Information Sciences 180 (2010) 4384–4400

Contents lists available at ScienceDirect

Information Sciences

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

Soft fuzzy rough sets for robust feature evaluation and selection

Qinghua Hu *, Shuang An, Daren Yu

Harbin Institute of Technology, Harbin 150001, PR China

ARTICLE INFO

Article history:
Received 29 June 2009
Received in revised form 1 June 2010
Accepted 19 July 2010

Keywords:
Fuzzy rough sets
Feature evaluation
Robust
Noise

ABSTRACT

The fuzzy dependency function proposed in the fuzzy rough set model is widely employed in feature evaluation and attribute reduction. It is shown that this function is not robust to noisy information in this paper. As datasets in real-world applications are usually contaminated by noise, robustness of data analysis models is very important in practice. In this work, we develop a new model of fuzzy rough sets, called soft fuzzy rough sets, which can reduce the influence of noise. We discuss the properties of the model and construct a new dependence function from the model. Then we use the function to evaluate and select features. The presented experimental results show the effectiveness of the new model.

1. Introduction

In classification learning, data are usually described with a great number of features. Typically, part of them are irrelevant or redundant with the classification task. These irrelevant features might confuse learning algorithms and deteriorate learning performance. Hence, it is useful to select relevant and indispensable features for designing classification systems.

So far, a number of algorithms have been developed for feature reduction [2,11,14,15,23,31]. Generally speaking, there are two key issues in constructing a feature selection algorithm: feature evaluation and search strategies. Feature evaluation is used to measure the quality of the candidate features. Obviously, evaluation functions have great influence on outputs of algorithms. A great number of functions were designed, such as dependency [45], neighborhood dependency [15] and fuzzy dependency in the rough set theory [19,39]; mutual information and symmetric uncertainty in information theory [2,23,31]; sample margin [57] and hypothesis margin [35,43] in statistical learning theory, and so on. As to the search strategy, it can be roughly divided into two categories. One guarantees to find the optimal subset of features in terms of the used evaluation function, such as the exhaustive search [25] and the branch-and-bound algorithm [28,40]. And the other is to find a suboptimal solution for efficiency, including sequential forward selection [21], sequential backward elimination [25], floating search [33,41], mRMR [31], etc.

The rough set theory provides a mathematical tool to handle uncertainty in data analysis [30]. It has been successfully used in attribute reduction and rule learning [32,45]. Moreover, this theory also provides practical solutions to many data analysis tasks, such as data mining [29] and rule discovery [32]. The classic rough set model is defined with equivalence relations, which leads to the limitation in handling data with numerical or fuzzy attributes, some generalized models were proposed, such as fuzzy rough sets [12,26,59] and neighborhood rough sets [15].

It is well known that datasets in real-world application are usually corrupted by noise [60,61]. The noisy samples may have great influence on outputs of the models. Accordingly, the performance of classification systems would be reduced. So robust models and algorithms are highly desirable in practice.

* Corresponding author.
E-mail addresses: huqinghua@hit.edu.cn (Q. Hu), yudaren@hit.edu.cn (D. Yu).

0020-0255/$ - see front matter © 2010 Elsevier Inc. All rights reserved.
doi:10.1016/j.ins.2010.07.010

In the framework of rough sets, dependency functions, defined as the ratio of the consistent samples over the universe, are used to compute the quality of features. This function plays the central role in rough set based learning algorithms. However, it is observed that the dependency function defined in Pawlak rough set model is not robust. This property is passed down to neighborhood rough sets and fuzzy rough sets [37,58,62], which limits the applications of these models.

In order to deal with this problem, some extended models were developed. First, Yao, Wong et al. proposed the decision-theoretic rough set model (DTRS) in 1990 [55] and applied this model to attribute reduction in 2008 [56]. This model considers the statistic information in data. In 1993, Ziarko developed the variable precision rough set model (VPRS) to tolerate noisy samples [62], where several mislabeled samples in an equivalence class are overlooked in computing lower and upper approximations. However, given a learning task, it is a big problem to set how many samples should be overlooked. In addition, information theory was also introduced to compute the significance of features [16,54]. These models are indeed more robust than rough sets, however, the granular structures are lost in these models. In [49], a comparative study between Pawlak's rough sets based reduction and the information-theoretic based reduction was conducted. Besides, Rolka et al. and Zhao, Tsang et al. showed the definitions of variable precision fuzzy rough sets [37] and fuzzy variable precision rough sets [58] to enhance robustness of fuzzy rough sets, respectively. Unfortunately, we find that the model in [58] is still sensitive to mislabeled samples. Although there are some models to deal with noise in datasets, it seems that handling noise is still an open problem in the rough set theory.

In intelligent data analysis, there are two ways to deal with noisy information. One is to remove noise in the step of data preprocessing, such as outlier detection [1,6,13,22,36,38,42], data cleaner [53] and impact-sensitive ranking [60]. And the other is to design robust algorithms, such as noise-tolerance feature selection [8,20], weighted k-Nearest Neighbor [44], Maxi–Min Margin Machine [18], robust minimax approach [24], Nearest Subclass Classifier [48], Cost-Sensitive Classification [61], Error-Aware Classification [52], robust clustering [4,10] and soft-margin SVM [5,46,47].

In recent years, soft-margin SVM becomes a popular and robust learning algorithm for classification modeling. As to hard-margin SVM, all the samples should be correctly classified with a margin, while soft-margin SVM allows some samples to be misclassified for obtaining a large-margin classifier by making tradeoff between margin and classification error. By this way, soft-margin SVM reduces the impact of noisy information on the final classifier.

In this work, we follow the idea of soft-margin SVM and introduce a robust rough set model, called soft fuzzy rough set. The classic fuzzy rough set model computes the membership of an object to a class with the nearest sample from different classes. However, this leads to the sensitivity to noisy samples. Our model improves the computation of approximations, where the membership is not calculated with the nearest sample from different classes, but the $k'th$ sample, where k is determined by tradeoff between the number of misclassified samples and the augmentation of membership. By this way, the proposed model is robust to the noisy samples. Some numerical experiments are conducted to test the robustness of the model in feature evaluation and selection.

The rest of the paper is organized as follows. Section 2 gives the basic notations of rough sets and analyzes the robustness of these models. Section 3 introduces the definition of soft fuzzy rough sets and discusses the properties of the model. Next, we define the soft fuzzy dependency and design a feature selection algorithm based on soft dependency in Section 4. And then we introduce some measures for evaluating robustness of algorithms in Section 5. Numerical experiments are presented in Section 6. Finally, the conclusions are given in Section 7.

2. Basic notations of rough sets and robustness analysis

$IS = \langle U, C\rangle$ is called an information table, where U is a finite and nonempty set of objects and C is a set of features used to characterize the objects. $\forall\ B \subseteq C$, a B-indiscernibility relation is defined as

$$IND(B) = \{(x,y) \in U^2 \quad |\forall a \in B,\ a(x) = a(y)\}. \tag{1}$$

Then the partition of U generated by $IND(B)$ is denoted by $U/IND(B)$ (or U/B). The equivalence class of x induced by B-indiscernible relation is denoted by $[x]_B$.

Given an arbitrary $X \subseteq U$, R is an equivalence relation on U induced by a set of attributes. The lower and upper approximations of X with respect to R are defined as

$$\begin{cases} \underline{R}X = \{x \in U | [x]_R \subseteq X\}, \\ \overline{R}X = \{x \in U | [x]_R \cap X \neq \phi\}. \end{cases} \tag{2}$$

$\mathrm{BN}_R(X) = \overline{R}X - \underline{R}X$ is called R-boundary region of X and $\mathrm{NEG}_R(X) = U - \overline{R}X$ is the R-negative region of X. The lower approximation is also called R-positive region of X, denoted by $\mathrm{POS}_R(X)$.

Given a decision table $DS = \langle U, C \cup D\rangle$, D is the decision attribute. For $\forall\ B \subseteq C$, the positive region of decision D on B, denoted by $\mathrm{POS}_B(D)$, is defined as

$$\mathrm{POS}_B(D) = \bigcup_{X \in U/D} \underline{B}X, \tag{3}$$

where U/D is the set of the equivalence classes generated by D. The dependency of decision D on B is defined as

$$\gamma_B(D) = \frac{|\mathrm{POS}_B(D)|}{|U|}. \tag{4}$$

Dependency is the ratio of the samples in the lower approximation over the universe. As the lower approximation is the set of objects with consistent decisions, dependency is used to measure the classification performance of attributes. It is expected that all the decisions of objects are consistent with respect to the given attributes. In practice, inconsistency widely exists in data.

The previous research shows that the lower and upper approximations in Pawlak's rough sets were sensitive to noise. According to the definition of lower approximations, the sample is grouped into lower approximation if all samples in its equivalence class consistently belong to a decision class. While the sample belongs to the upper approximation if one of the samples in its equivalence class comes from the decision class. Thus if there is one noisy sample, the whole equivalence class is grouped into the classification boundary. This leads to the sensitivity of dependency to noisy samples.

As to data with numerical-valued features, neighborhood relations and neighborhood rough sets are introduced [15]. Given a decision table $\langle U, C \cup D\rangle$, U is divided into N decision classes: $X_1, X_2, \ldots, X_N$. $\forall\ B \in C$, the neighborhood of sample x is defined as $\delta(x) = \{y | \Delta_B(x,y) \leqslant \delta, y \in U\}$, where Δ_B is a distance function defined in a feature space B. If sample y is contained by the neighborhood of x, we say y and x satisfy neighborhood relation N_B. We can see that neighborhood relation relaxes the equivalence relation to a similarity relation, and the similarity degree is characterized by distance functions. The lower and upper approximations of D in the neighborhood induced granular space are

$$\begin{cases} \underline{N_B}D = \{\underline{N_B}X_1, \underline{N_B}X_2, \ldots, \underline{N_B}X_N\}, \\ \overline{N_B}D = \{\overline{N_B}X_1, \overline{N_B}X_2, \ldots, \overline{N_B}X_N\}, \end{cases} \tag{5}$$

where $\underline{N_B}X = \{x_i | \delta_B(x_i) \subseteq X, x_i \in U\}$ and $\overline{N_B}X = \{x_i | \delta_B(x_i) \cap X \neq \phi, x_i \in U\}$. The neighborhood dependency of D on B is defined as

$$\gamma_B(D) = \frac{|\underline{N_B}D|}{|U|}. \tag{6}$$

Just like dependency in Pawlak's rough sets, neighborhood dependency is also sensitive to noisy samples. We can see if there is one sample with a different decision in the neighborhood of x_i, x_i should be grouped as classification boundary. In this sense, the lower approximation of neighborhood rough sets is sensitive to noise, which make neighborhood dependency is not robust to noise.

As to fuzzy cases, fuzzy rough sets were developed. 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: $R(x,y) \geqslant \mathrm{supmin}_{z\in U}\{R(x,z), R(z,y)\}$.

We say R is a fuzzy similarity relation. Fuzzy similarity relations are used to measure the similarity of the objects characterized with continuous features. The fuzzy similarity class $[x]_R$ associated with x and R is a fuzzy set, where $[x]_R(y) = R(x,y)$ for all $y \in U$. Fuzzy rough sets were first introduced by Dubois and Prade in [12] based on fuzzy similarity relations.

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

$$\begin{cases} \underline{R}A(x) = \inf\limits_{y\in U} \max\{1 - R(x,y), A(y)\}, \\ \overline{R}A(x) = \sup\limits_{y\in U} \min\{R(x,y), A(y)\}. \end{cases} \tag{7}$$

The approximation operators in (7) were studied in detail from the constructive and axiomatic approaches in [50,51]. In 1998, Morsi and Yakout replaced fuzzy equivalence relation with a T-equivalence relation and built an axiom system of the model [27]. In 2002, based on the negator operator δ and implicator operator θ, Radzikowska and Kerre defined fuzzy lower and upper approximations [34].

If A is a crisp set, then

$$A(y) = \begin{cases} 1, & y \in A, \\ 0, & y \notin A. \end{cases} \tag{8}$$

The fuzzy lower and upper approximations in (7) degenerate into the following formulae

$$\begin{cases} \underline{R}A(x) = \inf\limits_{y\in U-A} \{1 - R(x,y)\}, \\ \overline{R}A(x) = \sup\limits_{y\in A} R(x,y). \end{cases} \tag{9}$$

Considering the above definitions, we see that the membership of a sample $x \in U$ to the fuzzy lower approximation of A is the dissimilarity between x and the nearest sample $y \notin A$ and the membership of a sample $x \in U$ to the fuzzy upper approximation of A is the similarity between x and the nearest sample $y \in A$. If we take

$$R(x,y) = \exp\left(\frac{-\|x-y\|^2}{\delta}\right) \tag{10}$$

as a similarity function, then $1 - R(x,y)$ can be considered as a general distance function $d(x,y)$ between x and y. Then formula (9) can be expressed as

$$\begin{cases} \underline{R}A(x) = \inf\limits_{y \in U-A} \{d(x,y)\}, \\ \overline{R}A(x) = \sup\limits_{y \in A} \{1 - d(x,y)\} = 1 - \inf\limits_{y \in A} \{d(x,y)\}. \end{cases} \tag{11}$$

Fig. 1 shows a toy example. According to the above analysis, in Fig. 1, the membership of x to the fuzzy lower approximation of the class marked by squares is the distance between x and y_1. Unfortunately, y_1 is a noisy sample. If y_1 does not exist, the membership of x to the fuzzy lower approximation of the class equals to the distance between x and y_2. The membership of x to the fuzzy lower approximation of the class increases significantly in this case. However, if y_1 does exist, the memberships of the lower approximation of all samples marked by squares change. One noisy sample completely alters the lower approximation of a class. Correspondingly, the fuzzy dependency of D on feature subset B, defined as

$$\gamma_B(D) = \frac{\sum_{x \in U} \mathrm{POS}_B(D)(x)}{|U|} = \frac{\sum_{x \in U} (\sup\limits_{X \in U/D} \underline{B}(X)(x))}{|U|} \tag{12}$$

is sensitive to noise as well.

Zhao et al. [58] discussed the robustness of several rough set models, including VPRS [62] and VPFRS [37], generalized with a threshold from Pawlak rough sets, and they pointed out that all of them were sensitive to noise. Moreover, Zhao also referred to that it was difficult for VQRS [7] to design an attribute reduction method since the important property that monotonicity of approximation quality with features does not hold in this model. Then a robust model, called fuzzy variable precision rough sets (FVPRS), was developed [58]. For understandability, we describe the lower and upper approximations of FVPRS as

$$\begin{cases} \underline{R_\alpha}A(x) = \inf\limits_{A(y) \leqslant \alpha} \max(1 - R(x,y), \alpha) \wedge \inf\limits_{A(y) > \alpha} \max(1 - R(x,y), A(y)), \\ \overline{R_\alpha}A(x) = \sup\limits_{A(y) \geqslant 1-\alpha} \min(R(x,y), 1-\alpha) \vee \sup\limits_{A(y) < 1-\alpha} \min(R(x,y), A(y)). \end{cases} \tag{13}$$

In computing $\underline{R_\alpha}A(x)$, if $A(y) \leqslant \alpha$ ($y \in U$), then $A(y) = \alpha$. In other words, the samples with $A(y) < \alpha$ are overlooked. In computing $\overline{R_\alpha}A(x)$, if $A(y) \geqslant 1 - \alpha$ ($y \in U$), $A(y) = 1 - \alpha$ i.e. the samples with $A(y) > 1 - \alpha$ are overlooked. From the formulae of the lower and upper approximations we conclude that $\forall\, x \in U$, $\underline{R_\alpha}A(x) \geqslant \alpha$ by neglecting some samples that satisfy $A(y) \leqslant \alpha$. Similarly, $\forall\ x \in U, \overline{R_\alpha}A(x) \leqslant 1 - \alpha$ by neglecting some samples that satisfy $A(y) \geqslant 1 - \alpha$. Compared with fuzzy rough sets, $\underline{R_\alpha}A(x) \geqslant \underline{R}A(x)$ and $\overline{R_\alpha}A(x) \leqslant \overline{R}A(x)$.

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

$$\begin{cases} \underline{R_\alpha}A(x) = \inf\limits_{A(y)=0} \max\{1 - R(x,y), \alpha\}, \\ \overline{R_\alpha}A(x) = \sup\limits_{A(y)=1} \min\{R(x,y), 1-\alpha\}. \end{cases} \tag{14}$$

$\underline{R_\alpha}A(x) \geqslant \alpha$ and $\overline{R_\alpha}A(x) \leqslant 1 - \alpha$ still hold.

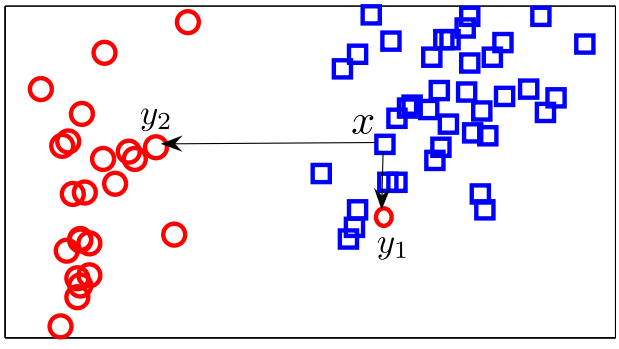

Fig. 1. The influence of noise on the membership of x to the fuzzy lower approximation of the class.

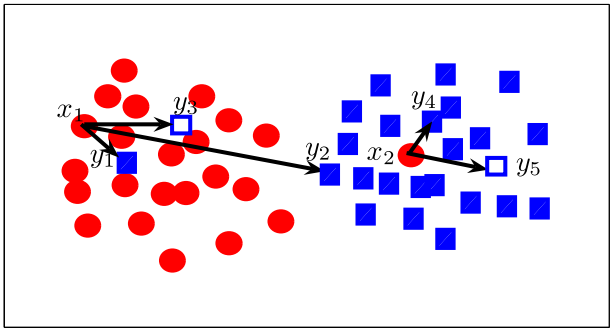

Fig. 2. The influence of noise on $\underline{R_\alpha}class_1(x)$.

However, the lower approximation of FVPRS is not robust to outliers, as shown in Fig. 2. x_i ($i = 1,2$) belongs to $class_1$ marked with balls and y_i ($i = 1,2,3,4,5$) comes from $class_2$ marked with squares, where y_3 and y_5 are imaginary samples. Here, we consider x_2 and y_1 as outliers. Suppose $\|x_1 - y_3\| = \|x_2 - y_5\| = \alpha$.

As to x_1,

$$\underline{R}class_1(x_1) = 1 - R(x_1, y_1) = \|x_1 - y_1\|,$$
$$\underline{R_\alpha}class_1(x_1) = 1 - R(x_1, y_3) = \|x_1 - y_3\| = \alpha.$$

$\underline{R}class_1(x_1) < \underline{R_\alpha}class_1(x_1)$. It seems that the lower approximation of FVPRS is more robust than fuzzy lower approximation. However, y_1 is a mislabeled sample. If we neglect y_1, the membership of x_1 to the fuzzy lower approximation of $class_1$ should be hu $\|x_1 - y_2\| > \alpha$.

As to x_2,

$$\underline{R}class_1(x_2) = 1 - R(x_2, y_4) = \|x_2 - y_4\|,$$
$$\underline{R_\alpha}class_1(x_2) = 1 - R(x_2, y_5) = \|x_2 - y_5\| = \alpha.$$

$\underline{R}class_1(x_2) < \underline{R_\alpha}class_1(x_2)$. That is to say we have to neglect some samples around x_2 to make $\underline{R_\alpha}class_1(x_2) \geqslant \alpha$. In fact the samples around x_2 should not be overlooked.

According to the above analysis, we see that the lower approximation of FVPRS is sensitive to mislabeled samples as well.

3. Soft fuzzy rough sets

Inspired by the idea of soft-margin SVM [9], we introduce a robust model of rough sets, named soft fuzzy rough sets. Soft-margin SVM is more robust than hard-margin SVM in classification. Hard-margin SVM finds the optimal classification hyperplane to make all the samples classified correctly with a margin. It is not applicable in many real-world problems where the data usually contain noise. And soft-margin SVM is to find an optimal classification hyperplane to make most samples classified correctly with a margin by neglecting a few samples. Soft-margin SVM is to find tradeoff between the size of margin and the classification error, which prevents the classifier overfitting noise.

First we introduce the definitions of hard distance and soft distance.

Definition 2. Given an object x and a set of objects Y, the hard distance between x and Y is defined as

$$HD(x, Y) = \min_{y \in Y} d(x, y), \tag{15}$$

where d is a distance function.

As we all know, the statistical minimum is sensitive to noise and not robust. We introduce a new definition of distance.

Definition 3. Given an object x and a set of objects Y, the soft distance between x and Y is defined as

$$SD(x, Y) = \arg_{d(x,y)} \sup_{y \in Y} \{d(x, y) - \beta m_Y\}, \tag{16}$$

where d is a distance function, β is a penalty factor and $m_Y = |\{y_i | d(x, y_i) < d(x, y)\}|$.

We explain the soft distance with Fig. 3. Sample x comes from $class1$ and the other samples are from $class2$, denoted by Y. Here, we suppose $d_1 < d_2 < d_3 < d_4$. We can see that $HD(x, Y)$ is d_1. 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; if y_2 is also taken as a noisy sample, $SD(x, Y)$ should be d_3. How many samples should be taken as noisy samples in this

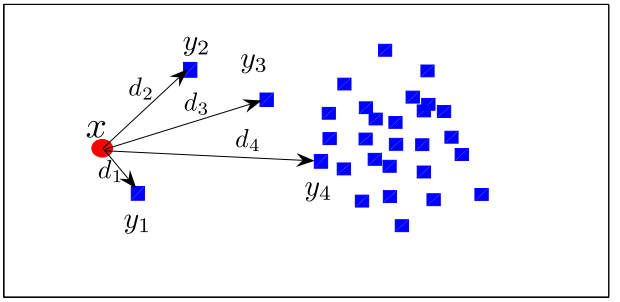

Fig. 3. Soft distance.

case? We here add a penalty term to the distance to solve the problem. If we overlook one noisy sample, $d(x,y)$ will be reduced by β. If $d(x,y') - \beta m'_Y$ ($y' \in Y$) is the largest of $d(x,y) - \beta m_Y$ ($\forall\ y \in Y$), this distance $d(x,y')$ is taken as the soft distance between x and Y.

Moreover, if β is larger than a certain value, the soft distance degenerates to the hard distance; and if β is smaller than a certain value, many samples would be overlooked. In other words, the larger β is, the less the noisy samples are neglected.

Next, we use an example to explain the definition of the soft distance.

Example. Given a set of objects $Y = \{y_1, y_2, y_3, y_4, y_5, y_6\}$ and sample x, $d(x,y_1) = 0.11$, $d(x,y_2) = 0.29$, $d(x,y_3) = 0.49$, $d(x,y_4) = 0.50$, $d(x,y_5) = 0.51$, $d(x,y_6) = 0.50$, $\beta = 0.06$. $HD(x,Y) = 0.11$, the soft distance $SD(x,Y)$ is

$$\begin{aligned} SD(x,Y) &= \arg_{d(x,y_i)} \max\{0.11, 0.29 - 0.06 \times 1, 0.49 - 0.06 \times 2, 0.50 - 0.06 \times 3, 0.51 - 0.06 \times 5\} \\ &= \arg_{d(x,y_i)} \max\{0.11, 0.23, 0.37, 0.32, 0.21\} = 0.49. \end{aligned}$$

Based on the soft distance we introduce a new model of fuzzy rough sets, named soft fuzzy rough sets. The new model is defined as follows.

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

$$\begin{cases} \underline{R^S}(A)(x) = 1 - R(x, \arg_y \sup_{A(y) \leqslant A(y_L)} \{1 - R(x,y) - \beta m_{Y_L}\}), \\ \overline{R^S}(A)(x) = R(x, \arg_y \inf_{A(y) \geqslant A(y_U)} \{R(x,y) + \beta n_{Y_U}\}), \end{cases} \tag{17}$$

where

$$\begin{cases} Y_L = \{y | A(y) \leqslant A(y_L), y \in U\},\ y_L = \arg_y \inf_{y \in U} \max\{1 - R(x,y), A(y)\}, \\ Y_U = \{y | A(y) \geqslant A(y_U), y \in U\},\ y_U = \arg_y \sup_{y \in U} \min\{R(x,y), A(y)\}. \end{cases} \tag{18}$$

$\beta$is a penalty factor, m_{Y_L} is the number of the samples overlooked in computing $\underline{R^S}(A)(x)$ and n_{Y_U} is the number of the samples overlooked in computing $\overline{R^S}(A)(x)$.

The essence of Definition 4 is to select two proper samples in U to compute $\underline{R^S}A(x)$ and $\overline{R^S}A(x)$, where the two samples satisfy $A(y) \leqslant A(y_L)$ and $A(y) \geqslant A(y_U)$, respectively. Fig. 4 illustrates this proposition. In the left figure, the two curves are $1 - R(x,y)$ and $A(y)$. According to the definition of fuzzy rough sets, $\underline{R}A(x) = 1 - R(x,y_L) = A(y_L)$. If y_L is a noisy sample, we

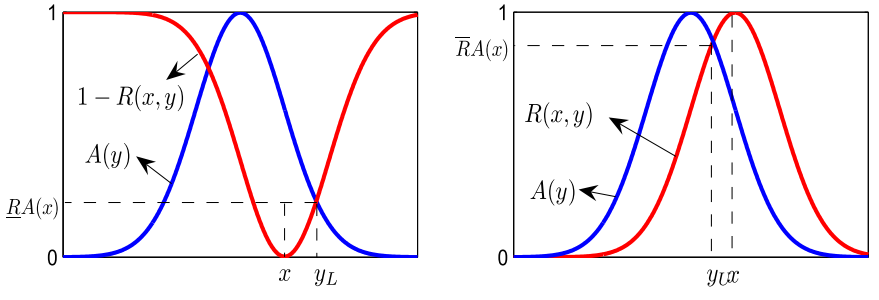

Fig. 4. Explanations of $A(y) \leqslant A(y_L)$ and $A(y) \geqslant A(y_U)$.

should use a sample that is farther away from x than y_L to compute $\underline{R^S}A(x)$. And such samples satisfy $A(y) \leqslant A(y_L)$. Similarly, if y_U is a noisy sample, we should use a sample satisfying $A(y) \geqslant A(y_U)$ to compute $\overline{R^S}A(x)$ (the right figure in Fig. 4).

Suppose A is a crisp set. The membership of x to the soft fuzzy lower approximation of A is

$$\underline{R^S}(A)(x) = 1 - R(x, y_{AL}), \tag{19}$$

where

$$y_{AL} = \arg_y \sup_{A(y)=0} \{1 - R(x,y) - \beta m_{Y_L}\} = \arg_y \sup_{A(y)=0} \{d(x,y) - \beta m_{Y_L}\} = \arg_y SD(x, U - A). \tag{20}$$

So $\underline{R^S}(A)(x)$ can be viewed as the soft distance from x to $U - A$.

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

$$\overline{R^S}(A)(x) = R(x, y_{AU}), \tag{21}$$

where

$$y_{AU} = \arg_y \inf_{A(y)=1} \{R(x,y) + \beta n_{Y_U}\} = \arg_y \sup_{A(y)=1} \{1 - R(x,y) - \beta n_{Y_U}\} = \arg_y \sup_{A(y)=1} \{d(x,y) - \beta n_{Y_U}\} = \arg_y SD(x, A). \tag{22}$$

Here $\overline{R^S}(A)(x)$ can be considered as the soft similarity between x and A.

Since the soft distance is more robust than the hard distance, soft fuzzy rough sets are more robust to noise than fuzzy rough sets.

Compared with Zhao's model, the advantage of our model is that it can automatically find optimal samples to compute the soft fuzzy memberships of the lower and upper approximations. In Fig. 2, FVPRS model lets $\underline{R_\alpha}class_1(x_1) = d(x_1, y_3) = \alpha$ because y_1 is a noisy sample, where α is subjectively set. And $\alpha = d(x_1, y_3)$ is much less than the real value $d(x_1, y_2)$. While our model can automatically find a balance between the memberships and the number of overlooked samples. If the enlargement $d(x_1, y_1) - d(x_1, y_2)$ of $\underline{R^S}class_1(x_1)$ is larger than the cost of misclassifying the sample y_1, the membership will be $d(x_1, y_2)$; otherwise, $\underline{R^S}class_1(x_1) = d(x_1, y_1)$.

Moreover, it is proven that soft fuzzy lower and upper approximations have the following properties.

Proposition 1. *For $\forall$ A, B $\in$ F(U), the following statements hold.*

$$(P11) \quad \underline{R^S}(A) \cap \underline{R^S}(B) = \underline{R^S}(A \cap B); \tag{23}$$

$$(P12) \quad \overline{R^S}(A) \cup \overline{R^S}(B) = \overline{R^S}(A \cup B). \tag{24}$$

Proof

(P11) $\forall\, x \in U$,

$$\underline{R^S}(A)(x) \wedge \underline{R^S}(B)(x) = (1 - R(x, \arg_{y_1} \sup_{A(y_1) \leqslant A(y_{L_A})} \{1 - R(x, y_1) - \beta m^1_{Y_L}\})) \wedge (1 - R(x, \arg_{y_2} \sup_{B(y_2) \leqslant B(y_{L_B})} \{1 - R(x, y_2) - \beta m^2_{Y_L}\}))$$

$$= 1 - R(x, \arg_y \sup_{(A\cap B)(y) \leqslant A(y_{L_A}) \wedge B(y_{L_B}) = (A\cap B)(y_{L_{(A\cap B)}})} \{1 - R(x,y) - \beta m_{Y_L}\}) = \underline{R^S}(A \cap B)(x).$$

Then $\underline{R^S}(A) \cap \underline{R^S}(B) = \underline{R^S}(A \cap B)$.

(P12) $\forall\, x \in U$,

$$\overline{R^S}(A)(x) \vee \overline{R^S}(B)(x) = R(x, \arg_{y_1} \inf_{A(y_1) \geqslant A(y_{U_A})} \{R(x, y_1) + \beta n^1_{Y_U}\}) \vee R(x, \arg_{y_2} \inf_{B(y_2) \geqslant B(y_{U_B})} \{R(x, y_2) + \beta n^2_{Y_U}\})$$

$$= R(x, \arg_y \inf_{(A\cup B)(y) \geqslant A(y_{U_A}) \vee B(y_{U_B}) = (A\cup B)(y_{U_{A\cup B}})} \{R(x,y) + \beta n_{Y_U}\}) = \overline{R^S}(A \cup B)(x).$$

Then $\overline{R^S}(A) \cup \overline{R^S}(B) = \overline{R^S}(A \cup B)$.

Fig. 5 illustrates (P11). A and B are two fuzzy sets. In terms of the definition of fuzzy rough sets, $\underline{R}A(x) = 1 - R(x, y_{L_A}) = A(y_{L_A}), \underline{R}B(x) = 1 - R(x, y_{L_B}) = A(y_{L_B})$ and $\underline{R}(A \bigcap B)(x) = (1 - R(x, y_{L_A})) \bigcap (1 - R(x, y_{L_B})) = A(y_{L_A})$. If y_{L_A} is a noisy sample, a sample y satisfying $(A \bigcap B)(y) < (A \bigcap B)(y_{L_A})$ will be used to compute $\underline{R^S}(A \bigcap B)(x)$. And the sample must be the sample that is used to compute the smaller one of $\underline{R^S}A(x)$ or $\underline{R^S}B(x)$. □

Proposition 2. *For $\forall A \in F$ (U), the following statements hold.*

$$(P21) \quad (\underline{R^S}(A))^c = \overline{R^S}(A^c); \tag{25}$$

$$(P22) \quad (\overline{R^S}(A))^c = \underline{R^S}(A^c). \tag{26}$$

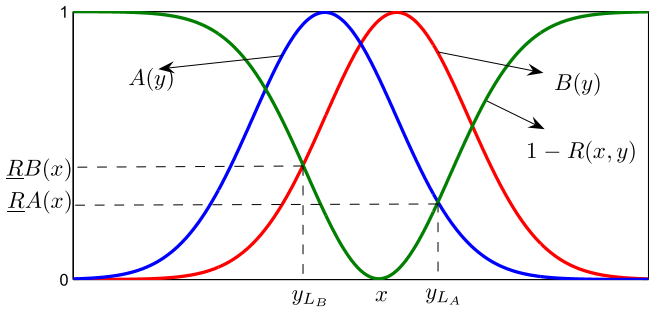

Fig. 5. $\underline{\underline{R}}^S(A) \cap \underline{R}^S(B) = \underline{R}^S(A \cap B)$.

Proof

(P21) $\forall\, x \in U$,

$$(\underline{R}^S(A)(x))^c = 1 - \underline{R}^S(A)(x) = R(x, \arg_y \sup_{A(y) \leqslant A(y_L)} \{1 - R(x,y) - \beta m_{Y_L}\}),$$

where

$$\arg_y \sup_{A(y) \leqslant A(y_L)} \{1 - R(x,y) - \beta m_{Y_L}\} = \arg_y \inf_{A(y) \leqslant A(y_L)} \{R(x,y) + \beta m_{Y_L}\} = \arg_y \inf_{A^c(y) \geqslant A^c(y_L)} \{R(x,y) + \beta m_{Y_L}\}.$$

Then

$$(\underline{R}^S(A)(x))^c = R(x, \arg_y \inf_{A^c(y) \geqslant A^c(y_L)} \{R(x,y) + \beta m_{Y_L}\}) = \overline{R^S}(A^c)(x).$$

(P22) $\forall\, x \in U$,

$$(\overline{R^S}(A)(x))^c = 1 - \overline{R^S}(A)(x) = 1 - R(x, \arg_y \inf_{A(y) \geqslant A(y_U)} \{R(x,y) + \beta n_{Y_U}\}),$$

where

$$\arg_y \inf_{A(y) \geqslant A(y_U)} \{R(x,y) + \beta n_{Y_U}\} = \arg_y \sup_{A(y) \geqslant A(y_U)} \{1 - R(x,y) - \beta n_{Y_U}\} = \arg_y \sup_{A^c(y) \leqslant A^c(y_U)} \{1 - R(x,y) - \beta n_{Y_U}\}.$$

Then

$$(\overline{R^S}(A)(x))^c = 1 - R(x, \arg_y \sup_{A^c(y) \leqslant A^c(y_U)} \{1 - R(x,y) - \beta n_{Y_U}\}) = \underline{R}^S(A^c)(x)$$

Therefore, $(\underline{R}^S(A))^c = \overline{R^S}(A^c)$ and $(\overline{R^S}(A))^c = \underline{R}^S(A^c)$ hold. □

4. Soft fuzzy dependency based feature selection

Definition 5. Given a decision table $\langle U, C \cup D \rangle$, U is a nonempty universe, C is the set of attributes and D is the decision attribute. $\forall\, B \in C$, the membership of an object $x \in U$ belonging to the soft positive region of D on B is defined as

$$\mathrm{POS}_B^S(D)(x) = \sup_{X \in U/D} \underline{B}^S(X)(x). \tag{27}$$

The soft fuzzy dependency of decision D on B is defined as

$$\gamma_B^S(D) = \frac{\sum_{x \in U} \mathrm{POS}_B^S(D)(x)}{|U|}. \tag{28}$$

Soft fuzzy dependency (SFD) can also be used to evaluate features. Section 3 illustrates that soft fuzzy lower approximation is robust to the mislabeled samples. We consider that soft fuzzy dependency is also robust to the mislabeled samples in feature evaluation.

Table 1
Feature selection algorithm.

Input:	X, F	X is a sample set and F is a feature set.
Output:	F'	F' is a feature ranking.
	Begin	
	Initialize	$F' = \phi$
	while	$F \neq \phi$
		Find $f = \arg_f \max_{f \in F}\{SFD_{(F' \cup \{f\})}(D)\}$,
		$F' = F' \cup \{f\}$,
		$F = F - \{f\}$
	End	
	Return	F'
	End	

Based on the soft fuzzy dependency we design a feature selection algorithm, shown in Table 1. The algorithm employs SFD as the feature evaluation function and the sequential forward selection as the search strategy. The output of the algorithm is a feature ranking $F' = \{f'_1, f'_2, \ldots, f'_{|F'|}\}$. Given the set F'_{k-1} with $k-1$ features selected, the $k'th$ feature is determined by

$$\max_{f \in F - F'_{k-1}} \{SFD_{(F'_{k-1} \cup \{f\})}(D)\}. \tag{29}$$

With the ranking, we can get n feature subsets $F'_1 = \{f'_1\},\ F'_2 = \{f'_1, f'_2\}, \ldots,\ F'_{|F'|} = \{f'_1, f'_2, \ldots, f'_{|F'|}\}$. Next, we use KNN classifier to cross-validate the classification accuracy of the data with these feature subsets. The feature subset with the highest classification accuracy is the final feature subset. In this work, we use this algorithm to validate the robustness of soft fuzzy dependency in Section 6.3.

5. Robustness evaluation

We wish that the feature quality computed with an evaluation function evaluation does not vary much if the samples are corrupted by a little noise. We take the robustness of measures as the similarity between the evaluation results computed with raw data and noisy data. Intuitively, the larger the similarity is, the more robust the evaluation function is.

In this work, we generate some noisy datasets from the given sets. The noisy datasets are generated as follows. Take the raw data containing n samples and m features as an example. Firstly, we randomly draw $3i\%$ ($i = 1, \ldots, k$) samples, and then give them labels that are distinct from their original labels. We get i-level noisy datasets for the raw dataset.

Assume $W = \{w_1, w_2, \ldots, w_m\}$ and $W' = \{w'_1, w'_2, \ldots, w_{m'}\}$ are the significance vectors of features computed with the raw data and the noisy data, respectively, where w_i ($i = 1,2,\ldots,m$) and w'_i ($i = 1,2,\ldots,m$) are the significance values of the i'th feature with the raw data and the noisy data. To compute similarity between W and W', we use Pearson's correlation coefficient

$$S_w(W, W') = \frac{\sum_{i=1}^{m}(w_i - \overline{W})(w'_i - \overline{W'})}{\left[\sum_{i=1}^{m}(w_i - \overline{W})^2 \sum_{i=1}^{m}(w'_i - \overline{W'})^2\right]^{1/2}}, \tag{30}$$

where $S_w(W, W')$ takes values in [-1,1]. The larger the value of $S_w(W, W')$ is, the larger the similarity is. $S_w(W, W') = 1$ means that W and W' are perfectly linearly correlated. $S_w(W, W') = 0$ means there is no linear correlation between W and W'. $S_w(W, W') = -1$ means W and W' are inverse-correlated.

As there are k evaluation results we compute the similarity between a pair of evaluations, and then we get a similarity matrix

$$\mathbf{S} = \begin{pmatrix} s_{11} & s_{12} & \cdots & s_{1k} \\ s_{21} & s_{22} & \cdots & s_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ s_{k1} & s_{k2} & \cdots & s_{kk} \end{pmatrix}, \tag{31}$$

where s_{ij} is the similarity between the i'th and j'th evaluation results.

In order to measure the similarity of all the evaluation results, in [17], the authors computed the similarity matrix with

$$TS = -\frac{1}{k}\sum_{j=1}^{k} \log \sum_{i=1}^{k} \frac{s_{ij}}{k}, \tag{32}$$

where $TS \in [0, \log k]$. If $\forall\ i, j,\ s_{ij} = 1$, which means the k evaluation results are the same, then $TS = 0$. In this case, the feature evaluation measure is robust. If $\forall\ i \neq j,\ s_{ij} = 0$, S is an identity matrix, then $TS = \log k$. We consider the similarity matrix is not stable and the measure is not robust. In Section 6.1, we use TS to measure the robustness of feature evaluation measures.

6. Experimental analysis

In this section, we first discuss the role of parameter β with an experiment. And then the robustness on soft fuzzy dependency is validated from two aspects. One is to validate the robustness on soft fuzzy dependency in evaluating a single feature using the methods described in Section 5, and the other one is to validate the robustness of soft fuzzy dependency in feature selection. The experiments are performed on nine data sets collected from UCI [3]. The summaries of the data sets are given in Table 2.

6.1. Parameter β

The parameter β in the definition of the soft distance is used to make tradeoff in computing the soft distance. If β is too small, more samples would be overlooked in computing soft distance. And if β is too large, the soft distance will degenerate to the hard distance. We give an experiment to show the relationship between β and soft distance. Data wine is used here.

The experiment is to compute the average soft distance of all the samples and the corresponding number of the overlooked samples in computing the soft distance. The results are shown in Fig. 6.

The first figure illustrates the relationship between the average soft distance of a sample and the value of β. It shows the average soft distance decreases if the value of β increases and it converges to a certain value. The second plot illustrates the relationship between the average number of the overlooked samples and the value of β. We can see that the average number of the overlooked samples also decreases along with the increasing of the value of β and converges to zero.

As we do not expect the soft distance just increases a little after many samples are ignored, we set that one sample is not ignored until the increment of soft distance is not less than β. For example, $\beta = 0.1$ means if the soft distance increases 0.1, we overlook one sample at most. In this work, we set $\beta = 0.1$. The subsequent experiments show that 0.1 is a good choice for β.

Table 2
Summaries of data sets.

Data	Samples	Features	Classes
Wine	178	13	3
WDBC	569	30	2
Sonar	208	60	2
Ionosphere	351	34	2
Glass	210	13	6
Musk	467	166	2
Pima-indians-diabetes	768	8	2
Shutter-trn	43500	10	5
Yeast	1484	7	10

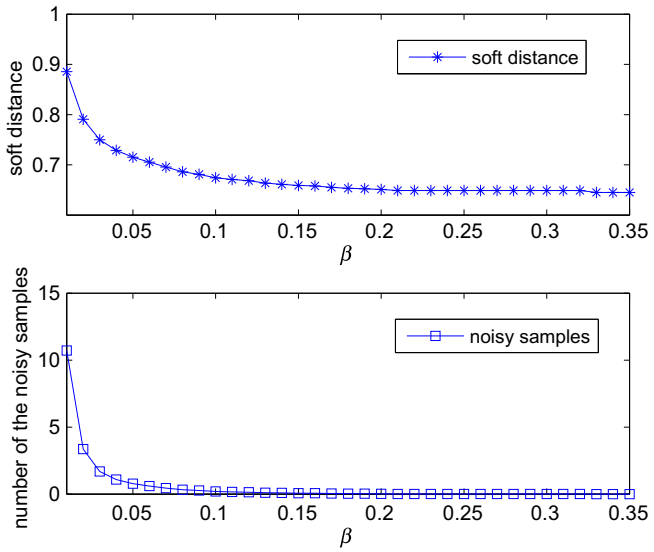

Fig. 6. Variation of soft distance and number of overlooked samples with β.

6.2. Robustness on soft fuzzy dependency for evaluating a single feature

We test the robustness of the soft fuzzy dependency in evaluating a single feature in this section. Here, we call the data sets downloaded from UCI raw datasets. With a raw dataset we can generate k noisy data sets using the method referred to in Section 5. In this work, $k = 10$ and the maximal noise level is 30%.

Firstly, we compute the soft fuzzy dependency of decision on each feature with the raw data sets and ten noisy data sets. Here, we take formula (10) as similarity function ($\delta = 0.15$ and $\|\cdot\|$ is 2-norm.) and let $\beta = 0.1$. Then we get a raw soft fuzzy dependency value (computed with a raw data set) and ten noisy soft fuzzy dependency values (computed with ten noisy data sets) of decision on each feature. Similarly, we also can compute a raw fuzzy dependency value and ten noisy fuzzy dependency values of decision on each feature. Fig. 7 shows the eleven evaluation results of decision on each feature.

For each subgraph, x axis is the feature index and y axis is the values of soft fuzzy dependency or fuzzy dependency. Eleven curves show eleven dependency values computed with a raw data set and ten noisy data sets. As to glass, we can see that the eleven curves of soft fuzzy dependency (SFD) are more similar than that of fuzzy dependency (FD). Notice that the dependency values are unitary. As Section 5 refers to, the larger the similarity is, the more robust the feature evaluation measure is. Thereby, soft fuzzy dependency function is more robust than fuzzy dependency on glass. As to musk, yeast and ionosphere, we can get the same conclusion.

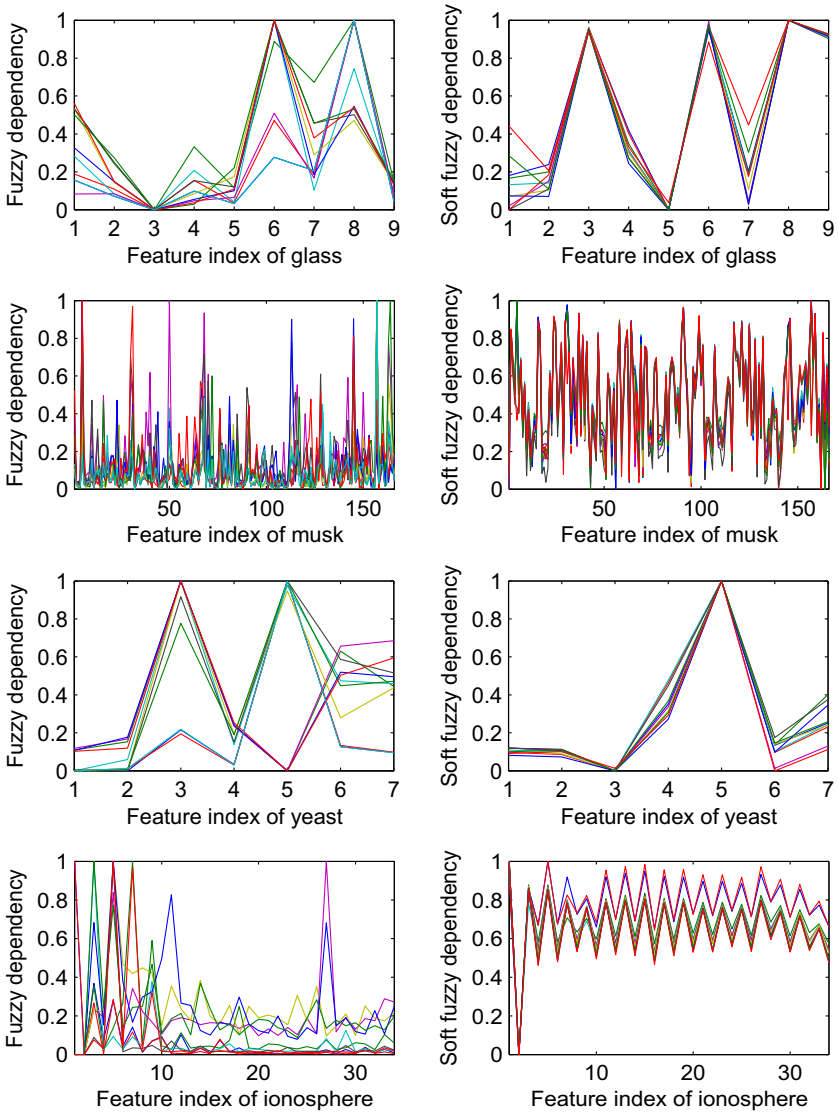

Fig. 7. Comparison of fuzzy dependency and soft fuzzy dependency.

Table 3
Similarity between raw fuzzy dependency and noisy fuzzy dependency.

Data	6%	12%	18%	24%	30%
Wine	0.98	0.74	0.39	0.05	0.34
WDBC	0.89	0.67	0.69	−0.30	−0.33
Sonar	0.79	0.77	0.55	0.57	0.37
Ionosphere	0.32	0.46	0.22	0.31	0.42
Glass	0.80	0.67	0.42	0.52	0.48
Musk	0.97	0.87	0.77	0.30	0.28
Pima-indians-diabetes	0.97	0.98	0.97	0.86	0.98
Shutter-trn	−0.15	0.02	0.08	0.99	0.81
Yeast	−0.26	0.72	0.72	−0.27	−0.26
Average	0.59	0.65	0.53	0.33	0.34

Next, we use Pearson's correlation coefficient to compute the similarity between a raw dependency and a noisy dependency, and compare the robustness of SFD with FD and neighborhood dependency (ND) by the values. Here, the raw dependency denotes a vector composed of the dependency values of decision on each feature, where the values are computed with a raw data set. And the values of the noisy dependency are computed with a noisy data set. The results, calculated with formula (30), are shown in Tables 3–5, where 6%, 12%, 18%, 24% and 30% are noise levels.

Table 3 shows the similarity between raw fuzzy dependency and noisy fuzzy dependency, Table 4 shows the similarity between raw neighborhood dependency and noisy neighborhood dependency and Table 5 shows the similarity between raw soft fuzzy dependency and noisy soft fuzzy dependency. The three tables show that the average similarity values of soft fuzzy dependency are the largest i.e. soft fuzzy dependency is more robust than other two dependency functions.

Moreover, we use *TS* to measure the robustness of FD, ND and SFD. The results are shown in Table 6. The average evaluation values of FD and ND are 0.58, while the average value of SFD is 0.20. As we know, the smaller the value is, the more robust the measure is. Consequently, soft fuzzy dependency is more robust than fuzzy dependency and neighborhood dependency.

6.3. Robustness of soft fuzzy dependency in feature selection

Next, we test the robustness on soft fuzzy dependency for feature selection with the algorithm in Table 1.

Table 4
Similarity between raw neighborhood dependency and noisy neighborhood dependency.

Data	6%	12%	18%	24%	30%
Wine	0.92	0.02	0.40	−0.17	0.00
WDBC	0.84	−0.11	0.63	−0.48	−0.26
Sonar	0.98	0.84	0.61	0.55	0.36
Ionosphere	0.30	0.37	0.00	0.40	0.00
Glass	0.20	0.12	0.04	0.22	0.16
Musk	0.85	0.27	0.21	0.05	0.07
Pima-indians-diabetes	0.77	0.77	0.77	0.00	0.77
Shutter-trn	0.98	−0.33	−0.33	−0.33	−0.33
Yeast	0.63	1.00	0.63	0.13	0.13
Average	0.71	0.32	0.32	0.04	0.10

Table 5
Similarity between raw soft fuzzy dependency and noisy soft fuzzy dependency.

Data	6%	12%	18%	24%	30%
Wine	0.89	0.81	0.47	0.77	0.71
WDBC	0.69	0.73	0.80	0.74	0.69
Sonar	0.87	0.96	0.49	0.68	0.50
Ionosphere	0.96	0.96	0.95	0.93	0.93
Glass	0.99	0.99	0.99	0.98	0.98
Musk	0.96	0.95	0.98	0.95	0.88
Pima-indians-diabetes	0.99	0.99	0.96	0.96	0.95
Shutter-trn	0.99	0.96	0.88	0.69	−0.03
Yeast	0.98	0.98	0.99	0.98	0.98
Average	0.92	0.92	0.83	0.85	0.73

Table 6
Robustness comparison of soft fuzzy dependency, fuzzy dependency and neighborhood dependency.

Data	FD	ND	SFD
Wine	0.56	0.67	0.07
WDBC	0.74	0.69	0.07
Sonar	0.77	0.88	0.71
Ionosphere	0.71	0.81	0.40
Glass	0.22	0.56	0.01
Musk	0.56	0.80	0.00
Pima-indians-diabetes	0.56	0.24	0.02
Shutter-trn	0.50	0.23	0.33
Yeast	0.61	0.34	0.00
Synthetic	100	13	2
Average	0.58	0.58	0.20

Table 7
Numbers of features selected and classification accuracies (%) on real-world data.

Data	FD		ND		SFD	
	n	RawAcc	n	RawAcc	n	RawAcc
Wine	7	95.9 ± 4.2	6	95.4 ± 5.3	6	97.1 ± 4.6
WDBC	7	89.9 ± 5.2	8	93.7 ± 2.6	8	94.7 ± 2.0
Sonar	8	81.2 ± 8.8	7	77.0 ± 9.5	10	84.9 ± 6.6
Ionosphere	10	94.2 ± 2.4	6	90.3 ± 5.6	9	91.0 ± 3.8
Glass	9	66.3 ± 8.0	6	68.5 ± 8.1	6	69.3 ± 7.7
Pima-indians-diabetes	8	70.8 ± 3.7	8	70.8 ± 3.7	4	71.6 ± 3.5
Average	8	83.1	7	82.6	7	84.8

6.3.1. Real-world data

Firstly, we select features with the algorithm in Table 1 with a real-world data set. Next, we use KNN ($k = 3$) classifier to cross-validate the classification accuracy of the data set with the feature subsets F'_m ($m = 1,2,\ldots,|F|$) composed of the first m features in the ranking. The feature subset with the highest classification accuracy is the final feature subset. Then we replace SFD with FD and ND, respectively, and select features with the same method.

The number of features selected and classification accuracies are shown in Table 7, where n is the number of features selected and RawAcc is classification accuracy. It is shown that, with SFD as the feature evaluation, the feature subsets selected can produce higher classification accuracies. In this work, we use the classification accuracies of feature subsets to evaluate the robustness of measures. The higher the classification accuracy is, the stronger the robustness of the measure is. Therefore, SFD is more robust than FD and ND.

6.3.2. Synthetic data

We conduct the proposed feature selection algorithm on the noisy synthetic data. We generate a set of data with 100 samples and 13 features. In addition, ten noisy data sets are generated from the synthetic data, where the noise levels are $i\%$ ($i = 1,2,\ldots,10$), respectively. With these set of data we perform select features using SFD, FD and ND as feature evaluation functions, respectively. The best four features are shown in Table 8.

Table 8
First four features of feature rankings.

Noise levels (%)	FD	ND	SFD
0	7, 12, 11, 8	7, 12, 11, 6	7, 12, 11, 13
1	7, 5, 3, 13	7, 5, 1, 3	7, 12, 11, 13
2	13, 9, 7, 8	13, 9, 6, 2	7, 12, 11, 13
3	11, 2, 12, 5	11, 2, 8, 1	7, 12, 11, 13
4	13, 8, 11, 5	13, 8, 9, 10	7, 12, 11, 13
5	13, 1, 12, 5	7, 1, 3, 10	7, 12, 11, 13
6	13, 2, 12, 3	13, 2, 8, 4	7, 12, 13, 11
7	12, 3, 13, 1	7, 3, 1, 5	7, 12, 11, 13
8	7, 3, 1, 5	7, 1, 8, 5	7, 12, 11, 8
9	7, 5, 1, 8	8, 6, 5, 13	7, 12, 11, 13
10	13, 9, 2, 10	13, 9, 10, 8	7, 2, 12, 10

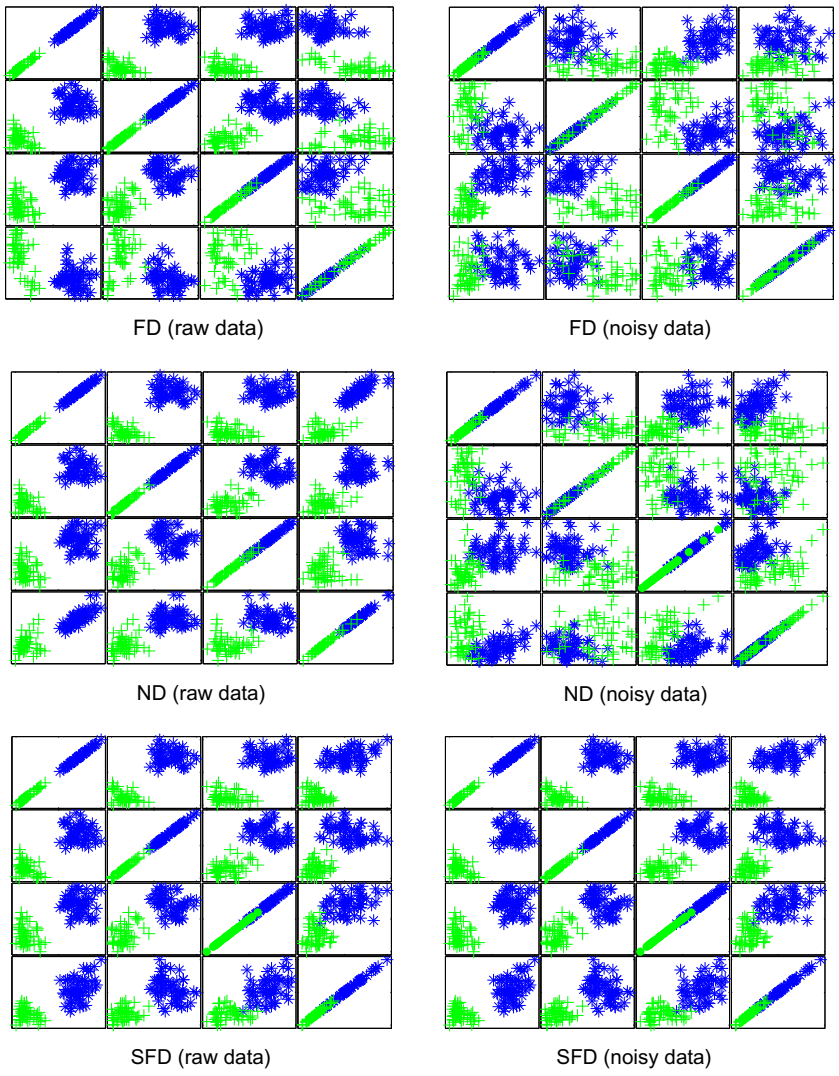

Fig. 8. Classification performance comparison.

It is shown that, with FD or ND as feature evaluation functions, the best four features are different from each other if the noise levels are different, while those with SFD are almost the same. Take the data containing 4% noise as an example. With the noisy data the best four features are {13,8,11,5}, {13,8,9,10} and {7,12,11,13} using FD, ND and SFD, respectively. With the raw data the feature subsets are {7,12,11,8}, {7,12,11,6} and {7,12,11,13}, respectively. Fig. 8 gives the comparison of classification performance.

For "FD (raw data)", "ND (raw data)" and "SFD (raw data)", the sixteen subfigures illustrate the two-dimensional distribution of the synthetic data with the first four features selected, where the features are selected with raw data and the feature evaluation are fuzzy dependency, neighborhood dependency and soft fuzzy dependency, respectively. While, for "FD (noisy data)", "ND (noisy data)" and "SFD (noisy data)", the features are selected with noisy data. We can see the classification performance of "SFD (noisy data)" is as well as "SFD (noisy data)" while "FD (noisy data)" and "ND (noisy data)" are worse than "FD (raw data)" and "ND (raw data)", respectively. That is to say noise has a great influence on the algorithms taking FD and ND as the feature evaluation. But the algorithm using SFD is more robust to noise.

The numbers of selected features and classification accuracies are shown in Table 9, where n is the number of features selected, Raw is the classification accuracy of raw data (synthetic data) (That is to say the features, selected with raw data, is used to classify the raw data) and Noisy is the classification accuracy of noisy data (That is to say the features, selected with noisy data, is used to classify the noisy data). We can see that if we use FD and ND the numbers of selected features are influenced by noise. But the numbers do not vary if SFD is used.

Table 9
Numbers of features selected and classification accuracies (%).

Noise levels (%)	FD				ND				SFD			
	Raw		Noisy		Raw		Noisy		Raw		Noisy	
	n	Acc	*n*	Acc	*n*	Acc	*n*	Acc	*n*	Acc	*n*	Acc
0	1	100	1	100	1	100	1	100	1	100	1	100
1	1	100	1	99	1	100	1	99	1	100	1	99
2	3	100	3	98	11	100	11	98	1	100	1	98
3	3	100	3	97	9	100	9	97	1	100	1	97
4	6	100	6	96	11	100	11	96	1	100	1	96
5	3	100	3	95	1	100	1	95	1	100	1	95
6	5	100	3	94	9	100	9	95	1	100	1	95
7	3	100	3	94	1	100	1	94	1	100	1	94
8	1	100	1	93	1	100	1	93	1	100	1	93
9	1	100	1	92	1	100	1	92	1	100	1	92
10	10	100	4	91	11	100	3	91	1	100	1	91

Table 10
Numbers of selected features and classification accuracies (%) on wine.

Noise levels (%)	FD		ND		SFD	
	n	RawAcc	*n*	RawAcc	*n*	RawAcc
3	3	95.0 ± 4.0	6	97.7 ± 4.2	5	97.2 ± 3.2
6	3	93.2 ± 5.3	6	93.8 ± 7.8	5	97.2 ± 3.2
9	6	97.1 ± 4.3	7	94.1 ± 4.1	6	97.7 ± 4.4
12	5	96.5 ± 4.2	5	92.7 ± 5.3	7	96.5 ± 3.8
15	9	97.2 ± 3.0	7	96.6 ± 5.2	6	96.6 ± 4.3
18	8	96.5 ± 5.6	8	96.6 ± 6.2	7	98.3 ± 3.0
21	7	95.4 ± 4.6	6	93.8 ± 6.8	6	96.6 ± 4.1
24	10	96.6 ± 3.0	4	91.6 ± 5.2	7	97.7 ± 4.7
27	9	97.1 ± 3.0	7	96.6 ± 5.0	6	93.4 ± 6.7
30	10	95.9 ± 5.8	8	97.1 ± 6.6	8	97.6 ± 6.1
Average	7	96.3	6	95.4	6	97.1

6.3.3. Noisy data created from real-world data

In this section, the classification accuracies are computed as follows. First, we select features with the noisy datasets and obtain feature ranking. Next, we use KNN ($k = 3$) classifier to compute the classification accuracy of the raw data with the selected feature subsets composed of the first m ($m = 1, 2, \ldots, |F|$) features in the ranking. The feature subset with the highest classification accuracy is output as the final feature subset. Take wine and sonar data as examples. In this work, we suppose there is no noise in the raw datasets of wine and sonar.

The numbers of the selected features and the corresponding performance are displayed in Tables 10 and 11, where n is the number of features selected and RawAcc is classification accuracy on raw data (wine and sonar). It is clear that the feature subsets selected, where SFD is taken as the measure, can produce higher classification accuracies. It shows that soft fuzzy dependency is more robust than fuzzy dependency and neighborhood dependency.

Table 11
Numbers of selected features and classification accuracies (%) on sonar.

Noise levels (%)	FD		ND		SFD	
	n	RawAcc	*n*	RawAcc	*n*	RawAcc
3	7	83.7 ± 4.5	6	80.3 ± 12.2	5	85.6 ± 8.1
6	7	78.4 ± 8.1	7	78.4 ± 10.3	15	85.6 ± 8.1
9	3	85.1 ± 8.9	6	80.7 ± 8.3	5	83.6 ± 6.0
12	9	77.4 ± 7.9	7	75.5 ± 6.9	17	87.5 ± 5.9
15	10	83.0 ± 9.8	7	81.1 ± 8.3	8	85.1 ± 4.9
18	11	85.6 ± 5.1	7	78.4 ± 10.1	9	84.6 ± 6.9
21	6	82.7 ± 11.2	6	76.4 ± 8.9	19	84.1 ± 4.8
24	12	77.4 ± 11.3	6	68.3 ± 7.5	11	83.1 ± 5.2
27	11	75.0 ± 11.6	7	66.8 ± 9.1	9	85.2 ± 10.2
30	6	80.6 ± 11.4	7	79.8 ± 10.8	5	82.2 ± 9.5
Average	8	81.2	7	77.0	10	84.9

7. Conclusions

Feature selection plays an important role in pattern classification systems. Feature evaluation functions used to compute the quality of features is a key issue in feature selection. In the rough set theory, dependency and fuzzy dependency has been successfully used to evaluate features. However, we find these function are not robust. In practice, data are usually corrupted by noise. So it is desirable to design robust models of rough sets.

Inspired by the idea of soft-margin SVM, we introduce a robust model of rough sets called soft fuzzy rough sets. The new model can reduce the influence of noise on the computation of soft fuzzy lower and upper approximations by overlooking some samples which are considered as noisy samples. In computing the membership of a sample to the soft fuzzy approximations, we make a tradeoff between the memberships and the number of the samples overlooked. Moreover, we discuss the properties of the new model. Then with the soft fuzzy lower approximation we give the definition of the soft fuzzy dependency and use it to evaluate features in feature selection. Finally, we test the robustness on soft fuzzy dependency function for feature evaluation and selection. The experimental results show that the soft fuzzy dependency function is effective in dealing with noisy data.

Acknowledgments

The authors would like to express their gratitude to the anonymous reviewers and Prof. Witold Pedrycz for their valuable comments. This work is partly 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, 2002, pp. 15–26.
[2] R. Battiti, Using mutual information for selecting features in supervised neural net learning, IEEE Transactions on Neural Networks 5 (1994) 531–549.
[3] C.L. Blake, C.J. Merz, UCI Repository of Machine Learning Databases, 1998. Available from: <http://www.ics.uci.edu/mlearn/MLRepository.html>.
[4] S. Chatzis, T. Varvarigou, Factor analysis latent subspace modeling and robust fuzzy clustering using *t*-distributions, IEEE Transactions on Fuzzy Systems 17 (2009) 505–516.
[5] 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, in: Rough Sets, Fuzzy Sets, Data Mining and Granular Computing, Lecture Notes in Artificial Intelligence, vol. 4482, Springer-Verlag, Berlin, Heidelberg, 2007, pp. 87–94.
[8] C. Cornelis, R. Jensen, A noise-tolerant approach to fuzzy-rough feature selection, in: Proceedings of the 17th International Conference on Fuzzy Systems, 2008, pp. 1598–1605.
[9] C. Cortes, V. Vapnik, Support-vector networks, Machine Learning 20 (1995) 273–297.
[10] R.N. Dave, S. Sen, Robust fuzzy clustering of relational data, IEEE Transactions on Fuzzy Systems 10 (2002) 713–727.
[11] A.B. David, H. Wang, A formalism for relevance and its application in feature subset selection, Machine Learning 41 (2000) 175–195.
[12] D. Dubois, H. Prade, Rough fuzzy sets and fuzzy rough sets, General Systems 17 (1990) 191–209.
[13] 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 the Second International Conference on Knowledge Discovery and Data Mining, 1996, pp. 226–231.
[14] M.A. Hall, Correlation-based feature selection for discrete and numeric class machine learning, in: Proceedings of the 17th International Conference on Machine Learning, 2000, pp. 359–366.
[15] Q.H. Hu, D.R. Yu, J.F. Liu, C.X. Wu, Neighborhood rough set based heterogeneous feature subset selection, Information Sciences 178 (2008) 3577–3594.
[16] Q.H. Hu, Z.X. Xie, D.Y. Yu, Hybrid attribute reduction based on a novel fuzzy-rough model and information granulation, Pattern Recognition 40 (2007) 3509–3521.
[17] Q.H. Hu, J.F. Liu, D.R. Yu, Stability analysis on rough set based feature evaluation, in: Rough Sets and Knowledge Technology, Lecture Notes in Computer Science, vol. 5009, Springer, Berlin, Heidelberg, 2008, pp. 88–96.
[18] 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.
[19] R. Jensen, Q. Shen, Fuzzy-rough sets for descriptive dimensionality reduction, in: IEEE International Conference on Fuzzy Systems, 2002, pp. 29–34.
[20] R. Jensen, Q. Shen, New approaches to fuzzy-rough feature selection, IEEE Transactions on Fuzzy Systems 17 (2009) 824–838.
[21] L.J. Ke, Z.R. Feng, Z.G. Ren, An efficient ant colony optimization approach to attribute reduction in rough set theory, Pattern Recognition Letters 29 (2008) 1351–1357.
[22] E.M. Knorr, R.T. Ng, V. Tucakov, Distance-based outliers: algorithms and applications, Very Large Databases 8 (2000) 237–253.
[23] N. Kwak, C.H. Choi, Input feature selection by mutual information based on Parzen window, IEEE Transactions on Pattern Analysis and Machine Intelligence 24 (2002) 1667–1671.
[24] 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] H. Liu, H. Motoda (Eds.), Feature Selection for Knowledge Discovery and Data Mining, Kluwer Academic Publishers., Boston, 1998.
[26] J.-S. Mi, Y. Leung, H.-Y. Zhao, T. Feng, Generalized fuzzy rough sets determined by a triangular norm, Information Sciences 178 (2008) 3203–3213.
[27] N.N. Morsi, M.M. Yakout, Axiomatics for fuzzy rough sets, Fuzzy Sets and Systems 100 (1998) 327–342.
[28] P. Narendra, K. Fukunaga, A branch and bound algorithm for feature subset selection, IEEE Transactions on Computers 26 (1977) 917–922.
[29] S.K. Pal, Soft data mining, computational theory of perceptions, and rough-fuzzy approach, Information Sciences 163 (2004) 5–12.
[30] Z. Pawlak, Rough sets, International Journal of Computer and Information Sciences 11 (1982) 341–356.
[31] H.C. Peng, F.H. Long, C. Ding, Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy, IEEE Transactions on Pattern Analysis and Machine Intelligence 27 (2005) 1226–1238.
[32] Studies in Fuzziness and Soft ComputingL. Polkowski, A. Skowron (Eds.), Rough Sets in Knowledge Discovery: Applications, Case Studies, and Software Systems, vol. 19, Physica-Verlag, Heidelberg, New York, 1998.

[33] P. Pudil, J. Novovicova, J. Kittler, Floating search methods in feature selection, Pattern Recognition Letters 15 (1994) 1119–1125.
[34] A.M. Radzikowskaa, E.E. Kerreb, A comparative study off fuzzy rough sets, Fuzzy Sets and Systems 126 (2002) 137–155.
[35] G.-B. Ran, N. Amir, T. Naftali, Margin based feature selection-theory and algorithms, in: ACM International Conference Proceeding Series, Proceedings of the 21st International Conference on Machine Learning, vol. 69, 2004.
[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.
[37] A.M. Rolka, L. Rolka, Variable precision fuzzy rough sets, in: F.P. James, S. Andrzej (Eds.), Transactions on Rough Sets I, Lecture Notes in Computer Science, vol. 3100, Springer, Berlin, Heidelberg, 2004, pp. 144–160.
[38] 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, 1998, pp. 428–439.
[39] Q. Shen, R. Jensen, Selecting informative features with fuzzy-rough sets and its application for complex systems monitoring, Pattern Recognition 37 (2004) 1351–1363.
[40] P. Somol, P. Pudil, J. Kittler, Fast branch and bound algorithms for optimal feature selection, IEEE Transactions on Pattern Analysis and Machine Intelligence 26 (2004) 900–912.
[41] P. Somol, P. Pudil, J. Novoviova, P. Paclik, Adaptive floating search methods in feature selection, Pattern Recognition Letters 20 (1999) 1157–1163.
[42] D.B. Stephen, S. Mark, Mining distance-based outliers in near linear time with randomization and a simple pruning rule, in: International Conference on Knowledge Discovery and Data Mining, Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, USA, 2003, pp. 29–38.
[43] Y.J. Sun, Iterative RELIEF for feature weighting: algorithms, theories, and applications, IEEE Transactions on Pattern Analysis and Machine Intelligence 29 (2007) 1035–1051.
[44] V. Suresh Babu, P. Viswanath, Weighted *k*-nearest leader classifier for large data sets, in: Pattern Recognition and Machine Intelligence, Lecture Notes in Computer Science, vol. 4815, Springer, Berlin, Heidelberg, 2007, pp. 17–24.
[45] R.W. Swiniarski, A. Skowron, Rough set methods in feature selection and recognition, Pattern Recognition Letters 24 (2003) 833–849.
[46] J.S. Taylor, N. Cristianini, On the generalization of soft margin algorithms, IEEE Transactions on Information Theory 48 (2002) 2721–2735.
[47] V.N. Vapnik (Ed.), Statistical Learning Theory, John Wiley and Sons, USA, 1998.
[48] 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.
[49] G.Y. Wang, J. Zhao, J.J. An, Y. Wu, A comparative study of algebra viewpoint and information viewpoint in attribute reduction, Fundamenta Informaticae 68 (2005) 289–301.
[50] W.-Z. Wu, W.-X. Zhang, Constructive and axiomatic approaches of fuzzy approximation operators, Information Sciences 159 (2004) 233–254.
[51] W.-Z. Wu, J.-S. Mi, W.-X. Zhang, Generalized fuzzy rough sets, Information Sciences 151 (2003) 263–282.
[52] 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.
[53] H. Xiong, G. Pandey, M. Steinbach, V. Kumar, Enhancing data analysis with noise removal, IEEE Transactions on Knowledge and Data Engineering 18 (3) (2006) 304–319.
[54] F.F. Xu, D.Q. Miao, L. Wei, An approach for fuzzy-rough sets attribute reduction via mutual information, in: Proceedings of the Fourth International Conference on Fuzzy Systems and Knowledge Discovery, USA, vol. 3, 2007, pp. 107–112.
[55] Y.Y. Yao, S.K.M. Wong, P. Lingras, A decision-theoretic rough set model, in: Z.W. Ras, M. Zemankova, M.L. Emrich (Eds.), Methodologies for Intelligent Systems, vol. 5, New York, 1990, pp. 17–24.
[56] Y.Y. Yao, Y. Zhao, Attribute reduction in decision-theoretic rough set models, Information Sciences 178 (2008) 3356–3373.
[57] L. Yun, L.L. Bao, Feature selection based on loss-margin of nearest neighbor classification, Pattern Recognition 42 (2009) 1914–1921.
[58] 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.
[59] L. Zhou, W.-Z. Wu, W.-X. Zhang, On characterization of intuitionistic fuzzy rough sets based on intuitionistic fuzzy implicators, Information Sciences 179 (2009) 883–898.
[60] X.Q. Zhu, X.D. Wu, Y. Yang, Error detection and impact-sensitive instance ranking in noisy datasets, in: Aaai Conference on Artificial Intelligence archive Proceedings of the 19th national conference on Artificial intelligence, 2004, pp. 378–383.
[61] 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.
[62] W. Ziarko, Variable precision rough set model, Journal of Computer and System Sciences 46 (1993) 39–59.