Available online at www.sciencedirect.com
ScienceDirect
Fuzzy Sets and Systems 183 (2011) 26–43
FUZZY sets and systems
www.elsevier.com/locate/fss

Robust fuzzy rough classifiers

Qinghua Hu[a,b,*], Shuang An[a], Xiao Yu[a], Daren Yu[a]

[a] *Harbin Institute of Technology, Harbin 150001, China*
[b] *Department of Computing, The Hong Kong Polytechnic University, Hong Kong, China*

Received 15 December 2009; received in revised form 15 January 2011; accepted 27 January 2011
Available online 5 February 2011

Abstract

Fuzzy rough sets, generalized from Pawlak's rough sets, were introduced for dealing with continuous or fuzzy data. This model has been widely discussed and applied these years. It is shown that the model of fuzzy rough sets is sensitive to noisy samples, especially sensitive to mislabeled samples. As data are usually contaminated with noise in practice, a robust model is desirable. We introduce a new model of fuzzy rough set model, called soft fuzzy rough sets, and design a robust classification algorithm based on the model. Experimental results show the effectiveness of the proposed algorithm.

Keywords: Approximate reasoning; Decision analysis; Fuzzy statistics and data analysis; Fuzzy rough sets; Robustness

1. Introduction

The rough set theory, proposed by Pawlak in 1982 [30], provides a mathematical tool to handle inconsistency in concept description and classification analysis [31,32]. This theory has been successfully applied in feature selection, attribute reduction and rule learning [28,33,34,43,44,62]. In addition, this theory was extended to the fuzzy case [10] and the new models have been used in various domains [3,17,18,39,42,58].

Unfortunately, it was reported that the fuzzy rough set model was sensitive to noisy samples when it was used in real-world tasks [16,59]. As we know there are two key definitions in the model: fuzzy lower and upper approximations. The membership of a sample to the fuzzy lower approximation depends on its nearest sample from other classes, and the membership to the fuzzy upper approximation of a set is computed with the nearest sample in that set. The essence of different fuzzy rough set models is the same though different fuzzy approximation operators were developed [27,52,53,56]. Data usually contain noisy samples in practice. This leads to performance degeneration of fuzzy rough set models as these models utilize the nearest neighbors in computing lower and upper approximations. If the nearest neighbor is a mislabeled sample, the values of fuzzy lower and upper approximations may be completely contaminated. We will further discuss this issue in the next section.

Some robust models of fuzzy rough sets were introduced these years. In [7], the model of vaguely quantified rough sets (VQRS) was introduced. This model is robust to noisy samples by disregarding some samples producing small

* Corresponding author at: Harbin Institute of Technology, Harbin 150001, China.
E-mail addresses: huqinghua@hit.edu.cn (Q. Hu), yudaren@hit.edu.cn (D.R.Yu).

0165-0114/$ - see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.fss.2011.01.016

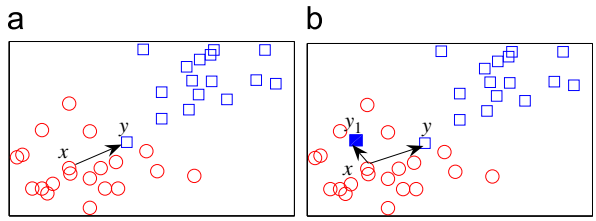

Fig. 1. $\underline{R}D_i(x)$.

memberships. Rolka and Zhao et al. introduced two robust models of rough sets, called variable precision fuzzy rough sets (VPFRS) [26] and fuzzy variable precision rough sets (FVPRS) [59], respectively. There are also some disadvantages with these models. The detailed discussion will be given in Section 4.

In the domains of machine learning and pattern recognition, there are two main approaches to reducing the influence of noise. One is to remove noisy samples before learning, such as outlier detection [1,6,11,22,36,41,45], data cleaner [55] and impact-sensitive ranking [61]. And the other one is to design robust algorithms for noisy data, such as weighted k-nearest neighbor method [46], Maxi-Min Margin Machine (M^4) [19], robust minimax approach [23], nearest subclass classifier [51], cost-sensitive classification [60], error-aware classification [54], noise tolerant classification [12], fuzzy rough nearest neighbor classification [20], a robust feature selection method [21] and soft margin SVM [4,47,50].

Generally speaking, it is difficult to know which samples are corrupted by noise. Thus, it seems more feasible to design noise-tolerant learning algorithms than noise removal. Soft margin SVM is a popular classification model to deal with noisy data [50]. In hard margin SVM, it is assumed that all the samples can be correctly classified with a margin [50]. However, noisy samples may lead to misclassification. So soft margin SVM allows several samples misclassified in training for obtaining a large-margin classifier. By this way, soft margin SVM reduces the influence of noise on final classification functions.

We know that the fuzzy rough set model is sensitive to mislabeled samples because it is constructed on two sensitive statistics *inf* and *sup*. Given a nonempty universe U, for a decision class $D_i \subset U$, $\underline{R}D_i(x) = \inf_{y\in U-D_i}\{1 - R(x, y)\}$. That is to say $\underline{R}D_i(x)$ is determined by the nearest sample y in $U - D_i$ (Fig. 1(a)).

If the dataset is corrupted with a mislabeled sample, such as y_1 in Fig. 1(b), $\underline{R}D_i(x)$ will be significantly influenced. The memberships of the samples belonging to the same class as x will all change. If we introduce the idea of soft margin SVM into fuzzy rough sets i.e. overlooking the sample y_1 in computing $\underline{R}D_i(x)$, the memberships of the samples in D_i do not vary.

In this work, we discuss the following issues. First, we simulate the idea of soft margin SVM and introduce a robust rough set model. As we know, outliers usually occur with small probabilities. In order to reduce the impact of outliers, we can overlook some samples in computing lower and upper approximations. That is to say, we do not compute the membership of a sample based on the nearest sample with different class labels, but the kth nearest sample, where k is adaptively determined by a tradeoff between k and the increase of memberships. In this case, the $k - 1$ nearest samples are discarded in computing memberships. Secondly, we discuss the similarities and differences between the proposed model and existing models, including VQRS, VPFRS and FVPRS. Then, we design a robust classifier based on the soft fuzzy lower approximation, called soft fuzzy rough classifier. Some numerical experiments are shown to test the proposed technique.

The paper is organized as follows. Section 2 introduces the preliminaries on fuzzy rough sets and discusses the robustness of the fuzzy approximation operators. Section 3 presents the soft fuzzy rough set model. And then we compare the proposed model with VQRS, VPFRS and FVPRS in Section 4. Next, we design the soft fuzzy rough classifier in Section 5 and experimental results are shown in Section 6. Finally, conclusions are given in Section 7.

2. Preliminaries on fuzzy rough sets

Given a nonempty universe U, R is a fuzzy binary relation on U. If R satisfies (1) reflexivity: $R(x, x) = 1$; (2) symmetry: $R(x, y) = R(y, x)$; (3) sup-min transitivity: $R(x, y) \geq \text{supmin}_{z\in U}\{R(x, z), R(z, y)\}$, we say R is a fuzzy equivalence relation. The fuzzy equivalence class $[x]_R$ associated with x and R is a fuzzy set on U, where $[x]_R(y) = R(x, y)$ for all $y \in U$. Based on fuzzy equivalence relations fuzzy rough sets were first introduced in [10].

Definition 1. Let U be a nonempty universe, R be a fuzzy equivalence relation on U and $F(U)$ be the fuzzy power set of U. Given a fuzzy set $F \in F(U)$, the lower and upper approximations are defined as

$$\begin{cases} \underline{R}F(x) = \inf\limits_{y\in U} \max\{1 - R(x, y), F(y)\}, \\ \overline{R}F(x) = \sup\limits_{y\in U} \min\{R(x, y), F(y)\}. \end{cases} \tag{1}$$

These approximation operators were discussed in the viewpoint of the constructive and axiomatic approaches in [52,53]. In 1998, Morsi and Yakout replaced fuzzy equivalence relation with a T-equivalence relation and built an axiom system of the model [27], where the lower and upper approximations of $F \in F(U)$ are

$$\begin{cases} \underline{R_\theta}F(x) = \inf\limits_{y\in U} \theta(R(x, y), F(y)), \\ \overline{R_T}F(x) = \sup\limits_{y\in U} T(R(x, y), F(y)), \end{cases} \tag{2}$$

where T is a triangular norm.

In 2002, based on σ and θ Radzikowska and Kerre introduced another model [35]:

$$\begin{cases} \underline{R_\theta}F(x) = \inf\limits_{y\in U} \theta(R(x, y), F(y)), \\ \overline{R_\sigma}F(x) = \sup\limits_{y\in U} \sigma(N(R(x, y)), F(y)). \end{cases} \tag{3}$$

In classification learning, samples are assigned with a class label and described with a group of features. Fuzzy equivalence relations can be generated with numerical or fuzzy features, while the decision variable divides the samples into some subsets. In this case, the task is to approximate these decision classes with the fuzzy equivalence classes induced with the features. Given a decision system $\langle U, R, D\rangle$, for a decision class $D_i \in U/D$, the membership of a sample x to D_i is

$$D_i(x) = \begin{cases} 1 & x \in D_i, \\ 0 & x \notin D_i. \end{cases} \tag{4}$$

Then the membership of sample x to the fuzzy lower approximation of D_i is

$$\begin{aligned} \underline{R}D_i(x) &= \inf_{y\in U} \max\{1 - R(x, y), D_i(y)\} = \inf_{y\in D_i} \max\{1 - R(x, y), 1\} \wedge \inf_{y\notin D_i} \max\{1 - R(x, y), 0\} \\ &= 1 \wedge \inf_{y\notin D_i} \{1 - R(x, y)\} = \inf_{y\notin D_i} \{1 - R(x, y)\}. \end{aligned} \tag{5}$$

If we introduce Gaussian function

$$G(x, y) = \exp\left(-\frac{\|x - y\|^2}{2\sigma^2}\right) \tag{6}$$

to compute the similarity R, $1 - G(x, y)$ can be considered as a pseudo-distance function.

Similarly, the membership of sample x to the fuzzy upper approximation of D_i is

$$\begin{aligned} \overline{R}D_i(x) &= \sup_{y\in U} \min\{R(x, y), D_i(y)\} = \sup_{y\in D_i} \min\{R(x, y), 1\} \vee \sup_{y\notin D_i} \min\{R(x, y), 0)\} \\ &= \sup_{y\in D_i} \{R(x, y)\} \vee 0 = \sup_{y\in D_i} \{R(x, y)\}. \end{aligned} \tag{7}$$

We can see that $\underline{R}D_i(x)$ is the distance from x to its nearest sample from different classes (Fig. 1(a)); while $\overline{R}D_i(x)$ is the similarity between x and the nearest sample in D_i.

Assume some samples are mislabeled in the given dataset. For example, y_1 is a mislabeled sample in Fig. 1(b). According to the definition of fuzzy lower approximation, the membership of x to the fuzzy lower approximation of the class equals the distance between x and y_1. However, y_1 is a noisy sample. If we remove y_1, the membership of x

to the fuzzy lower approximation of its class is the distance between x and y. In this case, the membership significantly increases. In fact the memberships of all the samples in D_i increase if y_1 is removed. We see that noise has a great influence on fuzzy lower approximations.

3. Soft fuzzy rough sets

Before introducing the new model, we discuss the idea of soft margin SVM. We then introduce the idea of this technique into fuzzy rough sets.

3.1. Soft margin SVM

Let $X = \{(x_1, y_1), (x_2, y_2), \ldots, (x_l, y_l)\} \subset \mathbb{R}^n \times \{1, -1\}$ be a training set, $\phi(x)$ be a mapping function from X to a feature space and $g(x) = \langle \omega, \phi(x) \rangle + b$ be a 1-norm linear classification function. In [40], a bound on the generalization error of the linear function $g(x)$ with 1-norm in a kernel feature space was given by

$$P_D(y \neq g(x)) \leq \frac{1}{l\gamma} \sum_{i=1}^{l} \xi_i + \frac{4}{\gamma} \sqrt{\mathrm{tr}(K)} + 3 \sqrt{\frac{\ln(1/\delta)}{2l}}, \tag{8}$$

where γ is the margin (Fig. 2(a)), K is the kernel matrix for the training set and $\xi_i = (\gamma - y_i g(x_i))_+$ $(i = 1, 2, \ldots, l)$ is a slack variable. If sample x_i can be correctly classified with margin γ, $\xi_i = 0$; otherwise, $\xi_i > 0$.

As to hard margin SVM, all the training samples should be correctly classified with margin γ, the first term on the right-hand side of (8) doest not exist. So reducing the generalization error is equivalent to increasing margin γ. If data are inseparable, the margin may be reduced. We are not able to find a hyperplane which can correctly separate all the training samples with margin γ. Soft margin SVM was introduced to deal with this problem (Fig. 2(b)):

$$\begin{aligned} \min \quad & -\gamma + C \sum_{i=1}^{l} \xi_i \\ \text{subject to} \quad & y_i(\langle \omega, \phi(x_i) \rangle + b) \geq \gamma - \xi_i, \quad i = 1, 2, \ldots, l \\ & \xi_i \geq 0, \end{aligned} \tag{9}$$

where ξ_i $(i = 1, 2, \ldots, l)$ are slack variables.

Soft margin SVM makes tradeoff between the size of the margin and the number of misclassified samples. Thus the influence of noise on the final separating hyperplane is reduced. By this way, the generalization performance of the trained model would not be greatly influenced by the noise.

3.2. Soft fuzzy rough sets

Now we simulate soft margin SVM and introduce a robust rough set model. We first introduce the definitions of hard distance and soft distance.

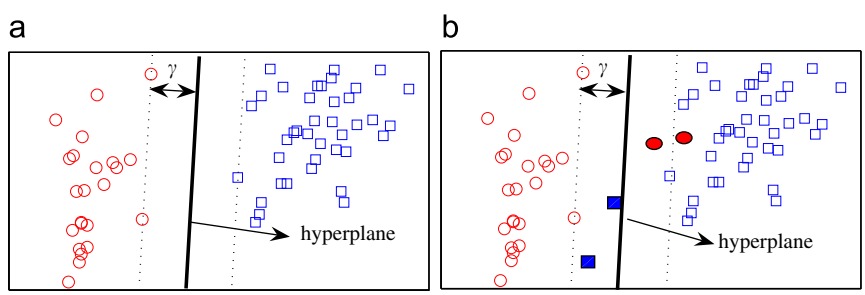

Fig. 2. Hard margin SVM and soft margin SVM.

Definition 2. Given an object x and a set of objects $Y = \{y_1, y_2, \ldots, y_n\}$, the hard distance between x and Y is defined as

$$HD(x, Y) = \min_{y_i \in Y} d(x, y_i), \tag{10}$$

where d is a distance function.

As we all know, the statistic *min* is sensitive to noise. We introduce a robust definition of distance here.

Definition 3. Given an object x and a set of objects $Y = \{y_1, y_2, \ldots, y_n\}$, the soft distance between x and Y is defined as

$$SD(x, Y) = \arg\max_{d(x,y_i)} \{d(x, y_i) - C \times m_i, y_i \in Y, i = 1, 2, \ldots, n\}, \tag{11}$$

where d is a distance function, C is a penalty factor and m_i is the number of samples which satisfy $d(x, y_j) < d(x, y_i)$, $j = 1, 2, \ldots, n$.

We give a toy example in Fig. 3, where x comes from $class_1$, the other samples belong to $class_2$ denoted by Y and $d_1 < d_2 < d_3 < d_4$. We see that $HD(x, Y)$ is d_1 according to the definition of hard distance. However, it seems that y_1 is a noisy sample. $HD(x, Y)$ may not exactly reflect the distance between x and Y. In this case soft distance can be used. If we take y_1 as a noisy sample and neglect it, $SD(x, Y)$ should be d_2; and if y_2 is also a noisy sample, $SD(x, Y)$ should be d_3. How many samples should be taken as noisy samples in this case? We add a penalty term to the distance. If we overlook one sample, $d(x, y_i)$ will minus C. For all candidate $d(x, y_i)$, we take $d'(x, y_i) = \arg_{d(x,y_i)} \max\{d(x, y_i) - C \times m_i\}$ as the soft distance between x and Y. That is to say the distance $d'(x, y_i)$ is the largest after punishing all the overlooked samples. Next, we show how to compute the soft distance between x and Y with an example.

Example 1. Given x and $Y = \{y_1, y_2, y_3, y_4, y_5, y_6\}$, $d(x, y_1) = 0.40$, $d(x, y_2) = 0.75$, $d(x, y_3) = 0.77$, $d(x, y_4) = 0.78$, $d(x, y_5) = 0.80$, $d(x, y_6) = 0.81$ and $C = 0.1$. $HD(x, Y) = 0.40$, the soft distance $SD(x, Y)$ is

$$\begin{aligned} SD(x, Y) &= \arg\max_{d(x,y_i)} \{0.40, 0.75 - 0.1 \times 1, 0.77 - 0.1 \times 2, 0.78 - 0.1 \times 3, 0.80 - 0.1 \times 4, 0.81 - 0.1 \times 5\} \\ &= \arg\max_{d(x,y_i)} \{0.40, 0.65, 0.57, 0.47, 0.40, 0.31\} = 0.75 \end{aligned}$$

Now we introduce the definition of soft fuzzy rough sets based on the soft distance.

Definition 4. Let U be a nonempty universe, R be a fuzzy equivalence relation on U and $F(U)$ be the fuzzy power set of U. Soft fuzzy lower and upper approximations of $F \in F(U)$ are defined as

$$\begin{cases} \underline{R}^S F(x) = 1 - R\left(x, \arg\sup_{y \ F(y) \le F(y_L)} \{1 - R(x, y) - C \times m\}\right), \\ \overline{R}^S F(x) = R\left(x, \arg\inf_{y \ F(y) \ge F(y_U)} \{R(x, y) + C \times n\}\right), \end{cases} \tag{12}$$

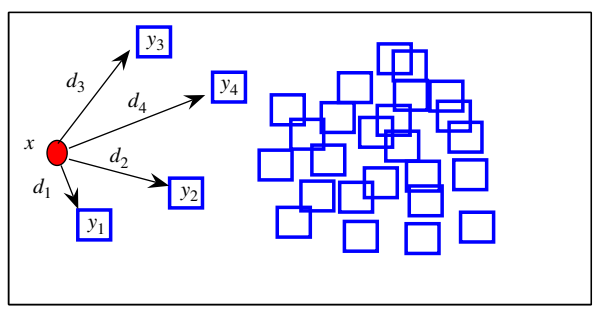

Fig. 3. Soft distance.

where

$$\begin{cases} y_L = \arg\limits_{y} \inf\limits_{y\in U} \max\{1 - R(x, y), F(y)\}, \\ y_U = \arg\limits_{y} \sup\limits_{y\in U} \min\{R(x, y), F(y)\}, \end{cases} \tag{13}$$

C is a penalty factor, m is the number of samples overlooked in computing $\underline{R}^S F(x)$ and n is the number of samples overlooked in computing $\overline{R}^S F(x)$.

If A is a crisp set, the membership of x to the soft fuzzy lower approximation of A is

$$\underline{R}^S A(x) = 1 - R(x, y_{AL}), \tag{14}$$

where

$$y_{AL} = \arg\limits_{y} \sup\limits_{A(y)=0} \{1 - R(x, y) - C \times m\} = \arg\limits_{y} \sup\limits_{A(y)=0} \{d(x, y) - C \times m\} = \arg\limits_{y} SD(x, U - A). \tag{15}$$

Obviously, $\underline{R}^S A(x)$ equals the soft distance from x to $U - A$.

Similarly, the membership of x to the soft fuzzy upper approximation of A is

$$\overline{R}^S A(x) = R(x, y_{AU}), \tag{16}$$

where

$$\begin{aligned} y_{AU} &= \arg\limits_{y} \inf\limits_{A(y)=1} \{R(x, y) + C \times n\} = \arg\limits_{y} \sup\limits_{A(y)=1} \{1 - R(x, y) - C \times n\} = \arg\limits_{y} \sup\limits_{A(y)=1} \{d(x, y) - C \times n\} \\ &= \arg\limits_{y} SD(x, A). \end{aligned} \tag{17}$$

$\overline{R}^S A(x)$ equals the similarity between x and the sample that is used to compute the soft distance from x to A.

We believe since the soft distance is more robust than the hard distance, soft fuzzy rough sets should be more robust to noise than the classical model.

3.3. Confidence analysis on soft fuzzy lower approximation

We first define a soft hypersphere associated with x. The center and radius of this hypersphere are x and $\underline{R}^S F(x)$, respectively. In the hypersphere, most samples come from the same class as x and several samples come from other classes. We call this hypersphere a soft hypersphere. While the hypersphere in which all the samples come from the same class of x is called a hard hypersphere. Here, we take the proportion of the samples with the same class label as x in the soft hypersphere as the confidence degree of the soft fuzzy lower approximation. In the following, we give the bound of confidence degree of soft fuzzy lower approximation. We introduce two theorems [40].

Theorem 1. *Given $\delta \in (0, 1)$, let $\mathcal{F}$ be a class of functions mapping from Z. Let $z_i{}_{i=1}^{l}$ be drawn independently according to a probability distribution $\mathcal{D}$. Then with probability at least $1 - \delta$ over random draws of samples of size l, every $f \in \mathcal{F}$ satisfies*

$$\mathbb{E}_{\mathcal{D}}[f(z)] \le \widehat{\mathbb{E}}[f(z)] + \mathcal{R}_l(\mathcal{F}) + \sqrt{\frac{\ln(2/\delta)}{2l}} \le \widehat{\mathbb{E}}[f(z)] + \widehat{\mathcal{R}}_l(\mathcal{F}) + 3\sqrt{\frac{\ln(2/\delta)}{2l}}. \tag{18}$$

In Theorem 1 $\mathbb{E}_{\mathcal{D}}(f(z))$ denotes the average of $f(z)$; $\widehat{\mathbb{E}}_{\mathcal{D}}(f(z))$ denotes the sample average of $f(z)$; $\mathcal{R}_l(\mathcal{F})$ denotes the Rademacher complexity of $\mathcal{F}$; and $\widehat{\mathcal{R}}_l(\mathcal{F})$ denotes the empirical Rademacher complexity of $\mathcal{F}$.

Theorem 2. *Let $\mathcal{F}$ be class of real functions. If $\mathcal{L}$: $\mathbb{R} \to \mathcal{R}$ is Lipschitz with constant L and satisfies: $\mathcal{L}(0) = 0$, then $\widehat{\mathcal{R}}_n(\mathcal{L} \circ \mathcal{F}) \le 2L\mathcal{R}_n(\mathcal{F})$.*

Let $U = X \cup Y$ be drawn independently from a probability distribution $\mathcal{D}$. $X = \{(x_1, d_x), \ldots, (x_m, d_x)\}$ is the set of the samples that have the same label as x and $Y = \{(y_1, d_{y_1}), \ldots, (y_n, d_{y_l})\}$ is the set of the samples that have different class labels from x's. According to Theorems 1 and 2 we give Theorem 3 for nonconfidence degree of soft fuzzy lower approximation.

Theorem 3. *Given $\gamma'_{x_0Y} > 0$, $\mathcal{F}$ is the class of functions given by*

$$f(\mathbf{x}) = -(\gamma_{x_0x} + d_x\gamma'_{x_0Y})/\gamma'_{x_0Y}, \tag{19}$$

where

$$d_x = \begin{cases} 1, & x \in X, \\ -1, & x \notin X. \end{cases} \tag{20}$$

For $x_0 \in X$, L_{x_0} is the degree of x_0 to the soft fuzzy lower approximation of the class, $|L_{x_0}| = l$. $\mathcal{H}(x)$ is Heaviside function,

$$\mathcal{H}(x) = \begin{cases} 1, & x > 0, \\ 0, & x \le 0. \end{cases} \tag{21}$$

Given $\delta \in (0, 1)$, with probability at least $1 - \delta$ over samples the nonconfidence degree of $x_0 \in X$ to the soft fuzzy lower approximation of the class satisfies

$$NConf(L_{x_0}) = \mathbb{E}_{\mathcal{D}}\{\mathcal{H}[f(x)]\} \le \frac{1}{l\gamma'_{x_0Y}}\sum_{i=1}^{l}\xi_i + \frac{2N}{l\gamma'_{x_0Y}}\left(1 - \frac{\gamma_{x_0Y}}{\gamma'_{x_0Y}}\right) + 3\sqrt{\frac{\ln(2/\delta)}{2l}}, \tag{22}$$

where $\xi_i = \max\{0, -(\gamma_{x_0x_i} + d_{x_i}\gamma'_{x_0Y})\}$, γ and γ' are HD and SD.

Proof. Consider the pattern function with loss function $\mathcal{L}$: $\mathbb{R} \to [0, 1]$, given by

$$\mathcal{L}(a) = a_+ = \begin{cases} a, & 0 < a < 1, \\ 0, & a \le 0. \end{cases} \tag{23}$$

It can be easily proven that $\mathcal{L}(f(x))$ satisfies Lipschitz and $|L| = 1/\gamma'_{x_0Y}$. Obviously, $\mathcal{H}(f(x)) \le \mathcal{L}(f(x))$. By Theorem 1 we have

$$\mathbb{E}_{\mathcal{D}}\{\mathcal{H}[f(x)]\} \le \mathbb{E}_{\mathcal{D}}\left\{L\left[-\left(\frac{\gamma_{x_0x_i} + d_{x_i}\gamma'_{x_0Y}}{\gamma'_{x_0Y}}\right)\right]\right\} \le \widehat{\mathbb{E}}\left\{L\left[-\left(\frac{\gamma_{x_0x_i} + d_{x_i}\gamma'_{x_0Y}}{\gamma'_{x_0Y}}\right)\right]\right\} + \widehat{\mathcal{R}}_l(\mathcal{L} \circ \mathcal{F}) + 3\sqrt{\frac{\ln(2/\delta)}{2l}}. \tag{24}$$

Due to $\mathcal{L}(a)$,

$$\mathbb{E}_{\mathcal{D}}\left\{L\left[-\left(\frac{\gamma_{x_0x_i} + d_{x_i}\gamma'_{x_0Y}}{\gamma'_{x_0Y}}\right)\right]\right\} \le \frac{1}{l\gamma'_{x_0Y}}\sum_{i=1}^{l}\xi_i + \widehat{\mathcal{R}}_l(\mathcal{L} \circ \mathcal{F}) + 3\sqrt{\frac{\ln(2/\delta)}{2l}}, \tag{25}$$

where $\xi_i = \max\{0, -(\gamma_{x_0x_i} + d_{x_i}\gamma'_{x_0Y})\}$. Let $L = 1/\gamma'_{x_0Y}$. Since $\mathcal{L}(0) = 0$, according to Theorem 2 we can get

$$\widehat{\mathcal{R}}_l(\mathcal{L} \circ \mathcal{F}) \le 2\widehat{\mathcal{R}}_l(\mathcal{F})/\gamma'_{x_0Y}. \tag{26}$$

Then the Rademacher complexity of the class $\mathcal{F}$ is bounded as

$$\begin{aligned}\widehat{\mathcal{R}}_l(\mathcal{F}) &= \mathbb{E}_\sigma\left[\sup_{f\in\mathcal{F}}\left|\frac{2}{l}\sum_{i=1}^{l}\sigma_i f(x_i)\right|\right] = \mathbb{E}_\sigma\left[\sup_{f\in\mathcal{F}'}\left|\frac{2}{l}\sum_{i=1}^{l}\sigma_i\left[-\left(\frac{\gamma_{x_0x_i} + d_{x_i}\gamma'_{x_0Y}}{\gamma'_{x_0Y}}\right)\right]\right|\right] \\ &= \mathbb{E}_\sigma\left[\sup_{f\in\mathcal{F}'}\left|\frac{2}{l}\sum_{i=1}^{l}\sigma_i\left(1 - \frac{\gamma_{x_0x_i}}{\gamma'_{x_0Y}}\right)\right|\right] = \frac{2N}{l}\left(1 - \frac{\gamma_{x_0Y}}{\gamma'_{x_0Y}}\right).\end{aligned} \tag{27}$$

Then the nonconfidence degree of the soft fuzzy lower approximation associated with x_0 satisfies

$$NConf(L_{x_0}) = \mathbb{E}_{\mathcal{D}}\{\mathcal{H}[f(x)]\} \le \frac{1}{l\gamma'_{x_0Y}}\sum_{i=1}^{l}\xi_i + \frac{2N}{l\gamma'_{x_0Y}}\left(1 - \frac{\gamma_{x_0Y}}{\gamma'_{x_0Y}}\right) + 3\sqrt{\frac{\ln(2/\delta)}{2l}}, \tag{28}$$

where $\xi_i = \max\{0, -(\gamma_{x_0x_i} + d_{x_i}\gamma'_{x_0Y})\}$.

From Theorem 3 we can see if $\gamma'_{x_0Y} = \gamma_{x_0Y}$ i.e. the degree of x_0 to its soft fuzzy lower approximation is 1, the first and second terms of (28) vanish. The nonconfidence degree of the soft fuzzy lower approximation is the least. With Theorem 3 the confidence degree of the soft fuzzy lower approximation satisfies

$$Conf(\mathcal{L}_{x_0}) \ge 1 - \left[\frac{1}{l\gamma'_{x_0Y}}\sum_{i=1}^{l}\xi_i + \frac{2N}{l\gamma'_{x_0Y}}\left(1 - \frac{\gamma_{x_0Y}}{\gamma'_{x_0Y}}\right) + 3\sqrt{\frac{\ln(2/\delta)}{2l}}\right], \tag{29}$$

where $\xi_i = \max\{0, -(\gamma_{x_0x_i} + d_{x_i}\gamma'_{x_0Y})\}$.

4. Comparison of soft fuzzy rough sets and other models

4.1. SFRS and VQRS

The vaguely quantified rough set (VQRS) model was constructed on a fuzzy quantification measure in [57].

Given a couple of fuzzy quantifiers (Q_l, Q_u) $(0 < l < u < 1)$, Q_l-upper and Q_u-lower approximations of A in approximation space (X, R) are defined as

$$\begin{cases} R \uparrow_{Q_l} A(y) = Q_l\left(\dfrac{|R_y \cap A|}{|R_y|}\right) = Q_l(R_A(y)), \\ R \downarrow_{Q_u} A(y) = Q_u\left(\dfrac{|R_y \cap A|}{|R_y|}\right) = Q_u(R_A(y)), \end{cases} \tag{30}$$

where A is a set and R_y is the equivalence class of y. If A is a fuzzy set, R_y is defined by $R_y(x) = R(x, y)$ for $x \in X$, $(R_y \cap A)(x) = \min(R_y(x), A(x))$ and the cardinality of a fuzzy set A is defined by $|A| = \sum_{x\in X} A(x)$. Then Q_l-upper and Q_u-lower approximations measure the degree of one fuzzy set R_y included by another fuzzy set A.

As to Pawlak's rough sets, if the equivalence class of x is completely contained by the crisp set $A \in X$, x is classified to the lower approximation of A. If a sample in the equivalence class of x is not contained by A, x is grouped into boundary, i.e. the membership of x to the lower approximation of A is 0 or 1. VQRS compute the membership in a robust way. It takes the proportion of samples in the equivalence class of x contained in the set A as a variable of Q_u to compute memberships. The larger the proportion is, the larger the membership of x to the lower approximation of A is. Then the membership of x to the lower approximation of A takes values in [0,1].

As to the fuzzy case, in VQRS, the sample proportion is replaced by the similarity proportion in computing the membership of x to the lower approximation of A. Given a similarity measure R, VQRS first computes the similarity between x and $y \in X$, i.e. $R(x, y)$. And then VQRS takes $\sum_{y\in A} R(x, y)/\sum_{y\in X} R(x, y)$ as a variable of Q_u to compute the lower approximation. The larger the similarity proportion is, the larger the lower approximation membership is. In this way, mislabeled samples and outliers will not have great influence on the memberships. However, we may get different memberships with different functions Q_u. And the value of u also has great impact on memberships. Besides, in [59], some limitations of VQRS were given. One main limitation is that some important properties about attribute reduction do not hold in this model. As to SFRS model, it reduces the influence of noise by overlooking some samples which we take as noisy samples. So our model is different from VQRS model.

4.2. SFRS and VPFRS

The variable precision rough set (VPRS) model is another robust model [25]. However, VPRS cannot deal with real-valued data. In [26], VPRS was extended into fuzzy rough sets, called variable precision fuzzy rough sets (VPFRS).

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

$$\mu_{\underline{R}_u}F(X_i)=\begin{cases}\inf\limits_{x\in S_{iu}} \vartheta(\mu_{X_i}(x), \mu_F(x)) & \text{if } \exists\alpha_u=\sup\{\alpha\in(0,1]: e_\alpha(X_i, F)\le 1-u\},\\ 0 & \text{otherwise,}\end{cases} \tag{31}$$

where $S_{i_u}=\sup p(X_i\cap X^F_{i_{\alpha_u}})$.

S_{i_u} contains the samples in X_i satisfying $\mu_F(x)\ge\alpha_u$ if such α_u exists. We can see that VPFRS is robust to mislabeled samples as $\mu_{\underline{R}_u}F(X_i)$ is computed with the samples that satisfy $\mu_F(x)>\alpha_u$, which is similar to SFRS model in the way of weakening the influence of noise. But, they are different in determining which samples should be neglected.

As to VPFRS, samples overlooked are determined by α_u if such α_u exists. And α_u is related to α-inclusion error $e(X_i, F)$. For two fuzzy sets X_i and F,

$$e(X_i, F)=1-\frac{|X_i\bigcap X_i{}^F_\alpha|}{|X_i|}=1-\frac{|X_i\bigcap(X_i\bigcap F)_\alpha|}{|X_i|}, \tag{32}$$

where $|X_i|=\sum_{i=1}^n X_i(x)$ and $(X_i\bigcap F)_\alpha=\{x\in X|\mu_{X_i\bigcap F}(x)>\alpha\}$. Here, α is computed with u. By this way, the memberships of samples to the lower and upper approximations of VPFRS would be influenced by the threshold α_u and so does the number of overlooked samples.

As to SFRS model, which samples to be ignored are determined by the tradeoff between the number and the memberships, which makes the number of samples ignored limited in augmenting or reducing memberships. Simultaneously, the memberships are also restricted. They would not enlarge or reduce too much.

4.3. SFRS and FVPRS

In [59], Zhao et al. introduced a robust model of fuzzy rough sets, called fuzzy variable precision rough sets (FVPRS). The lower and upper approximations of FVPRS were defined as

$$\begin{cases}\underline{R}_\alpha A(x)=\inf\limits_{A(u)\le\alpha}\max(1-R(x,u),\alpha)\wedge\inf\limits_{A(u)>\alpha}\max(1-R(x,u),A(u)),\\ \overline{R}_\alpha A(x)=\sup\limits_{A(u)\ge 1-\alpha}\min(R(x,u),1-\alpha)\vee\sup\limits_{A(u)<1-\alpha}\min(R(x,u),A(u)).\end{cases} \tag{33}$$

If $A(u)\le\alpha$, the model sets $A(u)=\alpha$ in computing $\underline{R}_\alpha A(x)$. In other words, the samples with $A(u)<\alpha$ are not considered. In computing $\overline{R}_\alpha A(x)$, if $A(u)\ge 1-\alpha$, the model sets $A(u)=1-\alpha$ i.e. the samples with $A(u)>1-\alpha$ are neglected. Consequently, the lower approximation of FVPRS increases and the upper approximation decreases.

If A is an arbitrary crisp subset of U, the lower and upper approximations of FVPRS degenerate into the following formulae. $\forall x\in U$,

$$\begin{cases}\underline{R}_\alpha A(x)=\inf\limits_{A(u)=0}\max(1-R(x,u),\alpha),\\ \overline{R}_\alpha A(x)=\sup\limits_{A(u)=1}\min(R(x,u),1-\alpha).\end{cases} \tag{34}$$

We can see that $\underline{R}_\alpha A(x)\ge\alpha$ and $\overline{R}_\alpha A(x)\le 1-\alpha$ for $\forall x\in U$.

Now we discuss the differences and similarities between FVPRS and SFRS. Given a sample set $D_i\in U/D$, both FVPRS and SFRS enlarge the membership of x to the lower approximation of A. FVPRS sets $\underline{R}_\alpha D_i(x)=\alpha$ if $\underline{R}_\alpha D_i(x)<\alpha$. In Fig. 4, x_1, x_2 and x_3 belong to D_i. In order to keep $\underline{R}_\alpha D_i(x_j)\ge\alpha$ $(j=1,2,3)$, some samples should not be considered in computing $\underline{R}_\alpha D_i(x_j)$. In terms of SFRS, we enlarge $\underline{R}^S D_i(x_j)$ by neglecting some samples, too.

We think there are some differences between these two models.

(1) In computing $\underline{R}_\alpha D_i(x)$, FVPRS sets $\underline{R}_\alpha D_i(x)=\alpha$ if $\underline{R}_\alpha D_i(x)<\alpha$ without any penalty, so it does not limit the number of the samples neglected. However, in computing $\underline{R}^S D_i(x)$, we punish $\underline{R}^S D_i(x)$ to limit the number of the samples neglected. We make tradeoff between $\underline{R}^S D_i(x)$ and the number of the misclassified samples.

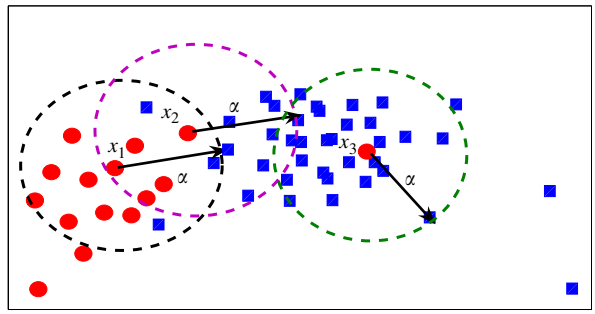

Fig. 4. Robustness of $\underline{R_\alpha} D_i(x)$.

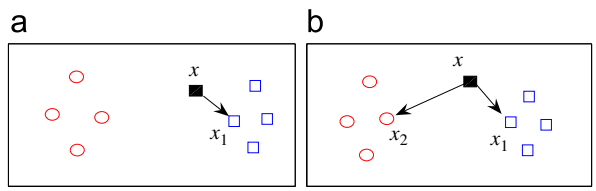

Fig. 5. Nearest neighbor rule and fuzzy rough classifier.

(2) It seems FVPRS is not robust to mislabeled samples, as shown in Fig. 4. Assume x_3 is a noisy sample. We should not consider it in computing the fuzzy lower approximation of samples around x_3. This is true if SFRS is used. But FVPRS sets the memberships of these samples to the fuzzy lower approximations α. In fact, the memberships should be much larger than α if x_3 is removed.

(3) If x_3 is considered as a noisy sample, $\underline{R}^S D_i(x_3)$ should be very small as x_3 is not much certain to be classified into D_i. This is true with respect to SFRS because we need to neglect a lot of samples to increase $\underline{R}^S D_i(x_3)$. In this case, a great penalty will be performed on it according to Definition 4. However, FVPRS sets $\underline{R_\alpha} D_i(x_3) = \alpha$ if $\underline{R_\alpha} D_i(x_3) \leq \alpha$. Obviously, it is not reasonable.

5. Robust classifier based on soft fuzzy rough sets

In this section, we design a robust classifier with soft fuzzy lower approximation. The idea of this classifier comes from nearest neighbor rule (NN) [8,9]. Sample x is classified to the class of the nearest neighbor of x. An example is given in Fig. 5(a). x_1 comes from class c_1 and x is an unseen sample. With NN rule, x should be classified into class c_1.

Fuzzy rough classifier (FRC) is designed as follows. Given a set of training samples with m classes, x is an unseen sample. We compute m memberships of x to fuzzy lower approximations of m classes. Finally, x is classified to the class producing the maximal membership as x belongs to this class with the greatest consistency.

We illustrate the fuzzy rough classifier with Fig. 5(b). x_1 and x_2 come from c_1 and c_2, respectively. x is a sample to be classified. If the nearest neighbor rule is used, x should be assigned to the class that the nearest neighbor of x belongs to. Clearly, x should be classified to c_1 for $\|x_1 - x\| < \|x_2 - x\|$. With FRC, x should be classified to the class with the largest membership of x to the fuzzy lower approximation of the class. If x belongs to c_1, the membership of x to the fuzzy lower approximation of c_1 is $1 - R(x_2, x)$. If x belongs to c_2, the membership of x to the fuzzy lower approximation of c_2 is $1 - R(x_1, x)$. If we introduce (6) to compute similarity between objects, we get $1 - R(x_2, x) > 1 - R(x_1, x)$ as $\|x_2 - x\| > \|x_1 - x\|$. Then x should be classified to c_1. In fact, FRC is the same as NN rule if we view $1 - R(x, y)$ as a generalized distance function.

Now we replace the fuzzy lower approximation with the soft fuzzy lower approximation for a robust classifier. We call it soft fuzzy rough classifier (SFRC). It works in a similar way with FRC. We compute the memberships of an unseen sample to the soft fuzzy lower approximations of each class. Given a training set with k classes, sample x is to be classified. First, suppose x belong to each class. We compute k memberships of x to the soft fuzzy lower approximations

of k classes. Finally, x is classified to the class with the maximal membership, formulated as

$$class_i(x) = \arg\max_{class_i} \{\underline{R}^S class_1(x), \underline{R}^S class_2(x), \ldots, \underline{R}^S class_k(x)\}, \tag{35}$$

where $\underline{R}^S class_i(x)$ is the membership of x to the soft fuzzy lower approximation of $class_i$.

Formally, the classification algorithm is given in Table 1.

In this algorithm, we compute the memberships of test samples to the fuzzy lower approximation of each class. In this step, we should compute the distance between the test sample and each training sample, and then sort the distance. The computational time is $n+nlogn$ for every test sample, where $nlogn$ is used in ranking samples. As to m test samples, the overall time is $m \times (n + nlogn)$.

6. Experiments

In this section, we discuss how to set parameter C, and we also conduct experiments on several datasets to test the robustness of the proposed classifier. The summary of the datasets is shown in Table 2.

Table 1
Soft fuzzy rough classifier.

Input	training set $X = \{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$ and test set $X' = \{x'_1, x'_2, \ldots, x'_m\}$
Process	$label \leftarrow \emptyset$
4	$classnum = \max(y_1, y_2, \ldots, y_n) - \min(y_1, y_2, \ldots, y_n) + 1$
5	for $i = 1 : m$
6	$degree \leftarrow \emptyset$
7	for $class = 1 : classnum$
8	$degree(class) \leftarrow \underline{R}^S class(x'_i)$
9	end
10	$class^* \leftarrow \arg\max_{class}(degree)$
11	$label(x'_i) \leftarrow class^*$
12	end
13	return $label$
Output	$label$

Table 2
Summary of datasets.

Data	Source	Size	Feature	Class
Australian-credit	[2]	690	14	2
Breast-cancer	[2]	699	9	2
Diabetes	[2]	768	8	2
Heart	[2]	270	13	2
Image-segmentation	[2]	2310	19	7
Ionosphere	[2]	351	34	2
Iris	[2]	150	4	3
Thyroid-gland	[2]	215	5	3
Sonar	[2]	208	60	2
Wine	[2]	178	13	3
WDBC	[2]	569	30	2
WPBC	[2]	198	33	2
ICU	[15]	200	19	2
Rice	[29]	105	5	2
Veteran-lung-cancer	[14]	137	7	2

6.1. Parameter discussion

Parameter C should be set in computing soft fuzzy approximations. From the definition of the soft distance we see that the number of the overlooked samples is determined by C. If C is too large, the soft fuzzy lower approximation will degenerate to fuzzy lower approximation and is sensitive to noise; and if C is too small, there may be too many overlooked samples, which would reduce the performance of SFRC. Thus, it is important to select a proper value for C. Now, we show the relationship between C and the number of the overlooked samples with two examples.

Example 2. Given a set of objects $Y=\{y_1, y_2, y_3, y_4, y_5\}$ and sample x, $d(x, y_1)=0.21$, $d(x, y_2)=0.49$, $d(x, y_3)=0.50$, $d(x, y_4)=0.52$, $d(x, y_5)=0.55$, $C=0.06$. $HD(x, Y)=0.21$ and the soft distance $SD(x, Y)$ is

$$\begin{aligned} SD(x, Y) &= \arg\max_{d(x,y_i)} \{0.21, 0.49-0.06\times 1, 0.50-0.06\times 2, 0.52-0.06\times 3, 0.55-0.06\times 4\} \\ &= \arg\max_{d(x,y_i)} \{0.21, 0.43, 0.38, 0.34, 0.31\} = 0.49 \end{aligned}$$

Example 3. Given a set of objects $Y=\{y_1, y_2, y_3, y_4, y_5\}$ and sample x, $d(x, y_1)=0.45$, $d(x, y_2)=0.49$, $d(x, y_3)=0.50$, $d(x, y_4)=0.52$, $d(x, y_5)=0.55$, $C=0.06$. $HD(x, Y)=0.45$ and the soft distance $SD(x, Y)$ is

$$\begin{aligned} SD(x, Y) &= \arg\max_{d(x,y_i)} \{0.45, 0.49-0.06\times 1, 0.50-0.06\times 2, 0.52-0.06\times 3, 0.55-0.06\times 4\} \\ &= \arg\max_{d(x,y_i)} \{0.45, 0.43, 0.38, 0.34, 0.31\} = 0.45 \end{aligned}$$

We can see that the soft distance satisfies

$$SD(x, Y) = \begin{cases} SD(x, Y), \max\limits_{y\in Y}\{d(x, y)-C\times m\} > HD(x, Y), \\ HD(x, Y), \max\limits_{y\in Y}\{d(x, y)-C\times m\} \le HD(x, Y), \end{cases} \tag{36}$$

where m is the number of samples with different labels in the soft hypersphere. From (36) we get

$$m < \frac{SD(x, Y)-HD(x, Y)}{C}. \tag{37}$$

That is to say if the radius increases C, at most one sample with different labels can be included in the soft hypersphere.

According to (37), we introduce an approach to selecting C. Fig. 6 shows the relationship between radiuses and nonconfidence degrees of soft hyperspheres, where x-axis is the rank of samples based on the soft distance from x_i ($i=1, 2$) to the samples with different labels. y-axis is the soft distance and nonconfidence degree of soft hypersphere (error rate). We can see that the nonconfidence degrees of soft hyperspheres increase along with soft distance. In addition, there are several peaks with error rate.

The corresponding dataset of Fig. 6 is given in Table 3. We consider x_1. If we require that the confidence degree of a soft hypersphere is greater than 95%, the sample error rate in the soft hypersphere should be less than or equal to 5%. From Table 3, the error rate is 3.7% and the radius of the soft hypersphere is 0.21. There are two samples with different labels in the soft hypersphere. In this case, the radius of the soft hypersphere enlarges 0.21 if two samples are overlooked. So C_{x_1}=0.11. Similarly, we can compute the corresponding values of C for each sample. Finally, we take the average or median of all the values of C as the final value of C for the dataset.

Now, we test this method with experiments. Here, we set the confidence degree of soft hypersphere as 97%. In Table 4, $CORR_1$ and $CORR_2$ are the average numbers of samples that are correctly classified in hard hyperspheres and soft hyperspheres, respectively. HD and SD are the average hard distance and soft distance. ERR_2 is the average number of the samples with different labels in soft hyperspheres. $\Delta D = SD - HD$ and R is the confidence degree of soft hypersphere.

We consider noise level 6%. $C=0.13$ means that any sample cannot be overlooked until SD enlarges 0.13. HD equals 0.54 and SD equals 0.58 in this case. So SD increases 0.04. If we allow a sample with different labels in the soft hypersphere, SD increases 0.2. The condition is satisfied. SD enlarges 0.13 if one sample is considered noisy. The confidence degree of soft hypersphere is greater than 97%. Therefore, we should set $C=0.13$.

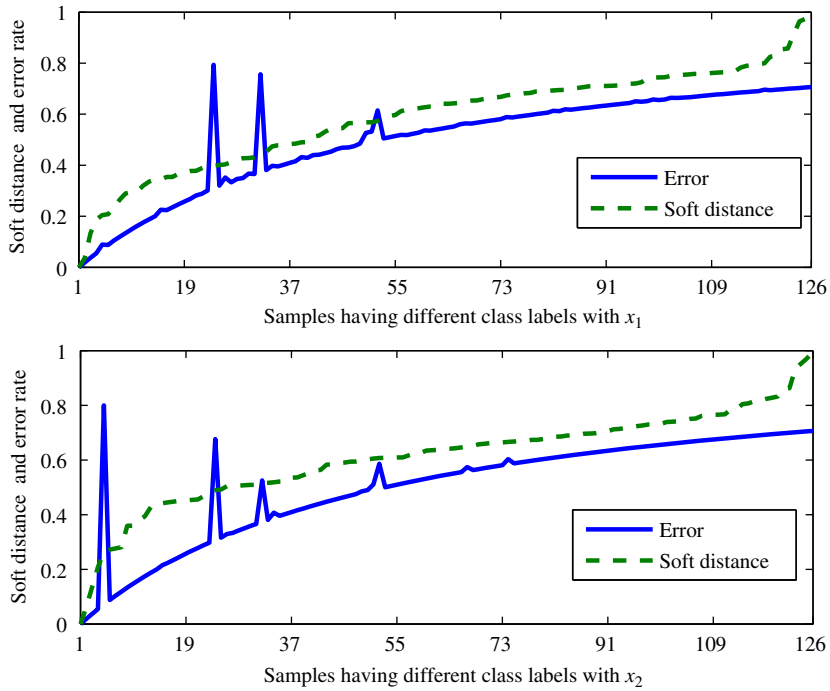

Fig. 6. Relationship between radiuses and nonconfidence degrees of soft hyperspheres.

Table 3
Corresponding data of Fig. 6

	No.	1	2	3	4	5	6	7	8	…
x_1	*Error*	0	1.8%	3.7%	5.4%	80.0%	88.0%	10.3%	11.9%	…
	Soft distance	0	0.14	0.21	0.26	0.27	0.28	0.28	0.28	…
x_2	*Error*	0	1.8%	3.8%	5.5%	8.8%	8.8%	10.5%	11.9%	…
	Soft distance	0	0.03	0.13	0.19	0.20	0.21	0.23	0.27	…

Table 4
Validating the method of setting C.

Noise levels	C	FRS		SFRS				
		$CORR_1$	HD	ERR_2	$CORR_2$	SD	ΔD	R (%)
3%	0.13	15.03	0.56	0.29	21.62	0.62	0.06	98.68
6%	0.14	15.84	0.54	0.20	18.75	0.58	0.04	98.94
9%	0.13	9.26	0.49	0.16	10.90	0.52	0.03	98.55
12%	0.12	7.57	0.48	0.22	10.38	0.52	0.04	97.45
15%	0.15	6.06	0.45	0.09	6.88	0.48	0.03	98.71
Average	0.11	7.46	0.47	0.23	9.57	0.51	0.04	97.65

6.2. Comparison of classification performance

Now we conduct experiments to compare the soft fuzzy rough classifier with other algorithms.

First, we compute the classification performance of SFRC, FRC, NN, CART and BN (BayesNet) on the datasets with 10-fold cross-validation technique. The confidence degree is set as 90%. Table 5 gives the classification accuracies

Table 5
Classification accuracies (%) on real-world tasks.

Data	SFRC	FRC	NN	CART	BN
Wine	95.4 ± 4.6	94.4 ± 0.9	94.4 ± 0.9	89.3 ± 2.5	98.3 ± 0.9
WDBC	96.9 ± 2.3	95.8 ± 1.1	95.8 ± 1.1	91.9 ± 3.0	95.1 ± 3.2
WPBC	76.8 ± 4.2	66.2 ± 4.9	66.2 ± 4.9	69.2 ± 8.9	74.7 ± 4.2
Diabetes	75.4 ± 2.7	70.6 ± 2.2	70.6 ± 2.2	70.2 ± 2.9	74.3 ± 2.0
Heart	80.4 ± 4.3	76.7 ± 9.4	76.7 ± 9.4	74.1 ± 6.3	81.1 ± 8.9
Ionosphere	86.4 ± 2.7	86.4 ± 5.0	86.4 ± 5.0	87.3 ± 7.4	89.5 ± 5.6
Sonar	87.1 ± 7.6	87.1 ± 7.6	87.1 ± 7.6	72.1 ± 13.9	79.3 ± 7.0
Australian-credit	87.6 ± 3.2	79.6 ± 5.5	79.6 ± 5.5	82.6 ± 4.5	84.9 ± 5.7
Breast-cancer	95.3 ± 3.8	94.9 ± 4.0	94.9 ± 4.0	93.6 ± 4.9	97.1 ± 3.0
Iris	96.7 ± 4.7	95.3 ± 7.1	95.3 ± 7.1	96.7 ± 3.5	92.7 ± 4.1
Thyroid-gland	95.3 ± 3.9	95.3 ± 3.9	95.3 ± 3.9	93.0 ± 7.0	94.4 ± 4.4
Image-segmentation	97.2 ± 1.8	97.2 ± 1.8	97.2 ± 1.8	95.4 ± 1.8	91.5 ± 2.2
Rice	89.6 ± 11.3	80.0 ± 11.0	80.0 ± 11.0	82.1 ± 11.7	76.9 ± 9.8
ICU	94.1 ± 8.5	86.8 ± 12.3	86.8 ± 12.3	79.4 ± 31.6	91.5 ± 11.3
Average	89.6	86.2	86.2	84.1	87.2

Table 6
Performance comparison of fuzzy ID3, FRCT, FRC and SFRC.

Data	Fuzzy ID3 [3]	FRCT [3]	FRC	SFRC
Diabetes	72.4	71.5	70.8	75.4
Iris	96.0	96.2	95.3	96.7
Wine	70.0	77.0	94.9	95.4
Breast-cancer	90.9	95.3	94.9	95.3
Heart	80.3	76.1	76.7	80.4
Australian-credit	85.5	83.0	79.6	87.6
Thyroid-gland	82.4	82.4	95.3	95.3
Image-segmentation	44.3	70.0	97.2	97.2
Rice	94.3	95.0	80.0	89.6
ICU	84.0	84.0	86.8	94.1
Average	81.2	83.5	84.7	88.8

computed with different algorithms. We see that FRC produces the same accuracy as NN; SFRC obtains the highest accuracies among these classifiers in most tasks.

The average accuracy does not show the confidence degree on which classifier is better than another. Now we use t-test to compare different classifiers [49].

Let A and B be two classifiers. Given a dataset, we split it into training set and test set. P_A and P_B are test accuracies with classifiers A and B, respectively. Repeat the above process k times. And we get k accuracies $P_A^{(i)}$ $(i = 1, 2, \ldots, k)$ and $P_B^{(i)}$ $(i = 1, 2, \ldots, k)$ with classifiers A and B, respectively. Let $P^{(i)} = P_A^{(i)} - P_B^{(i)}$ $(i = 1, 2, \ldots, k)$. Assume that $P^{(i)}$ is independently drawn from a normal distribution. Then we can apply t-test, by computing the statistic

$$t = \frac{\overline{P}\sqrt{(k)}}{\sqrt{\sum_{i=1}^{k}(P^{(i)} - \overline{P})^2/(k-1)}}, \tag{38}$$

where $\overline{P} = \sum_{i=1}^{k} P^{(i)}$. Under the null hypothesis (H_0: There is no difference between classifiers A and B), this statistic satisfies t distribution with $k - 1$ degrees of freedom. If t is greater than the tabulated value for the given confidence degree with $k - 1$ degrees of freedom, we reject H_0 and accept that there are significant difference between the two models.

Now we use the above method to compare SFRC with other classifiers.

Table 7
Classification accuracies (%) of datasets in noisy environment.

Data	Noise levels	Classifiers				
		SFR	FR	NN	CART	BN
Wine	3%	96.3 ± 5.4	92.5 ± 5.5	92.5 ± 5.5	90.2 ± 5.9	95.2 ± 0.5
	6%	95.6 ± 5.5	89.1 ± 7.3	89.1 ± 7.3	87.5 ± 7.6	92.1 ± 1.1
	9%	95.6 ± 5.8	86.7 ± 7.7	86.7 ± 7.7	85.6 ± 8.3	88.9 ± 1.5
	12%	95.0 ± 5.3	83.9 ± 7.9	83.9 ± 7.9	83.4 ± 7.5	83.9 ± 1.3
	15%	94.7 ± 5.7	81.3 ± 9.0	81.3 ± 9.0	80.3 ± 9.5	81.4 ± 1.0
WDBC	3%	96.0 ± 3.4	93.0 ± 3.3	93.0 ± 3.3	89.7 ± 4.0	92.1 ± 0.9
	6%	95.3 ± 3.6	90.4 ± 4.7	90.4 ± 4.7	87.8 ± 4.6	88.9 ± 0.9
	9%	94.2 ± 3.6	87.4 ± 5.2	87.4 ± 5.2	84.5 ± 5.7	85.8 ± 0.3
	12%	93.1 ± 3.9	84.9 ± 4.9	84.9 ± 4.9	81.9 ± 5.7	83.3 ± 1.2
	15%	90.2 ± 4.4	81.9 ± 5.4	81.9 ± 5.4	80.3 ± 5.7	80.7 ± 1.3
WPBC	3%	76.3 ± 4.1	69.3 ± 7.3	69.3 ± 7.3	66.7 ± 12.0	72.9 ± 1.9
	6%	75.7 ± 4.5	67.3 ± 7.4	67.3 ± 7.4	65.5 ± 10.7	71.6 ± 1.9
	9%	75.1 ± 5.6	65.5 ± 8.9	65.5 ± 8.9	62.9 ± 12.1	70.3 ± 2.5
	12%	76.1 ± 5.1	65.1 ± 8.9	65.1 ± 8.9	62.8 ± 11.3	69.8 ± 2.5
	15%	73.8 ± 5.3	64.7 ± 9.8	64.7 ± 9.8	62.9 ± 11.1	68.3 ± 3.2
Diabetes	3%	73.1 ± 4.0	69.7 ± 4.1	69.7 ± 4.1	70.4 ± 4.6	75.2 ± 2.1
	6%	71.9 ± 3.6	68.3 ± 4.2	68.3 ± 4.2	68.9 ± 4.8	71.2 ± 1.3
	9%	70.5 ± 4.1	67.2 ± 4.4	67.2 ± 4.4	66.4 ± 4.7	69.1 ± 1.2
	12%	68.9 ± 3.8	65.9 ± 4.4	65.9 ± 4.4	65.5 ± 5.3	66.0 ± 1.7
	15%	68.0 ± 4.9	64.4 ± 4.8	64.4 ± 4.8	64.6 ± 5.1	64.1 ± 4.4
Heart	3%	80.9 ± 5.4	75.0 ± 9.4	75.0 ± 9.4	74.0 ± 7.9	80.1 ± 1.0
	6%	78.4 ± 6.2	73.0 ± 9.3	73.0 ± 9.3	74.0 ± 9.4	78.7 ± 2.6
	9%	76.8 ± 7.0	71.7 ± 10.0	71.7 ± 10.0	71.6 ± 8.2	77.1 ± 2.7
	12%	76.1 ± 7.1	69.6 ± 9.0	69.6 ± 9.0	69.4 ± 8.1	77.0 ± 2.1
	15%	74.9 ± 7.6	68.7 ± 8.8	68.7 ± 8.8	68.6 ± 8.6	72.0 ± 2.6
Ionosphere	3%	85.2 ± 5.0	84.2 ± 5.4	84.2 ± 5.4	85.9 ± 6.7	87.0 ± 0.7
	6%	83.5 ± 6.2	81.9 ± 4.8	81.9 ± 4.8	84.6 ± 7.4	84.4 ± 0.6
	9%	82.4 ± 6.4	80.1 ± 6.7	80.1 ± 6.7	81.3 ± 8.5	82.0 ± 0.9
	12%	79.3 ± 7.3	77.8 ± 7.7	77.8 ± 7.7	79.2 ± 8.2	78.9 ± 0.8
	15%	78.8 ± 7.3	75.8 ± 8.1	75.8 ± 8.1	77.0 ± 9.2	77.9 ± 1.6
Sonar	3%	85.2 ± 7.8	84.4 ± 7.8	84.4 ± 7.8	72.1 ± 10.3	73.8 ± 1.4
	6%	83.6 ± 7.2	82.8 ± 7.8	82.8 ± 7.8	69.2 ± 10.3	72.9 ± 2.4
	9%	81.6 ± 8.5	80.9 ± 8.9	80.9 ± 8.9	68.2 ± 10.4	69.3 ± 3.4
	12%	79.2 ± 7.8	77.7 ± 8.7	77.7 ± 8.7	67.4 ± 11.9	67.3 ± 4.5
	15%	77.7 ± 9.4	76.5 ± 9.8	76.5 ± 9.8	65.0 ± 11.1	63.0 ± 4.0
Average		82.3	77.1	77.1	74.8	77.5

Accuracies computed with SFRC, FRC, CART and BN are denoted by $P^{(i)}_{SFRC}$, $P^{(i)}_{FRC}$, $P^{(i)}_{CART}$ and $P^{(i)}_{BN}$ ($i = 1, 2, \ldots, 14$). First, we assume that SFRC and FRC (NN) are the same. With (38), $t = 3.1531 > t_{0.995}(13) = 3.0123$. So we reject the assumption and accept SFRC and FRC (NN) are different with 99.5% confidence degree. Similarly, for SFRC and CART, $t = 4.1829 > t_{0.995}(13) = 3.0123$; and for SFRC and BN, $t = 1.9770 > t_{0.975}(13) = 1.7709$. Consequently, we accept that SFRC is different from FRC, NN, CART and BN with high confidence degrees.

From Table 5 we know the average accuracy of SFRC is higher than other classifiers. The above analysis shows SFRC outperforms FRC, NN, CART and BN.

Now we compare SFRC with FRCT [3] and fuzzy ID3. The datasets used here are the same as those in [3]. Table 6 shows classification performance comparison of fuzzy ID3, FRCT, FRC and SFRC, where accuracies computed with

fuzzy ID3 and FRCT were collected from [3]. We can see that SFRC produces higher average accuracy than FRCT and fuzzy ID3 on nine of 10 tasks.

Now we also use the statistical method to compare SFRC with FRCT, fuzzy ID3 and FRC. We assume that SFRC is the same as other classifiers. For SFRC and FRCT, with (38), $t = 2.3067 > t_{0.975}(9) = 2.2622$. So we reject the assumption. As to SFRC and fuzzy ID3, $t = 1.8522 > t_{0.95}(9) = 1.8331$; and for SFRS and FRC, $t = 3.0590 > t_{0.99}(9) = 2.8214$. We accept that SFRC is different from fuzzy ID3 and FRC with 95% and 99% confidence degrees, respectively. SFRC is better than FRC, FRCT and fuzzy ID3.

In order to test the robustness of SFRC, we add noise into training sets in this work. The experimental process is described as follows.

Given a set of data, we divide it into 10 subsets. One subset is taken as the test set and the others are taken as the training samples. We randomly draw $i\%$ training samples and revise their class labels. We consider the training set is corrupted by noise and the noise level is $i\%$. The noisy set is used to train a model, and we predict labels of test samples with SFRC and compute the accuracy. And then we take each subset as a test set, and get 10 accuracies. We compute the average accuracy as final outputs. In order to reduce variance, we repeat the above process 10 times and take the average values as the final results.

Table 7 presents the classification performance in the noisy environment. Noise levels are 3%, 6%, 9%, 12% and 15%, respectively. We can see the soft fuzzy rough classifier produces the highest accuracies on all the datasets.

We also perform t-test to compare the classifiers. The accuracies computed with SFRC is denoted by $P^{(i)}_{SFRC}$ $(i = 1, 2, \ldots, 35)$ and those with FRC is by $P^{(i)}_{FRC}$ $(i = 1, 2, \ldots, 35)$. And $P^{(i)} = P^{(i)}_{SFRC} - P^{(i)}_{FRC}$ $(i = 1, 2, \ldots, 35)$. With (38), for SFRC and FRC, $t = 3.6656 > t_{0.995}(34) = 2.7284$. So we reject the assumption and accept SFRC and FRC are different with 99.5% confidence degree. Similarly, for SFRC and CART, $t = 3.9796 > t_{0.995}(34) = 2.7284$. And for SFRC and BN, $t = 2.3752 > t_{0.975}(34) = 2.0322$. Obviously, SFRC is different from CART and BN classifiers at high confidence.

7. Conclusions and future work

Fuzzy rough sets were proposed as a mathematical tool to deal with uncertainty. Noise is a main source of uncertainty in real-world applications. We show that fuzzy rough sets are very sensitive to mislabeled samples. we introduce a robust model of fuzzy rough sets in this work, called soft fuzzy rough sets and discuss the connections between the soft fuzzy rough set model and other models. In addition, we design a soft fuzzy rough classifier based on the model. Some experiments are conducted to show the effectiveness of the model. The following conclusions are drawn from this work.

(1) The classical model of fuzzy rough sets is sensitive to class noise; one mislabeled sample may completely change fuzzy approximations.

(2) Soft fuzzy rough sets reduce the influence of the mislabeled samples in training data by overlooking some outliers in computing fuzzy lower and upper approximations. This strategy is effective in dealing with *class* noise.

(3) The parameter used in soft fuzzy rough sets can be computed from data according to the level of noise. The experiments show soft fuzzy rough classifier outperforms NN rule, CART, FRCT and BayesNet in the noisy context.

In this work we just discuss the *class* noise, while attribute noise is not considered. We will focus on designing a robust rough classifier for data corrupted with attribute noise in the future.

Acknowledgments

This work is partially supported by National Natural Science Foundation of China under Grants 60703013, 10978011 and The Hong Kong Polytechnic University (G-YX3B). Prof. Yu is supported by National Science Fund for Distinguished Young Scholars under Grant 50925625.

References

[1] F. Angiulli, C. Pizzuti, Fast outlier detection in high dimensional spaces, in: Proceedings of the Sixth European Conference on the Principles of Data Mining and Knowledge Discovery, vol. 2341, August 2002, pp. 15–26.

[2] C.L. Blake, C.J. Merz, UCI Repository of Machine Learning Databases, 1998. Available: ⟨http://www.ics.uci.edu/mlearn/MLRepository.html⟩.

[3] R.B. Bhatt, M. Gopal, FRCT: fuzzy-rough classification trees, Pattern Analysis and Applications 11 (2008) 73–88.
[4] D.R. Chen, Q.W. Yi, M. Ying, D.X. Zhou, Support vector machine soft margin classifiers: error analysis, Journal of Machine Learning Research 5 (2004) 1143–1175.
[6] Y. Chen, X. Dang, H. Peng, H.L. Bart Jr., Outlier detection with the kernelized spatial depth function, IEEE Transactions on Pattern Analysis and Machine Intelligence 31 (2009) 288–305.
[7] C. Cornelis, M.D. Cock, A.M. Radzikowska, Vaguely quantified rough sets, Lecture Notes in Computer Science in Rough Sets, Fuzzy Sets, Data Mining and Granular Computing, vol. 4482, Springer, Berlin, Heidelberg, 2007, pp. 87–94.
[8] T.M. Cover, P.E. Hart, Nearest neighbor pattern classification, IEEE Transactions on Information Theory 13 (1) (1967) 21–27.
[9] B.V. Dasarathy (Ed.), Nearest Neighbor NN Norms: NN Pattern Classification Techniques, IEEE Computer Society, 1990.
[10] D. Dubois, H. Prade, Rough fuzzy sets and fuzzy rough sets, International Journal of General Systems 17 (1990) 191–209.
[11] M. Ester, H.P. Kriegel, J. Sander, X. Xu, A density-based algorithm for discovering clusters in large spatial databases with noise, in: Proceedings of 2nd International Conference on Knowledge Discovery and Data Mining, 1996, pp. 226–231.
[12] H. Fan, K. Ramamohanarao, Noise tolerant classification by chi emerging patterns, Lecture Notes in Computer Science, Advances in Knowledge Discovery and Data Mining, vol. 3056, Springer, Berlin, Heidelberg, 2004, pp. 201–206.
[14] ⟨http://lib.stat.cmu.edu/datasets/veteran⟩.
[15] ⟨http://lib.stat.cmu.edu/DASL/Datafiles/ICU.html⟩.
[16] Q.H. Hu, S. An, D.R. Yu, Soft fuzzy rough sets for robust feature evaluation and selection, Information Sciences 180 (2010) 4384–4400.
[17] Q.H. Hu, Z.X. Xie, D.R. Yu, Hybrid attribute reduction based on a novel fuzzy-rough model and information granulation, Pattern Recognition 40 (2007) 3509–3521.
[18] Q.H. Hu, D.R. Yu, Z.X. Xie, Information-preserving hybrid data reduction based on fuzzy-rough techniques, Pattern Recognition Letters 27 (2006) 414–423.
[19] K.Z. Huang, H.Q. Yang, I. King, M.R. Lyu, Max-min margin machine: learning large margin classifiers locally and globally, IEEE Transactions on Neural Networks 19 (2008) 260–272.
[20] R. Jensen, C. Cornelis, A new approach to fuzzy-rough nearest neighbour classification, in: Proceedings of the 6th International Conference on Rough Sets and Current Trends in Computing, USA, 2008, vol. 5306, pp. 310–319.
[21] R. Jensen, Q. Shen, New approaches to fuzzy-rough feature selection, IEEE Transactions on Fuzzy Systems f17 (2009) 824–838.
[22] E.M. Knorr, R.T. Ng, V. Tucakov, Distance-based outliers: algorithms and applications, VLDB Journal: Very Large Databases 8 (2000) 237–253.
[23] G.R.G. Lanckriet, L.E. Ghaoui, C. Bhattacharyya, M.I. Jordan, A robust minimax approach to classification, Journal of Machine Learning Research 3 (2002) 555–582.
[25] A. Mieszkowicz-Rolka, L. Rolka, Variable precision rough sets, in: J. Soldek, L. Drobiazgiewicz (Eds.), Evaluation of Human Operator's Decision Model, Artificial Intelligence and Security in Computing Systems, Kluwer Academic Publishers, Boston, Dordrecht, London, 2003.
[26] A. Mieszkowicz-Rolka, L. Rolka, Variable precision fuzzy rough sets, in: F.P. James, A. Skowron (Eds.), Transactions on Rough Sets I, vol. 3100, Springer, Berlin, Heidelberg, 2004, pp. 144–160.
[27] N.N. Morsi, M.M. Yakout, Axiomatics for fuzzy rough sets, Fuzzy Sets and Systems 100 (1998) 327–342.
[28] H.S. Nguyen, Approximate boolean reasoning: foundations and applications in data mining, Transactions on Rough Sets V, Lecture Notes in Computer Science, vol. 4100, 2006, pp. 344–523.
[29] K. Nozaki, H. Ishibuchi, H. Tanaka, A simple but powerful heuristic method for generating fuzzy rules from numerical data, Fuzzy Sets and Systems 86 (1997) 251–270.
[30] Z. Pawlak, Rough sets, International Journal of Computer and Information Sciences 11 (1982) 341–356.
[31] Z. Pawlak, A. Skowron, Rudiments of rough sets, Information Sciences 177 (2007) 3–27.
[32] Z. Pawlak, A. Skowron, Rough sets: some extensions, Information Sciences 177 (2007) 28–40.
[33] Z. Pawlak, A. Skowron, Rough sets and boolean reasoning, Information Sciences 177 (2007) 41–73.
[34] Y.H. Qian, J.Y. Liang, W. Pedrycz, C.Y. Dang, Positive approximation: an accelerator for attribute reduction in rough set theory, Artificial Intelligence 174 (2010) 597–618.
[35] A.M. Radzikowska, E.E. Kerre, A comparative study of fuzzy rough sets, Fuzzy Sets and Systems 126 (2002) 137–155.
[36] S. Ramaswamy, R. Rastogi, S. Kyuseok, Efficient algorithms for mining outliers from large data sets, in: Proceedings of ACM SIGMOD International Conference on Management of Data, vol. 29, 2000, pp. 427–438.
[39] M. Sarkar, Fuzzy-rough nearest neighbor algorithms in classification, Fuzzy Sets and Systems 158 (2007) 2134–2152.
[40] J. Shawe-Taylor, N. Cristianini (Eds.), Kernel Methods for Pattern Analysis, Cambridge University, 2004.
[41] G. Sheikholeslami, S. Chatterjee, A. Zhang, Wavecluster: a multi-resolution clustering approach for very large spatial databases, in: Proceedings of International Conference on Very Large Databases, New York, USA, 1998, pp. 428–439.
[42] Q. Shen, A. Chouchoulas, A rough-fuzzy approach for generating classification rules, Pattern Recognition 35 (2002) 2425–2438.
[43] A. Skowron, L. Polkowski (Eds.), Rough Sets in Knowledge Discovery, Berlin, Germany, Springer-Verlag, 1998, pp. 1–2.
[44] R. Slowinski (Ed.), Intelligent Decision Support: Handbook of Applications and Advances of the Rough Sets Theory, 1992.
[45] D.B. Stephen, S. Mark, Mining distance-based outliers in near linear time with randomization and a simple pruning rule, in: KDD'03: Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, New York, NY, USA, 2003, pp. 29–38.
[46] V. Suresh Babu, P. Viswanath, Weighted k-nearest leader classifier for large data sets, Pattern Recognition and Machine Intelligence, Lecture Notes in Computer Science, vol. 4815, Springer, Berlin, Heidelberg, 2007, pp. 17–24.
[47] J.S. Taylor, N. Cristianini, On the generalization of soft margin algorithms, IEEE Transactions on Information Theory 48 (2002) 2721–2735.
[49] G.D. Thomas, Approximate statistical tests for comparing supervised classification learning algorithms, Neural Computation 10 (1998) 1895–1923.

[50] V.N. Vapnik, Statistical Learning Theory, John Wiley and Sons, 1998.
[51] C.J. Veenman, M.J.T. Reinders, The nearest subclass classifier: a compromise between the nearest mean and nearest neighbor classifier, IEEE Transactions on Pattern Analysis and Machine Intelligence 27 (2005) 1417–1429.
[52] W.-Z. Wu, W.-X. Zhang, Constructive and axiomatic approaches of fuzzy approximation operators, Information Sciences 159 (2004) 233–254.
[53] W.-Z. Wu, J.-S. Mi, W.-X. Zhang, Generalized fuzzy rough sets, Information Sciences 151 (2003) 263–282.
[54] X.D. Wu, X.Q. Zhu, Mining with noise knowledge: error-aware data mining, IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans 38 (2008) 917–931.
[55] H. Xiong, G. Pandey, M. Steinbach, V. Kumar, Enhancing data analysis with noise removal, IEEE Transactions on Knowledge and Data Engineering 18 (2006) 304–319.
[56] D.S. Yeung, D.G. Chen, E.C.C. Tsang, J.W.T. Lee, X.Z. Wang, On the generalization of fuzzy rough sets, IEEE Transactions on Fuzzy Systems 13 (2005) 343–361.
[57] L.A. Zadeh, A computational approach to fuzzy quantifiers in natural languages, Computers and Mathematics with Applications 9 (1983) 149–184.
[58] S.Y. Zhao, W.W.Y. Ng, E.C.C. Tsang, Rule induction from numerical data based on rough sets theory, in: Proceedings of 2006 International Conference on Machine Learning and Cybernetics, vols. 1–7, 2006, pp. 2294–2299.
[59] S.Y. Zhao, E.C.C. Tsang, D.G. Chen, The model of fuzzy variable precision rough sets, IEEE Transactions on Fuzzy Systems 17 (2009) 451–467.
[60] X.Q. Zhu, X.D. Wu, Class noise handling for effective cost-sensitive learning by cost-guided iterative classification filtering, IEEE Transactions on Knowledge and Data Engineering 18 (2006) 1435–1440.
[61] X.Q. Zhu, X.D. Wu, Y. Yang, Error detection and impact-sensitive instance ranking in noisy data sets, in: Proceedings of 19th National Conference on Artificial, 2004, pp. 378–383.
[62] W.P. Ziarko, C.J.V. Rijsbergen (Eds.), Rough Sets, Fuzzy Sets, and Knowledge Discovery, Springer-Verlag, New York, USA, 1994.