4	95.3	95.3

TABLE IV
CLASSIFICATION ACCURACY (%) COMPARISON ON CLASS-NOISY DATASETS

Data	Noise levels	FRS -LA	β-PFRS -LA	VPFRS -LA	FVPRS -LA	VQRS -LA	SFRS -LA	k-trimmed FRS-LA	k-mean FRS-LA	k-median FRS-LA
wine	5%	90.0	96.7	85.6	90.0	94.3	94.8	97.6	97.3	96.2
	10%	85.2	95.4	79.3	85.2	91.9	94.3	97.2	96.7	95.8
	15%	81.1	95.3	72.7	81.1	90.3	93.1	97.0	96.4	95.2
WDBC	5%	91.0	95.9	87.1	91.4	96.3	96.0	95.9	97.0	96.6
	10%	86.7	95.4	84.2	86.7	95.3	96.1	95.9	96.7	96.6
	15%	82.5	95.3	81.7	82.5	93.8	95.3	95.3	96.2	95.7
WPBC	5%	67.8	50.0	67.6	67.6	69.6	72.6	70.2	73.9	72.5
	10%	66.5	47.9	66.5	66.5	69.0	71.1	71.0	73.4	71.3
	15%	63.2	48.5	63.2	63.2	64.0	68.8	70.3	70.9	68.8
ionosphere	5%	68.5	67.7	67.0	68.5	70.5	72.5	65.6	70.3	70.4
	10%	65.5	67.7	63.8	65.6	69.8	73.5	66.4	70.3	70.3
	15%	62.5	66.7	61.2	62.7	68.2	72.0	65.9	70.1	69.7
diabetes	5%	68.5	72.0	72.8	67.9	74.8	73.6	74.3	74.9	74.0
	10%	66.8	71.9	72.8	66.3	73.9	72.5	73.5	73.9	73.0
	15%	64.7	72.3	71.2	64.4	73.3	72.4	73.9	73.0	72.5
iris	5%	92.5	96.4	92.1	89.5	96.1	96.0	95.6	95.3	95.5
	10%	86.2	96.4	90.9	83.2	95.9	96.0	95.9	95.7	96.0
	15%	82.5	96.0	90.3	78.1	95.6	96.0	95.6	96.4	96.1
thyroid-gland	5%	91.2	94.4	80.2	84.8	88.2	93.1	89.3	92.5	92.5
	10%	87.1	94.4	92.6	81.6	88.1	93.1	89.2	93.5	92.8
	15%	82.0	92.0	93.1	91.2	80.0	91.0	87.4	92.1	92.5
Aver.		77.7	81.3	77.9	77.5	82.8	84.9	84.0	85.5	85.0

TABLE V
CLASSIFICATION ACCURACY (%) COMPARISON ON ATTRIBUTE-NOISY DATASETS

Data	Noise levels	FRS -LA	β-PFRS -LA	VPFRS -LA	FVPRS -LA	VQRS -LA	SFRS -LA	k-trimmed FRS-LA	k-mean FRS-LA	k-median FRS-LA
wine	5%	95.8	95.1	94.4	95.8	96.6	95.8	96.0	96.4	96.7
	10%	94.2	95.5	91.9	94.2	95.4	95.7	95.9	95.7	95.6
	15%	92.5	93.4	90.4	92.5	94.6	94.0	93.4	94.9	93.7
WDBC	5%	92.1	92.2	89.8	92.1	91.7	92.1	91.9	92.3	92.9
	10%	84.2	83.6	84.2	84.2	83.6	83.6	84.5	84.5	84.0
	15%	77.2	70.1	77.2	77.3	78.6	79.9	79.5	79.6	79.9
WPBC	5%	66.0	54.4	66.0	66.0	68.3	73.2	74.2	72.3	71.2
	10%	65.2	58.5	65.2	68.1	72.4	72.8	72.6	72.6	72.7
	15%	65.9	52.4	65.9	65.9	66.8	71.5	70.7	71.7	71.0
ionosphere	5%	69.6	69.1	69.1	69.6	70.0	70.0	70.0	70.0	70.0
	10%	69.3	68.5	69.3	69.3	69.6	69.4	68.3	69.4	69.4
	15%	68.6	67.7	68.6	68.6	68.6	66.5	67.4	68.3	68.3
diabetes	5%	69.5	73.7	71.8	69.5	72.6	73.2	73.0	71.8	71.8
	10%	67.9	72.0	67.8	78.0	70.2	72.3	71.0	70.2	70.2
	15%	66.1	71.5	63.5	66.1	68.9	71.1	70.1	67.7	67.7
iris	5%	95.5	95.5	95.9	92.8	95.7	95.3	95.6	95.5	95.6
	10%	91.3	94.8	92.1	89.1	94.7	95.3	94.8	95.1	94.2
	15%	88.0	92.7	88.8	86.7	93.3	94.7	93.3	92.5	91.2
thyroid-gland	5%	93.1	94.5	92.8	93.0	86.3	93.2	90.9	93.0	93.0
	10%	87.4	90.2	88.2	83.3	83.3	87.5	86.3	88.1	88.6
	15%	81.1	82.4	80.2	81.1	80.9	82.0	81.2	83.1	83.1
Aver.		80.0	75.6	79.7	80.1	81.1	82.3	82.0	82.1	81.9

VI. CONCLUSION

Fuzzy rough set theory has attracted much attention in recent years. Noise is one of the main sources of uncertainty in applications. It has been shown that most of the current fuzzy rough operators are not robust to noise. In this paper, we systematically discuss why the models of rough sets are sensitive to noise and develop some robust models of fuzzy rough sets. Some numerical experiments are described. The following conclusions can be drawn from the analysis.

1) The classical fuzzy rough operators are sensitive to mislabel samples instead of small perturbation in attribute values. Mislabeled samples have great impact on fuzzy dependence functions, while the influence of attribute noise is restrained.

2) The classical fuzzy rough set model computes the membership of a sample to the lower approximation of its decision based on the nearest sample from different classes. Here, the statistic of minimum is used. It is sensitive to outliers. This is the essential reason why the classical model is not robust to class-noisy samples.

3) We discuss the disadvantage of VPRS and FVPRS. VPRS cannot reflect the noise level in dealing with noisy tasks as they group the mislabeled samples into classification positive region, while FVPRS does not work in dealing with noise unless the membership of a sample is less than the parameter α in it. We find that there is contradiction in setting the parameter value.

4) β-PFRS, soft fuzzy rough sets, k-trimmed, k-mean, and k-median fuzzy rough set models are not only able to reduce the influence of class noise but can also reflect the level of noise. Therefore, these models are effective in dealing with noisy tasks. The corresponding classifiers are better than those developed with FRS, VPFRS, FVPRS, and VQRS. Comparing with β-precision FRS and soft FRS, we find that k-mean and k-median fuzzy rough set models are usually more effective. Furthermore, the semantics of the parameters used in k-mean and k-median models are clear, which is important for applicability.

REFERENCES

[1] Z. Pawlak, *Rough Sets-Theoretical Aspects of Reasoning About Data*. Dordrecht, The Netherlands: Kluwer, 1991.

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

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

[4] D. S. Yeung, D. G. Chen, E. C. C. Tsang, J. W. T. Lee, and X. Z Wang, "On the generalization of fuzzy rough sets," *IEEE Trans. Fuzzy Syst.*, vol. 13, no. 3, pp. 343–361, Jun. 2005.

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

[6] D. C. Martine, C. Cornelis, and E. E. Kerre, "Fuzzy rough sets: The forgotten step," *IEEE Trans. Fuzzy Syst.*, vol. 15, no. 1, pp. 121–130, Feb. 2007.

[7] E. C. C. Tsang, D. G. Chen, D. S. Yeung, X.-Z. Wang, and J. Lee, "Attributes reduction using fuzzy rough sets," *IEEE Trans. Fuzzy Syst.*, vol. 16, no. 5, pp. 1130–1141, Oct. 2008.

[8] H. Y. Wu, Y. Y. Wu, and J. P. Luo, "An interval type-2 fuzzy rough set model for attribute reduction," *IEEE Trans. Fuzzy Syst.*, vol. 17, no. 2, pp. 301–315, Apr. 2009.

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

[10] X. Liu, W. Pedrycz, and M. Song, "The development of fuzzy rough sets with the use of structures and algebras of axiomatic fuzzy sets," *IEEE Trans. Knowl. Data Eng.*, vol. 21, no. 3, pp. 443–462, Mar. 2009.

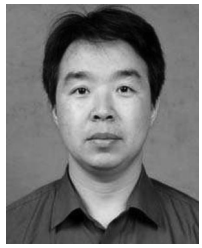
[11] L. A. Zadeh, "Fuzzy logic = Computing with words," *IEEE Trans. Fuzzy Syst.*, vol. 4, no. 2, pp. 103–111, May 1996.

[12] L. Zadeh, "Toward a theory of fuzzy information granulation and its centrality in human reasoning and fuzzy logic," *Fuzzy Sets Syst.*, vol. 90, pp. 111–127, 1997.

[13] P. Maji and S. K. Pal, "Rough-fuzzy C-medoids algorithm and selection of bio-basis for amino acid sequence analysis," *IEEE Trans. Knowl. Data Eng.*, vol. 19, no. 6, pp. 859–872, Jun. 2007.

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

[15] Q. H. Hu, L. Zhang, D. G. Chen, W. Pedrycz, and D. Yu, "Gaussian kernel based fuzzy rough sets: Model, uncertainty and applications," *Int. J. Approx. Reason.*, vol. 51, no. 4, pp. 453–471, 2010.
[16] T. P. Hong, T. T. Wang, S. L. Wang, and B. C. Chien, "Learning a coverage set of maximally general fuzzy rules by rough sets," *Expert Syst. Appl.*, vol. 19, no. 2, pp. 97–103, 2000.
[17] X. Z. Wang, E. C. C. Tsang, S. Y. Zhao, D. Chen, and D. S. Yeung, "Learning fuzzy rules from fuzzy samples based on rough set technique," *Inf. Sci.*, vol. 177, no. 20, pp. 4493–4514, 2007.
[18] S. Y. Zhao, E. C. C. Tsang, D. G. Chen, and X. Z. Wang, "Building a rule-based classifier—A fuzzy-rough set approach," *IEEE Trans. Knowl. Data Eng.*, vol. 22, no. 5, pp. 624–638, May 2010.
[19] R. B. Bhatt and M. Gopal, "FRCT: Fuzzy-rough classification trees," *Pattern Anal. Appl.*, vol. 11, no. 1, pp. 73–88, 2008.
[20] A. Globerson and S. Roweis, "Nightmare at test time: Robust learning by feature deletion," in *Proc. 23rd Int. Conf. Mach. Learning*, 2006, pp. 353–360.
[21] S. Agarwal, S. Godbole, and D. Punjani, "How much noise is too much: A study in automatic text classification," in *Proc. 7th IEEE Int. Conf. Data Mining*, 2007, pp. 3–12.
[22] X. D. Wu and X. Q. Zhu, "Mining with noise knowledge: Error-aware data mining," *IEEE Trans. Syst., Man, Cybern. A: Syst. Humans*, vol. 38, no. 4, pp. 917–932, Jul. 2008.
[23] D. Gamberger, N. Lavrac, and S. Dzeroski, "Noise detection and elimination in data reprocessing: Experiments in medical domains," *Appl. Artif. Intell.*, vol. 14, pp. 205–223, 2000.
[24] Y. Chen, X. Dang, H. Peng, and H. L. Bart Jr., "Outlier detection with the kernelized spatial depth function," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 2, pp. 288–305, Feb. 2009.
[25] M. Kearns, "Efficient noise-tolerant learning from statistical queries," *J. Assoc. Comput. Mach.*, vol. 45, no. 6, pp. 983–1006, 1998.
[26] H. Xu, C. Caramanis, and S. Mannor, "Robustness and regularization of support vector machines," *J. Mach. Learn. Res.*, vol. 10, pp. 1485–1510, 2009.
[27] Y. Y. Yao, S. K. M. Wong, and P. Lingras, "A decision-theoretic rough set model," in *Methodologies for Intelligent Systems*, vol. 5, Z. W. Ras, M. Zemankova, and M. L. Emrich, Eds. New York: North-Holland, 1990, pp. 17–24.
[28] W. Ziarko, "Variable precision rough set model," *J. Comput. Syst. Sci.*, vol. 46, pp. 39–59, 1993.
[29] M. Beynon, "Reducts within the variable precision rough sets model: A further investigation," *Eur. J. Oper. Res.*, vol. 134, no. 3, pp. 592–605, 2001.
[30] J.-S. Mi, W.-Z Wu, and W.-X. Zhang, "Approaches to knowledge reduction based on variable precision rough set model," *Inf. Sci*, vol. 159, no. 3–4, pp. 255–272, 2004.
[31] M. Ningler, G. Stockmanns, G. Schneider, H. D. Kochs, and E. Kochs, "Adapted variable precision rough set approach for EEG analysis," *Artif. Intell. Med.*, vol. 47, no. 3, pp. 239–261, 2009.
[32] J. M. F. Salido and S. Murakami, "Rough set analysis of a general type of fuzzy data using transitive aggregations of fuzzy similarity relations," *Fuzzy Sets Syst.*, vol. 139, pp. 635–660, 2003.
[33] A. Mieszkowicz-Rolka and L. Rolka, *Variable Precision Fuzzy Rough Sets. Transactions on Rough Sets I*. vol. LNCS-3100, Berlin, Germany: Springer, 2004, pp. 144–160.
[34] C. Cornelis, M. De Cock, and A. M. Radzikowska, "Vaguely quantified rough sets," in *Proc. 11th Int. Conf. Rough Sets, Fuzzy Sets, Data Mining, Granular Comput.*, 2007, pp. 87–94.
[35] C. Cornelis and R. Jensen, "A noise-tolerant approach to fuzzy-rough feature selection," in *Proc. IEEE Int. Conf. Fuzzy Syst.*, 2008, pp. 1598–1605.
[36] X. Q. Zhu and X. D. Wu, "Class noise versus attribute noise: A quantitative study of their impacts," *Artif. Intell. Rev.*, vol. 22, pp. 177–210, 2004.
[37] D. F. Nettleton, A. Orriols-Puig, and A. Fornells, "A study of the effect of different types of noise on the precision of supervised learning techniques," *Artif. Intell. Rev.*, vol. 33, pp. 275–306, 2010.
[38] J. S. Mi and W. X. Zhang, "An axiomatic characterization of a fuzzy generalization of rough sets," *Inf. Sci.*, vol. 160, pp. 235–249, 2004.
[39] W.-Z. Wu and W. Zhang, "Constructive and axiomatic approaches of fuzzy approximation operators," *Inf. Sci.*, vol. 159, pp. 233–254, 2004.
[40] Q. H. Hu, D. R. Yu, W. Pedrycz, and D. G. Chen, "Kernelized fuzzy rough sets and their applications," *IEEE Trans. Knowl. Data Eng.*, vol. 23, no. 11, pp. 1649–1667, Nov. 2011.
[41] M. Robnik-Sikonja and I. Kononenko, "Theoretical and empirical analysis of reliefF and rReliefF," *Mach. Learn.*, vol. 53, pp. 23–69, 2003.
[42] Y. Y. Yao and Y. Zhao, "Attribute reduction in decision-theoretic rough set models," *Inf. Sci.*, vol. 178, no. 17, pp. 3356–3373, 2008.
[43] D. C. Hoaglin, F. Mosteller, and J. W. Tukey, *Understanding Robust and Exploratory Data Analysis*. New York: Wiley, 1983.
[44] J. W. Tukey, *Exploratory Data Analysis*. Reading, MA: Addison-Wesley, 1977.
[45] A. Asuncion and D. J. Newman (2007). UCI machine learning repository, School Inf. Comput. Sci., Univ. California, Irvine, [Online]. Available: http://www.ics.uci.edu/mlearn/ MLRepository.html
[46] L. Zhou, W. -Z. Wu, and W. -X. Zhang, "On characterization of intuitionistic fuzzy rough sets based on intuitionistic fuzzy implicators," *Inf. Sci.*, vol. 179, no. 7, pp. 883–898, 2009.
[47] S. Y. Zhao, E. C. C. Tsang, and D. G. Chen, "The model of fuzzy variable precision rough sets," *IEEE Trans. Fuzzy Syst.*, vol. 17, no. 2, pp. 451–467, Apr. 2009.
[48] Y. H. Qian, J. Y. Liang, W. Z. Wu, and C. Y. Dang, "Information granularity in fuzzy binary GrC model," *IEEE Trans. Fuzzy Syst.*, vol. 19, no. 2, pp. 253–264, Apr. 2011.
[49] J. Lawry and Y. C. Tang, "Granular knowledge representation and inference using labels and label expressions," *IEEE Trans. Fuzzy Syst.*, vol. 18, no. 3, pp. 500–514, Jun. 2010.
[50] A. Kalousis, J. Prados, and M. Hilario, "Stability of feature selection algorithms: A study on high-dimensional spaces," *Knowl. Inf. Syst.*, vol. 12, no. 1, pp. 95–116, 2007.
[51] X.-Z Wang and C.-R. Dong, "Improving generalization of fuzzy if-then rules by maximizing fuzzy entropy," *IEEE Trans. Fuzzy Syst.*, vol. 17, no. 3, pp. 556–567, Jun. 2009.
[52] X.-Z. Wang, J.-H. Zhai, and S.-X. Lu, "Induction of multiple fuzzy decision trees based on rough set technique," *Inf. Sci.*, vol. 178, pp. 3188–3202, 2008.
[53] Q. H. Hu, S. An, and D. R. Yu, "Soft fuzzy rough sets for robust feature evaluation and selection," *Inf. Sci.*, vol. 180, pp. 4384–4400, 2010.
[54] C. Cornelis, N. Verbiest, and R. Jensen, "Ordered weighted average based fuzzy rough sets," in *Proc. 5th Int. Conf. Rough Sets Knowl. Technol.*, 2010, pp. 78–85.
[55] W. Zhu and S. P. Wang, "Matroidal approaches to generalized rough sets based on relations," *Int. J. Mach. Learn. Cybern.*, vol. 2, no. 4, pp. 273–279, 2011.
[56] J.-H. Zhai, "Fuzzy decision tree based on fuzzy-rough technique," *Soft Comput.*, vol. 15, no. 6, pp. 1087–1096, 2011.
[57] Z. H. Deng, F.-L. Chung, and S. T. Wang, "Robust relief-feature weighting, margin maximization, and fuzzy optimization," *IEEE Trans. Fuzzy Syst.*, vol. 18, no. 4, pp. 726–744, Aug. 2010.
[58] Y. Liu, Y. L. Jiang, and L. C. Huang, "Modeling complex architectures based on granular computing on ontology," *IEEE Trans. Fuzzy Syst.*, vol. 18, no. 3, pp. 585–598, Jun. 2010.
[59] W. Pedrycz, V. Loia, and S. Senatore, "Fuzzy clustering with viewpoints," *IEEE Trans. Fuzzy Syst.*, vol. 18, no. 2, pp. 274–284, Apr. 2010.
[60] V. Vagin and M. Fomina, "Problem of knowledge discovery in noisy databases," *Int. J. Mach. Learn. Cybern.*, vol. 2, no. 3, pp. 135–145, 2011.



Qinghua Hu (M'10) received the B.S., M.S., and Ph.D. degrees from the Harbin Institute of Technology, Harbin, China, in 1999, 2002, and 2008, respectively.

He was an Associate Professor with the Harbin Institute of Technology from 2008 to 2011. He is currently a Full Professor with the School of Computer Science and Technology, Tianjin University, Tianjin, China. His research interests include intelligent modeling, data mining, and knowledge discovery for classification and regression. He chaired 2010 International Conference on Rough Sets and Current Trends in Computing Publications Committee and severs as referee for many journals and conferences. He has published more than 70 journal and conference papers in the areas of pattern recognition and fault diagnosis.

Lei Zhang (M'04) received the B.S. degree from the Shenyang Institute of Aeronautical Engineering, Shenyang, China, in 1995 and the M.S. and Ph.D. degrees in electrical and engineering from Northwestern Polytechnical University, Xi'an, China, in 1998 and 2001, respectively.

He was a Research Associate with the Department of Computing, The Hong Kong Polytechnic University, Kowloon, Hong Kong, from 2001 to 2002. From January 2003 to January 2006, he was a Postdoctoral Fellow with the Department of Electrical and Computer Engineering, McMaster University, Hamilton, ON, Canada. Since January 2006, he has been an Assistant Professor with the Department of Computing, The Hong Kong Polytechnic University. His research interests include image and video processing, biometrics, pattern recognition, multisensor data fusion, machine learning, and optimal estimation theory.

Shuang An received the B.S. degree from Shenyang Normal University, Shenyang, China, in 2005, the M.S. degree from Northeastern University, Shenyang, in 2008, and the Ph.D. degree from Harbin Institute of Technology, Harbin, China, in 2011.

She is currently with Northeastern University at Qinhuangdao, Qinhuangdao, China. Her interests include robust machine learning, and pattern recognition.

David Zhang (F'08) received the B.Sc. degree in computer science from Peking University, Beijing, China. He received the M.Sc. degree in computer science in 1982 and the Ph.D. degree in 1985 from the Harbin Institute of Technology, Harbin, China. He also received the Ph.D. degree in electrical and computer engineering from the University of Waterloo, Waterloo, ON, Canada, in 1994.

From 1986 to 1988, he was a Postdoctoral Fellow with Tsinghua University, Beijing, and then an Associate Professor with the Academia Sinica, Beijing. He is currently a Head of the Department of Computing and a Chair Professor with the Hong Kong Polytechnic University, where he is also the Founding Director of the Biometrics Technology Centre, which is supported by the Hong Kong SAR Government in 1998. He also serves as a Visiting Chair Professor with Tsinghua University and Adjunct Professor with Shanghai Jiao Tong University, Shanghai, China; Peking University; Harbin Institute of Technology; and the University of Waterloo. He is the author of more than ten books and 200 journal papers. He is the Editor of the *Springer International Series on Biometrics* (Berlin, Germany: Springer).

Dr. Zhang is the Founder and Editor-in-Chief of the *International Journal of Image and Graphics* and the Organizer of the first International Conference on Biometrics Authentication. He is an Associate Editor of more than ten international journals, including the IEEE TRANSACTIONS AND PATTERN RECOGNITION. He is the Technical Committee Chair of the IEEE Computational Intelligence Society. He is a Croucher Senior Research Fellow, a Distinguished Speaker of the IEEE Computer Society, and a Fellow of the International Association for Pattern Recognition.

Daren Yu received the M.E. and Ph.D. degrees from the Harbin Institute of Technology, Harbin, China, in 1988 and 1996, respectively.

He has been with the School of Energy Science and Engineering, Harbin Institute of Technology, since 1988. He has published more than 100 conference and journal papers on power control and fault diagnosis. His main research interests include modeling, simulation, and the control of power systems.