Knowledge-Based Systems 000 (2016) 1–7

Contents lists available at ScienceDirect

Knowledge-Based Systems

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

Feature subset selection based on fuzzy neighborhood rough sets

Changzhong Wang[a,*], Mingwen Shao[b], Qiang He[c], Yuhua Qian[d], Yali Qi[a]

[a] Department of Mathematics, Bohai university, Jinzhou, 121000, PR China
[b] College of Computer and Communication Engineering, Chinese University of Petroleum, Qingdao, Shandong, 266580, P.R China
[c] College of Science, Beijing University of Civil Engineering and Architecture, Beijing 100044, P.R China
[d] School of Computer and Information Technology, Shanxi University, Taiyuan 030006, P.R China

ARTICLE INFO

Article history:
Received 21 March 2016
Revised 4 August 2016
Accepted 7 August 2016
Available online xxx

Keywords:
Fuzzy neighborhood
Fuzzy decision
Feature selection
Rough set model

ABSTRACT

Rough set theory has been extensively discussed in machine learning and pattern recognition. It provides us another important theoretical tool for feature selection. In this paper, we construct a novel rough set model for feature subset selection. First, we define the fuzzy decision of a sample by using the concept of fuzzy neighborhood. A parameterized fuzzy relation is introduced to characterize fuzzy information granules for analysis of real-valued data. Then, we use the relationship between fuzzy neighborhood and fuzzy decision to construct a new rough set model: fuzzy neighborhood rough set model. Based on this model, the definitions of upper and lower approximation, boundary region and positive region are given, and the effects of parameters on these concepts are discussed. To make the new model tolerate noises in data, we introduce a variable-precision fuzzy neighborhood rough set model. This model can decrease the possibility that a sample is classified into a wrong category. Finally, we define the dependency between fuzzy decision and condition attributes and employ the dependency to evaluate the significance of a candidate feature, using which a greedy feature subset selection algorithm is designed. The proposed algorithm is compared with some classical algorithms. The experiments show that the proposed algorithm gets higher classification performance and the numbers of selected features are relatively small.

1. Introduction

Nowadays, databases expand quickly, more and more attributes are obtained in production practice. Some of attributes may be redundant or irrelevant to a classification task, they need to be removed before any further data processing can be carried out. Feature selection or attribute reduction is a technique for reducing features. Its aim is to find an optimal feature subset to predict sample categories. Feature subset selection can also facilitate data visualization and data understanding [11]. In recent years, much attention has been paid to feature selection in machine learning, data analysis and pattern recognition.

There is a key issue in feature selection process: feature evaluation. How to construct an effective evaluation function is one of the most important steps. It directly affects the performance of a classifier. A lots of feature evaluation measures, such as information entropy [8,12], dependency [3,9–11], correlation [7], and consistency [4], have been proposed for feature selection until now. In general, different evaluation measures may lead to different optimal feature subsets. However, every measure is aimed to determine the discriminating ability of a subset of features.

The classical rough set theory [17] has been proven to be an effective tool for feature selection. It employs a dependency function to evaluate the classification quality of a subset of attributes. However, this model is just applicable to nominal data. In practical problems, it is most often the case that the values of attributes may be both crisp and real-valued. The real-valued features need to be discretized before the dependency is calculated. The inherent error that exists in discretization process is of major concern. This is where the traditional rough set theory encounters a problem.

Some generalizations of the model were proposed to deal with this problem [5,6,13–16,19–25]. Neighborhood rough set and fuzzy rough set are considered two important models. Lin generalized the classical rough set with neighborhood operators and introduced a neighborhood rough set model [14]. Dubois and Prade defined fuzzy rough approximation operators by combining rough sets and fuzzy sets and proposed a fuzzy rough set model [5]. Recently, some feature selection algorithms based on the generalized models have been proposed [1–3,8–11,18,20,26].

As we know, the core idea of rough set theory is based on granulation and approximation. In a neighborhood rough set, neighborhood similarity classes are used to approximately characterize decision equivalence classes. The limitation of this model is that it

* Corresponding author.
E-mail address: changzhongwang@126.com (C. Wang).

http://dx.doi.org/10.1016/j.knosys.2016.08.009
0950-7051/

cannot describe the fuzziness of samples in fuzzy background. In classical fuzzy rough set model, fuzzy information granules are the elemental granules. The membership degrees of a sample to different decision classes are computed by min-max operations. That is to say, the decision of a sample is based on a nearest sample. However, there may be some risks in computation of fuzzy lower approximations when a data set has noise. Data noise can destroy the accuracy of calculation of membership degrees and lead to an increase in classification error rate. To better describe sample decisions by using fuzzy information granules, a new rough set model, named fuzzy neighborhood rough set model, is introduced in this paper.

We first define the fuzzy decision of a sample and employ a parameterized fuzzy relation to characterize its fuzzy information granule. We then use the inclusion relation of them to decide whether the sample is classified into one of decision classes. Because this way of decision-making fully utilizes the classification information of multiple samples, it overcomes the disadvantage of fuzzy rough set model by using a nearest neighbor to determine the membership degree of a sample to different decision classes. The proposed model is a nature generalization of neighborhood rough sets. This is the main difference from the classical fuzzy rough set theory. As the proposed model is too strict to tolerate noise in data, a variable precision fuzzy neighborhood rough set model is introduced. This model is more effective to process the fuzzy or uncertain knowledge because it can decrease the possibility that a sample is classified into a wrong class. Finally, we define the dependency between features and decision and design a feature selection algorithm. Numerical experiments show that the proposed algorithm yields better performance.

The paper is organized as follows. In Section 2, we review some relevant literature about neighborhood rough sets and fuzzy rough sets. In Section 3, we develop a new model: fuzzy neighborhood rough set model. In Section 4, we design a heuristic algorithm of attribute reduction. In Section 5, we verify the feasibility and stability of the proposed algorithm. Section 6 concludes the paper.

2. Literature reviews

Neighborhood is one of important concepts in classification learning and reasoning with uncertainty. A neighborhood relation can be used to generate a family of neighborhood granules characterized with numerical features [15]. In 1997, Lin pointed out that neighborhoods are more general information granules than equivalence classes and introduced neighborhood relations into rough set methodology [14]. Based on this observation, a neighborhood rough set model was constructed. Then, Wu and Zhang studied some properties of neighborhood approximation spaces [22]. Yao discussed the relationship between neighborhood operators and rough approximation operators and presented the axiomatic properties of this model [23]. In 2008, Hu employed the neighborhood rough set model to deal with feature subset selection in real-valued sample space [9]. In fact, the neighborhood model is a natural generalization of classical rough sets. The model can be used to deal with mixed numerical and categorical data within a uniform framework and overcomes the drawback of discretization of data in classical rough sets. However, it cannot describe the fuzziness of samples in fuzzy background.

Fuzzy rough sets, as proposed by Dubois and Prade [5], can also deal with numerical or continuous data sets directly. Numerical attribute values are no longer needed for discretization. In this model, a fuzzy similarity relation is defined to measure the similarity between samples. The fuzzy upper and lower approximations of a decision are then defined by using the fuzzy similarity relation. The fuzzy positive region is defined as the union of the fuzzy lower approximations of decision equivalence classes. As the fuzziness is introduced into the rough set theory, more information of continuous attribute values is easily kept. So, feature selection with fuzzy rough sets becomes another important tool in handling dataset with real-valued attributes. In recent years, a series of feature selection algorithms based on fuzzy rough sets have been proposed. Jensen introduced the dependency function in classical rough sets into fuzzy rough sets and proposed an a greedy algorithm for reducing redundant attributes [11]. Bhatt and Gopal presented the concept of compact computational domain for Jensen's algorithm to improve computational efficiency [1]. Chen used fuzzy rough sets to define fuzzy discernibility matrix by which all attribute reducts are computed [2]. For data-based attribute selection, Cornelis generalized the classical rough set model using fuzzy tolerance relations within the context of fuzzy rough set theory [3]. Hu et al. employed kernel functions to define fuzzy similarity relations and constructed a greedy algorithm for dimensionality reduction [10]. Meanwhile, the classical fuzzy rough set model was improved to analyze noisy data. Mieszkowicz Rolka introduced the model of variable precision fuzzy rough sets to deal with noisy data [19], where the fuzzy memberships of a sample to the lower and upper approximations were computed with fuzzy inclusion. Zhao et al. defined the concept of fuzzy variable precision rough sets to handle noise of misclassification and perturbation [26]. To solve the problem of data fitting in classical fuzzy rough sets, Wang proposed a fitting fuzzy rough set model to conduct feature selection [20]. However, in all kinds of fuzzy rough set models, the fuzzy upper and lower approximations of a decision is computed by using a nearest sample, there may be some risks when a data set has noise. This is the main drawback of fuzzy rough set models.

3. Fuzzy neighborhood rough set model

Let $\langle U, A, D\rangle$ be a decision table, where $U = \{x_1, x_2, \ldots, x_n\}$ is called a sample space, A is a set of attributes or features characterizing samples and D is a decision attribute. Assume that the samples are partitioned into r mutually exclusive decision classes by D, that is, $U/D = \{D_1, D_2, \ldots, D_r\}$. In this section, the fuzzy decision of a sample is defined and parameterized fuzzy information granules associated with samples are introduced. The task is to approximate the fuzzy decision classes with parameterized fuzzy information granules.

Let $B \subseteq A$ be a subset of attributes on U, and then B can induce a fuzzy binary relation R_B on U. R_B is called a fuzzy similarity relation if it satisfies

(1) Reflectivity: $R_B(x, x) = 1$, $\forall x \in U$; (2) Symmetry: $R_B(x, y) = R_B(y, x)$, $\forall x, y \in U$.

Let $a \in B$ and R_a be a fuzzy similarity relation induced by a, we denote $R_B = \bigcap_{a\in B} R_a$. For any $x \in U$, the fuzzy neighborhood of x is defined as $[x]_B(y) = R_B(x, y)$, $y \in U$.

Definition 1. Given a decision table $\langle U, A, D\rangle$, $U/D = \{D_1, D_2, \ldots D_r\}$. R_A is the fuzzy similarity relation on U induced by A, $\forall x \in U$, the fuzzy decision of x is defined as follows.

$$\tilde{D}_i(x) = \frac{\left|[x]_A \cap D_i\right|}{\left|[x]_A\right|}, \quad i = 1, 2, \ldots, r,$$

where $\tilde{D}_i$ is a fuzzy set and $\tilde{D}_i(x)$ indicates the membership degree of x to D_i. We call $\{\tilde{D}_1, \tilde{D}_2, \ldots \tilde{D}_r\}$ the fuzzy decisions of samples induced by D.

Please cite this article as: C. Wang et al., Feature subset selection based on fuzzy neighborhood rough sets, Knowledge-Based Systems (2016), http://dx.doi.org/10.1016/j.knosys.2016.08.009

Example 1. Given a decision table $\langle U, A, D\rangle$, where $U = \{x_1, x_2, \ldots, x_5\}$. R_A is the fuzzy similarity relation induced by A and

$$R_A = \begin{bmatrix} 1 & 0.5 & 0.7 & 0 & 0 \\ 0.5 & 1 & 0.7 & 0.2 & 0.6 \\ 0.7 & 0.7 & 1 & 0.1 & 0.5 \\ 0 & 0.2 & 0.1 & 1 & 0.7 \\ 0 & 0.6 & 0.5 & 0.7 & 1 \end{bmatrix}.$$

Suppose that $U/D = \{D_1, D_2\}$ such that $D_1 = \{x_1, x_2, x_3\}$ and $D_2 = \{x_4, x_5\}$, then

$$\tilde{D}_1 = \frac{1}{x_1} + \frac{0.73}{x_2} + \frac{0.8}{x_3} + \frac{0.15}{x_4} + \frac{0.39}{x_5},$$
$$\tilde{D}_2 = \frac{0}{x_1} + \frac{0.27}{x_2} + \frac{0.2}{x_3} + \frac{0.85}{x_4} + \frac{0.61}{x_5}$$

So, we get the fuzzy decisions $\{\tilde{D}_1, \tilde{D}_2\}$ of samples.

To analyze a classification task under different information granularity, we need to introduce a parameter λ to characterize the similarity of samples. Let $B \subseteq A$, R_B is the fuzzy similarity relation on U induced by B. For any $x \in U$, a parameterized fuzzy information granule associated with x is constructed as follows.

$$[x]_B^\lambda(y) = \begin{cases} 0, & R_B(x,y) < \lambda; \\ R_A(x,y), & R_B(x,y) \geq \lambda. \end{cases}$$

We call λ the radius of the fuzzy neighborhood of samples. There are two factors λ and B that impact on the membership degrees. Obviously, the following properties hold.

(1)$R_A \subseteq R_B$ for $B \subseteq A$, (2)$[x]_B^{\lambda_2} \subseteq [x]_B^{\lambda_1}$ for $\lambda_1 \leq \lambda_2$ and any $x \in U$.

In the following, we use the relationship between fuzzy information granule and fuzzy decision to define the fuzzy lower and upper approximations of a decision.

Definition 2. Given a decision table $\langle U, A, D\rangle$, $B \subseteq A$, $U/D = \{D_1, D_2, \ldots D_r\}$ and a neighborhood radius λ. $\{\tilde{D}_1, \tilde{D}_2, \ldots \tilde{D}_r\}$ are the fuzzy decisions of samples induced by D, R_B is the fuzzy similarity relation on U induced by B, the lower and upper approximations of D with respect to B are defined as

$$\underline{R}_B^\lambda(D) = \left\{\underline{R}_B^\lambda(\tilde{D}_1), \underline{R}_B^\lambda(\tilde{D}_2), \ldots \underline{R}_B^\lambda(\tilde{D}_r)\right\},$$
$$\overline{R}_B^\lambda(D) = \left\{\overline{R}_B^\lambda(\tilde{D}_1), \overline{R}_B^\lambda(\tilde{D}_2), \ldots \overline{R}_B^\lambda(\tilde{D}_r)\right\}.$$

where

$$\underline{R}_B^\lambda\left(\tilde{D}_i\right) = \left\{x_i \in D_i | [x_i]_B^\lambda \subseteq \tilde{D}_i\right\},\ \overline{R}_B^\lambda\left(\tilde{D}_i\right) = \left\{x_i \in D_i | [x_i]_B^\lambda \cap \tilde{D}_i \neq \emptyset\right\}.$$

$\underline{R}_B^\lambda(\tilde{D}_i)$ and $\overline{R}_B^\lambda(\tilde{D}_i)$ are called the fuzzy neighborhood lower approximation and upper approximation, respectively.

They share the same idea of approximating a decision class as the classical rough set model. Whether a sample can be correctly classified into its own category depends on the relationship between its fuzzy similarity class and decision class. If the fuzzy similarity class is completely contained in its decision class, then the sample can be classified into its own category with certainty. If the fuzzy similarity class is partly included in the decision class, then the sample probably belongs to its own category. Obviously, $\underline{R}_B^\lambda(\tilde{D}_i)$ is a set of samples which definitely belong to D_i. $\overline{R}_B^\lambda(\tilde{D}_i)$ is a set of samples which possibly belong to D_i. The fuzzy neighborhood lower approximation of D_i is also called fuzzy positive region of D_i.

If the fuzzy neighborhoods and fuzzy decisions respectively degrade to similarity classes and equivalence classes, the proposed approximations degenerate to the corresponding ones in neighborhood rough set model. Thus, the proposed model is a generalization of neighborhood rough sets [8].

In classical fuzzy rough set model, fuzzy neighborhoods and fuzzy decisions can also be used to construct the fuzzy rough approximations of a decision [5], but the model determines the membership degree of each sample to different decision classes based on a nearest sample. In our proposed model, the decision of a sample is made by the relationship between its fuzzy decision and fuzzy neighborhood. This is the point where the proposed model is different.

Definition 3. Given a neighborhood radius λ, $B \subseteq A$, and fuzzy decision $\{\tilde{D}_1, \tilde{D}_2, \ldots \tilde{D}_r\}$ induced by D, R_B is the fuzzy similarity relation on U induced by B. The positive and boundary regions of D with respect to B are defined as follows, respectively.

$$POS_B^\lambda(D) = \bigcup_{i=1}^{r} \underline{R}_B^\lambda\left(\tilde{D}_i\right), \quad BN_B^\lambda(D) = U - POS_B^\lambda(D).$$

The size of the positive region $POS_B^\lambda(D)$ reflects the classification ability of B.

Definition 4. Given a neighborhood radius λ and $B \subseteq A$, the dependency degree of D upon B is defined as $\partial_B^\lambda(D) = \frac{|POS_B^\lambda(D)|}{|U|}$.

The dependency function is defined as the ratio of the sizes of the positive region over all samples. It is used to determine the relevance between decision and conditional attributes.

The sizes of neighborhood radius and feature subset have great impacts on the positive region and dependency function.

Theorem 1. *Given a neighborhood radius λ, if $B_1 \subseteq B_2 \subseteq A$, then $POS_{B_1}^\lambda(D) \subseteq POS_{B_2}^\lambda(D)$.*

Proof. Since $B_1 \subseteq B_2$, we have $R_{B_2} \subseteq R_{B_1}$, which implies that $[x]_{B_2}^\lambda \subseteq [x]_{B_1}^\lambda$ for any $x \in U$. It follows from Definition 2 that $\underline{R}_{B_1}^\lambda(D_i) \subseteq \underline{R}_{B_2}^\lambda(D_i)$ for any $D_i \in U/D$. Hence, $POS_{B_1}^\lambda(D) \subseteq POS_{B_2}^\lambda(D)$. □

Theorem 2. *Given $B \subseteq A$, if $\lambda_1 \leq \lambda_2$, then $POS_B^{\lambda_1}(D) \subseteq POS_B^{\lambda_2}(D)$.*

Proof. Since $\lambda_1 \leq \lambda_2$, we have $[x]_B^{\lambda_2} \subseteq [x]_B^{\lambda_1}$ for any $x \in U$. By the definition of lower approximation, we have $\underline{R}_B^{\lambda_1}(D_i) \subseteq \underline{R}_B^{\lambda_2}(D_i)$ for any $D_i \in U/D$. Hence, $POS_B^{\lambda_1}(D) \subseteq POS_B^{\lambda_2}(D)$. □

According to Theorems 1 and 2, we easily get the following properties.

Theorem 3. *If $B_1 \subseteq B_2 \subseteq \cdots \subseteq B_m \subseteq A$, then $\partial_{B_1}^\lambda(D) \leq \partial_{B_2}^\lambda(D) \leq \cdots \leq \partial_{B_m}^\lambda(D)$.*

Theorem 4. *Given $B \subseteq A$, if $\lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_m$, then $\partial_B^{\lambda_1}(D) \leq \partial_B^{\lambda_2}(D) \leq \cdots \leq \partial_B^{\lambda_m}(D)$.*

Definition 5. Given a neighborhood radius λ and $B \subseteq A$, for any $a \in B$, if $\partial_{B-a}^\lambda(D) \neq \partial_B^\lambda(D)$, we say attribute a is indispensable in B. Otherwise, we say a is redundant in B.

A redundant attribute not only cannot provide more classification information, but also will cut down the classification accuracy of a decision table. Therefore, it must be deleted from the attribute set before classification learning.

Definition 6. Given a neighborhood radius λ and $B \subseteq A$, we say B is a reduct of A if it satisfies

$\partial_B^\lambda(D) = \partial_A^\lambda(D)$, (2) $\forall a \in B$, $\partial_{B-a}^\lambda(D) < \partial_B^\lambda(D)$.

The first condition means that a reduct has the same classification ability as the whole attribute set. The second one ensures there is no redundant attribute in the reduct.

In practice, the above definitions of lower and upper approximations are too strict to tolerate noise in data. In the following,

Please cite this article as: C. Wang et al., Feature subset selection based on fuzzy neighborhood rough sets, Knowledge-Based Systems (2016), http://dx.doi.org/10.1016/j.knosys.2016.08.009

we introduce a variable precision fuzzy neighborhood rough set model.

Definition 7. Let A and B be two fuzzy sets on U, the inclusion $I(A, B)$ is defined as

$$I(A,B)=\frac{|A\subseteq B|}{|U|},$$

where $|A \subseteq B|$ denotes the number of samples whose membership degrees to A are not greater than those to B. We call $I(A, B)$ the inclusion degree of A in B.

Example 2. Given a set $X=\{x_1,x_2,\ldots,x_{10}\}$, A and B are two fuzzy sets defined on X, where

$$A=\frac{0.3}{x_1}+\frac{0}{x_2}+\frac{0.8}{x_3}+\frac{0.7}{x_4}+\frac{0}{x_5}+\frac{0.6}{x_6}+\frac{0}{x_7}+\frac{1}{x_8}+\frac{0.2}{x_9}+\frac{0.5}{x_{10}},$$
$$B=\frac{0.2}{x_1}+\frac{0}{x_2}+\frac{1}{x_3}+\frac{0.5}{x_4}+\frac{0.4}{x_5}+\frac{0.9}{x_6}+\frac{0.1}{x_7}+\frac{1}{x_8}+\frac{0.3}{x_9}+\frac{0.2}{x_{10}}.$$

Then, we can get $|A \subseteq B| = 7$ and $|B \subseteq A| = 5$. Thus, $I(A, B) = 0.7$ and $I(B, A) = 0.5$.

Definition 8. Given a neighborhood radius λ, $B \subseteq A$, and the fuzzy decision $\{\tilde{D}_1, \tilde{D}_2, \ldots \tilde{D}_r\}$ induced by D. R_B is the fuzzy similarity relation on U induced by B. Then the variable precision lower and upper approximations of D with respect to B are defined as follows, respectively.

$$\underline{R}_B^{\lambda,\alpha}(D)=\left\{\underline{R}_B^{\lambda,\alpha}(\tilde{D}_1),\underline{R}_B^{\lambda,\alpha}(\tilde{D}_2),\ldots\underline{R}_B^{\lambda,\alpha}(\tilde{D}_r)\right\}$$
$$\overline{R}_B^{\lambda,\beta}(D)=\left\{\overline{R}_B^{\lambda,\beta}(\tilde{D}_1),\overline{R}_B^{\lambda,\beta}(\tilde{D}_2),\ldots\overline{R}_B^{\lambda,\beta}(\tilde{D}_r)\right\},$$

where

$$\underline{R}_B^{\lambda,\alpha}(\tilde{D}_i)=\{x_i\in D_i | I([x_i]_B^{\lambda},\tilde{D}_i)\geq\alpha\},\ 0.5\leq\alpha\leq 1$$
$$\overline{R}_B^{\lambda,\beta}(\tilde{D}_i)=\{x_i\in D_i | I([x_i]_B^{\lambda},\tilde{D}_i)>\beta\},\ 0\leq\beta<0.5.$$

Obviously, $\underline{R}_B^{\lambda,\alpha}(\tilde{D}_i)\subseteq\overline{R}_B^{\lambda,\beta}(\tilde{D}_i)$ for any α, β and decision equivalence class $D_i \in U/D$. The variable precision boundary region of D with respect to B is defined as

$$BN_B^{\lambda,\alpha\beta}(D)=\overline{R}_B^{\lambda,\beta}(D)-\underline{R}_B^{\lambda,a}(D),$$

where $\overline{R}_B^{\lambda,\beta}(D)=\bigcup_{i=1}^{r}\overline{R}_B^{\lambda,\beta}(\tilde{D}_i)$ and $\underline{R}_B^{\lambda,\alpha}(D)=\bigcup_{i=1}^{r}\underline{R}_B^{\lambda,\alpha}(\tilde{D}_i)$. Here, $\underline{R}_B^{\lambda,\alpha}(D)$ are also called variable precision positive region of D with respect to B, denoted as $POS_B^{\lambda,\alpha}(D)$.

Definition 9. The variable precision dependency of D on B is defined as

$$\partial_B^{\lambda,\alpha}(D)=\frac{\left|POS_B^{\lambda,\alpha}(D)\right|}{|U|}.$$

Similarly, we also have the following theorem as to monotonicity.

Theorem 5. *For given parameters λ and α, if $B_1 \subseteq B_2 \subseteq A$, then we have*

(1) $POS_{B_1}^{\lambda,\alpha}(D)\subseteq POS_{B_2}^{\lambda,\alpha}(D)$, (2) $\partial_{B_1}^{\lambda,\alpha}(D)\leq\partial_{B_2}^{\lambda,\alpha}(D)$.

Definition 10. Given a neighborhood radius λ, and $B \subseteq A$. We say B is a variable precision reduct, if it satisfies (1) $\partial_B^{\lambda,\alpha}(D)=\partial_A^{\lambda,\alpha}(D)$, (2) $\forall a\in B$, $\partial_{B-a}^{\lambda,\alpha}(D)<\partial_B^{\lambda,\alpha}(D)$.

4. Attribute reduction algorithm based on fuzzy neighborhood rough set model

As discussed above, the dependency function reflects the classification power of an attribute subset. It can be used to measure the significance of a candidate attribute.

Algorithm Heuristic algorithm based on fuzzy neighborhood rough sets (FNRS).

Input: Decision table $\langle U, A, D\rangle$, parameters λ and α, //λ controls the size of radius of fuzzy neighborhood, α is the threshold for computing inclusion degree.
Output: one reduct *red*.
1: $\forall a \in A$, compute the relation matrix R_a^{λ};
2: Compute the fuzzy decision $\tilde{D}=\{\tilde{D}_1,\tilde{D}_2,\ldots\tilde{D}_r\}$;
3: Initialize: $red=\emptyset$, $B=A-red$, start=1; // *red* is the pool to contain the selected attributes and B is for the left attributes.
4: while start
5: for each $a_i \in B$
6: Compute the fuzzy similarity relation $R_{red\cup\{a_i\}}^{\lambda}$.
7: for each $x_j \in U$, suppose $x_j \in D_i$;
8: Compute the lower approximation function $\underline{R_{red\cup\{a_i\}}^{\lambda,\alpha}}(\tilde{D}_i)$.
9: end for
10: $\partial_{red\cup\{a_i\}}^{\lambda,\alpha}(D)=|\bigcup_{D_i\in U/D}\underline{R_{red\cup\{a_i\}}^{\lambda,\alpha}}(\tilde{D}_i)|/|U|$;
11: end for
12: Find attribute a_k with maximum value $\partial_{red\cup\{a_k\}}^{\lambda,\alpha}(D)$.
13: Compute $SIG^{\lambda,\alpha}(a_k, red, D)=\partial_{red\cup\{a_k\}}^{\lambda,\alpha}(D)-\partial_{red}^{\lambda,\alpha}(D)$.
14: if $SIG^{\lambda,\alpha}(a_k, red, D)>0$
15: $red\leftarrow red\cup a_k$;
16: $B\leftarrow B-red$;
17: else
18: start=0;
19: end if
20: end while
21: return *red*.

Table 1
Description of data sets.

No	Data sets	Sample	Attributes	Classes
1	Wine	178	13	3
2	Glass	214	10	6
3	Cancer	683	9	2
4	Ionos	351	33	2
5	CT	221	36	2
6	Wdbc	569	30	2
7	Colon	62	1225	2
8	Wpbc	198	32	2

Table 2
Numbers of selected features with four reduction algorithms.

Data sets	Raw data	FCMRS	FISEN	FNRS	FRSINT
Wine	13	5	5	9	8
Glass	10	8	4	5	6
Cancer	9	7	5	5	6
Ionos	33	10	7	8	9
CT	36	7	7	9	8
Wdbc	30	9	16	18	19
Colon	1225	4	10	3	26
Wpbc	32	7	6	8	8
Avaerage	173.5	7.13	7.5	8.13	11.25

Definition 11. Given a neighborhood radius λ, $B \subseteq A$, and $a \in A - B$, the significance of a with respect to B is defined as $SIG^{\lambda,\alpha}(a, B, D)=\partial_{B\cup\{a\}}^{\lambda,\alpha}(D)-\partial_B^{\lambda,\alpha}(D)$.

The objective of attribute reduction is to find a minimal subset of attributes which has the same discriminating power as the original data. Although there are usually multiple reducts for a given decision table, it is enough to find one in most of applications. With the proposed measure of attribute significance, a greedy forward Algorithm can be formally designed as follows.

As described above, this algorithm stops when the addition of any left feature does not make the dependency $\partial_B^{\lambda,\alpha}(D)$ increase. If there are n samples and m condition features, the computational complexity for a fuzzy similarity relation is $\frac{1}{2}n^2$, the worst search time for a reduct will bring about m^2 evaluations of the

Please cite this article as: C. Wang et al., Feature subset selection based on fuzzy neighborhood rough sets, Knowledge-Based Systems (2016), http://dx.doi.org/10.1016/j.knosys.2016.08.009

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

Data sets	Raw data	FCMRS	FISEN	FNRS	FRSINT
wine	97.44 ± 3.43	94.86 ± 2.75	97.71 ± 3.96	97.22 ± 4.72	96.60 ± 3.51
glass	93.49 ± 5.05	92.53 ± 3.27	94.29 ± 5.85	94.29 ± 3.01	92.06 ± 7.30
cancer	95.83 ± 2.14	97.06 ± 2.68	96.93 ± 2.97	96.79 ± 2.82	97.21 ± 1.83
Ionos	90.42 ± 4.93	94.59 ± 3.91	94.31 ± 3.09	94.31 ± 3.54	93.17 ± 3.15
CT	87.78 ± 6.59	89.13 ± 6.51	90.49 ± 7.86	92.79 ± 5.86	91.82 ± 6.05
wdbc	96.77 ± 2.17	96.49 ± 2.34	97.01 ± 2.62	97.36 ± 2.23	97.18 ± 2.09
colon	76.10 ± 17.57	82.92 ± 14.22	82.92 ± 11.86	85.00 ± 16.57	80.42 ± 10.77
wpbc	77.35 ± 8.78	78.33 ± 8.05	80.28 ± 7.30	82.83 ± 7.86	79.78 ± 14.53
Avaerage	89.40 ± 6.33	90.73 ± 5.47	91.74 ± 5.69	92.57 ± 5.83	91.03 ± 6.15

dependency function. Therefore, the overall computational complexity of the proposed algorithm is about $O(\frac{1}{2}n^2m + m^2)$.

5. Experimental analysis

In this section, we evaluate the performance of the proposed method by comparing it with existing methods. These methods are classical rough set based algorithm (FCMRS) [17], fuzzy entropy based algorithm (FISEN) [8] and fuzzy rough set based algorithm (FRINT)[11]. We first compare the numbers of selected features with different algorithms. Then, we present the comparative results of classification accuracies. Finally, we discuss the influences of the parameters λ and α on classification performance with our proposed algorithm. All of the algorithms are performed in Matlab 2007 and run in the hardware environment with Pentium (R) Core 2, CPU E5200, 2.50 GHz and 2.0GB RAM.

Two classification learning algorithms are introduced to evaluate the performance of different algorithms. The classifiers are support vector machine (RBF-SVM) and k-nearest neighbor rule (K-NN, K=3). To compute the classification accuracy of different classifiers, the 10-fold cross validation is used. Eight data sets are used in the experimental analysis. They are selected from UCI Machine Learning Repository. The information of these data sets is outlined in Table 1. All the numerical attributes are first normalized into the interval [0, 1] with the formula $a' = (a - a_{\min})/a_{\max}$. The value of the fuzzy similarity degree r_{ij} between objects x_i and x_j with respect to an attribute a is computed as

$$r_{ij} = \begin{cases} \rho * (1 - |x_i - x_j|), & |x_i - x_j| \le 1 - \lambda; \\ 0 & |x_i - x_j| > 1 - \lambda. \end{cases}$$

Here, ρ is an adjustable constant coefficient and $0 < \rho \le 1$. As $r_{ij} = r_{ji}$ and $0 \le r_{ij} < 1$, the matrix $M_a^\lambda = (r_{ij})_{n\times n}$ is a fuzzy similarity relation. To make more samples fall into the positive region, we set $\rho = 0.5$ in the following series of experiments.

Since the classical rough set considers only categorical data, a fuzzy C-means clustering (FCM) technique is employed to discretize numerical data. The numeric attributes are discretized into four intervals. In the FNRS algorithm, there are two parameters λ and α. The parameter λ is used to control the size of fuzzy neighborhood. We set the value of λ to vary from 0.1 to 0.5 with a step of 0.05. The parameter α is introduced to compute the inclusion degree and variable precision lower approximation. We set α to vary from 0.5 to 1 with a step of 0.05. As different learning algorithms may require different feature subsets to produce the best classification accuracy, all the experimental results reported in the following tables are presented at the highest classification accuracy.

From Table 2, we can find that these reduction methods can effectively reduce attributes. The numbers of selected features with FCMRS are fewer compared to the other algorithms in most of the cases. The reason for this result may be due to the information loss caused by data discretization, because the resulting classification accuracies are lower as shown in Tables 3 and 4. The numbers of selected features with FISEN, FRDMA and FRSINT are comparable except for Colon data set.

Tables 3 and 4 show the classification accuracies of the raw data and the reduced data sets, where the underlined symbols indicate the highest classification accuracies among the reduced data sets. From the results of Tables 3 and 4, it is easily seen that the classification accuracies based on FCMRS method are obviously lower than the other methods. Out of 16 cases of 10-fold cross validation, the FNRS and FISEN methods achieve the highest classification accuracy in 9 and 5 cases, respectively. The FRSINT method obtains it in 2 cases, while FCMRS attains it for only once. For the FNRS algorithm, there are 14 cases higher than the FCMRS algorithm. There are 9 cases higher than and 2 cases the same as the FISEN algorithm. There are 13 cases higher than the FRSINT algorithm.

The classification performances are improved for all the original data sets. As to SVM, FNRS outperforms the raw data 7 times over the 8 classification tasks. In the same time, FNRS outperforms the raw data 6 times with respect to 3NN. Moreover, the average accuracy of FNRS outperforms that of any other algorithm in terms of SVM and 3NN.

The feature subsets with the greatest accuracies, selected by FISEN and FNRS according to SVM, are shown in Table 5. The last column shows the corresponding values of λ and α in FNRS.

From Table 5, we can find that most of the best features for FISEN and FNRS are the same in most cases, especially for

Table 4
Comparison of classification accuracies of reduced data with 3NN.

Data sets	Raw data	FCMRS	FISEN	FNRS	FRSINT
Wine	96.52 ± 4.33	93.82 ± 4.09	97.71 ± 2.97	98.33 ± 2.69	97.78 ± 2.93
glass	90.73 ± 6.72	90.63 ± 5.52	92.53	92.38 ± 4.60	89.83 ± 7.12
cancer	96.95 ± 1.89	96.62 ± 2.31	96.62 ± 2.87	96.64 ± 3.31	95.75 ± 1.72
Ionos	85.57 ± 6.41	87.77 ± 4.72	89.46 ± 4.72	90.89 ± 4.41	91.44 ± 4.72
CT	89.47 ± 6.58	88.68 ± 6.53	90.04 ± 4.90	89.17 ± 4.76	89.15 ± 5.22
wdbc	96.83 ± 2.57	94.91 ± 2.40	96.84 ± 1.99	97.37 ± 1.24	96.62 ± 2.61
colon	76.25 ± 16.39	86.25 ± 13.90	86.25 ± 13.90	86.67 ± 15.32	73.92 ± 19.76
wpbc	74.84 ± 9.79	70.61 ± 10.87	77.72 ± 7.33	75.72 ± 7.47	76.72 ± 14.53
Avaerage	88.40 ± 6.84	88.79 ± 6.29	90.89 ± 5.77	90.90 ± 5.48	88.90 ± 7.33

Please cite this article as: C. Wang et al., Feature subset selection based on fuzzy neighborhood rough sets, Knowledge-Based Systems (2016), http://dx.doi.org/10.1016/j.knosys.2016.08.009

Table 5
The best feature subsets of FISEN and FNRS algorithm.

Data sets	FISEN	FNRS	(λ, α)
Wine	12, 13, 1,7, 10	1, 12, 8, 2, 10, 13, 7, 9, 11	(0.15,0.6)
Glass	1, 4, 7, 9	1, 10, 4, 7, 9	(0.1,0.6)
Cancer	6, 2, 8, 3, 1	6, 2, 8, 1, 3	(0.35,0.75)
Ionos	4, 5, 33, 28, 7, 22, 2	4, 23, 14, 32, 2, 7, 6, 16	(0.3,0.6)
CT	31, 33, 32, 30, 20, 4, 29	31, 33, 20, 29, 35, 3, 30, 5	(0.2,1)
Wdbc	28, 21, 22, 11, 7, 29, 16, 12, 19, 9, 2, 27, 26, 5, 8, 23	8, 26, 21, 28, 22, 7, 16, 29, 5, 10, 6, 5, 11, 19, 18, 27, 2, 4	(0.3,0.95)
Colon	1224, 1173, 329, 951, 1084, 500, 1050, 1216, 1208, 121	1205, 1224, 4	(0.4,0.95)
Wpbc	1, 13, 24, 16, 12, 32	5, 32, 20, 12, 1, 13, 9, 24	(0.2,0.55)

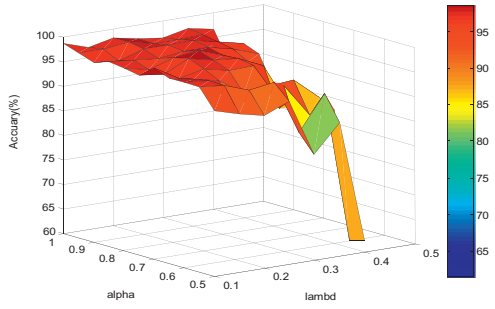

Fig. 1. The accuracy varying with thresholds λ and α (Wine).

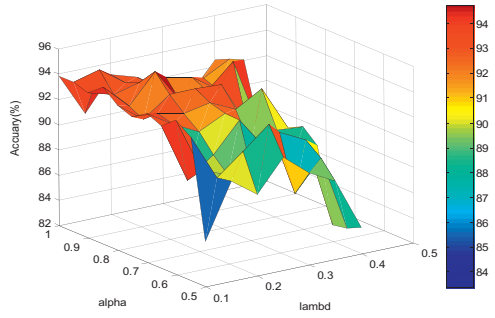

Fig. 2. The accuracy varying with thresholds λ and α (Glass).

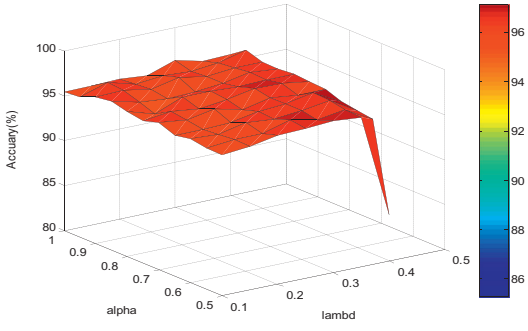

Fig. 3. The accuracy varying with thresholds λ and α (Cancer).

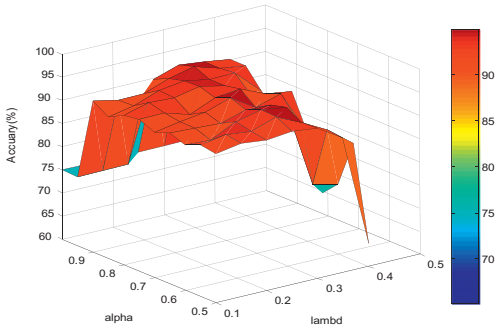

Fig. 4. The accuracy varying with thresholds λ and α (Ionos).

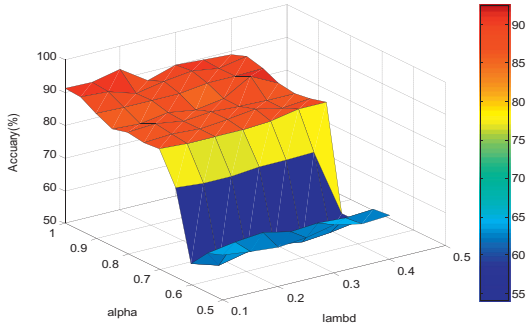

Fig. 5. The accuracy varying with thresholds λ and α (CT).

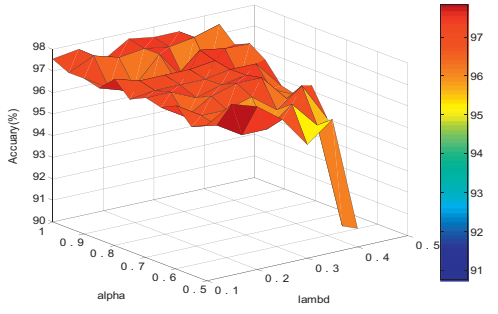

Fig. 6. The accuracy varying with thresholds λ and α (Wdbc).

Wine, Glass, Cancer, CT data sets. Comparing their classification accuracies shown in Tables 3 and 4, we can say that the proposed method can be comparable to fuzzy entropy based method.

The thresholds λ and α play an important role in FNRS algorithm. λ is considered as a parameter to control the size of fuzzy neighborhood radius, the parameter α is used to control the inclusion degree and overcome the bad affects caused by noises in data. Figs. 1–8 show classification accuracies of SVM varying with λ and α. We can select the suitable value of λ and α for each data set according to these figures. The experimental results obtained using 3KNN are roughly consistent with SVM. From the Figs. 1–8, we can see that most of data sets achieve higher precision in a larger area. Thus, the FNRS algorithm is of feasibility and stability.

6. Conclusions and future works

Reducing redundant features can improve classification performance and decrease the cost of classification. In this paper, we first introduced a new rough set model: fuzzy neighborhood rough set. As the model is too strict to tolerate noise in the data, we then proposed the variable precision model. This model overcomes

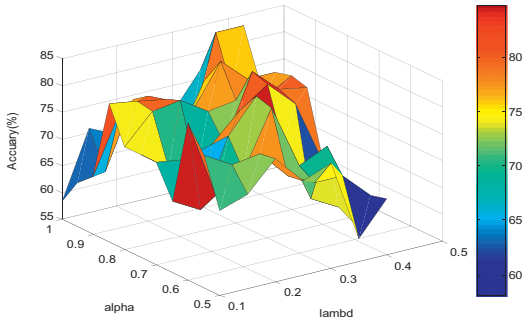

Fig. 7. The accuracy varying with thresholds λ and α (Colon).

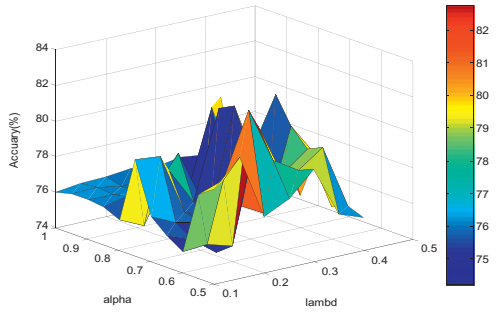

Fig. 8. The accuracy varying with thresholds λ and α (Wpbc).

the possibility that a sample belongs to several classes. Finally, we define the dependency between fuzzy decision and condition attributes and employ the dependency to evaluate the significance of a candidate feature, using which a greedy feature subset selection algorithm is designed. The experimental results show that the algorithm can find a small and effective subset of features and obtain high classification accuracy. We also find that the two parameters have great impact on the performance of the proposed attribute reduction algorithm. We should select the suitable values of parameters for each data set according to the numbers of selected features and classification accuracies.

Future works may include 1) How can the proposed model be applied to the fields of classification learning and reasoning with uncertainty? 2) In the proposed model, the two parameters have important impact on the performance of the proposed algorithm. They need to be set by users in advance. How to automatically set the optimal solutions of two parameters for each data set is also an interesting work.

Acknowledgement

This work was supported by the National Natural Science Foundation of China under Grants 61572082, 61473111, 61363056, the Program for Liaoning Excellent Talents in University Grant (LR2012039), the Natural Science Foundation of Liaoning Province (2014020142).

References

[1] R.B. Bhatt, M. Gopal, On the compact computational domain of fuzzy rough sets, Pattern Recognit. Lett. 26 (2005) 1632–1640.
[2] D.G. Chen, L. Zhang, S.Y. Zhao, Q.H. Hu, P.F. Zhu, A novel algorithm for finding reducts with fuzzy rough sets, IEEE Trans. Fuzzy Syst. 20 (2) (2012) 385–389.
[3] C. Cornelis, R. Jensen, G. Hurtado, et al., Attribute select with fuzzy decision reducts, Inf. Sci. 177 (2007) 3–20.
[4] M. Dash, H. Liu, Consistency-based search in feature selection, Artif. Intell. 151 (1-2) (2003) 155–176.
[5] D. Dubois, H. Prade, Rough fuzzy sets and fuzzy rough sets, Int. J. Gen. Syst. 17 (1990) 191–208.
[6] W.A. Dudek, Y.B. Jun, Rough subalgebras of some binary algebras connected with logics, Int. J. Math. Math. Sci. 3 (2005) 437–447.
[7] M.A. Hall, Correlation-based feature selection for discrete and numeric class machine learning, in: Proc. 17th Int. Conf. Machine Learning, 2000, pp. 359–366.
[8] Q.H. Hu, D.R. Yu, Z.X. Xie, Information-preserving hybrid data reduction based on fuzzy-rough techniques, Pattern Recognit. Lett. 27 (5) (2006) 414–423.
[9] Q.H. Hu, D. Yu, J.F. Liu, C. Wu, Neighborhood-rough-set based heterogeneous feature subset selection, Inf. Sci. 178 (18) (2008) 3577–3594.
[10] Q. Hu, D. Yu, W. Pedrycz, D. Chen, Kernelized fuzzy rough sets and their applications, IEEE Trans. Knowl. Data Eng. 23 (11) (2011) 1649–1667.
[11] R Jensen, Q Shen, Fuzzy-rough attributes reduction with application to web categorization, Fuzzy Sets Syst. 141 (2004) 469–485.
[12] J.Y. Liang, F. Wang, C.Y. Dang, Y.H. Qian, A group incremental approach to feature selection applying rough set technique, IEEE Trans. Knowl. Data Eng. 26 (2) (2014) 294–304.
[13] G. Lin, J.Y. Liang, Y.H. Qian, J. Li, A fuzzy multigranulation decision-theoretic approach to multi-source fuzzy information systems, Knowl. Based Syst. 91 (2016) 102–113.
[14] T.Y. Lin, Neighborhood systems – application to qualitative fuzzy and rough sets, in: P.P. Wang (Ed.), Advances in machine intelligence and soft computing, Department of Electrical Engineering, Duke University, Durham, North Carolina, USA, 1997, pp. 132–155.
[15] T.Y. Lin, Granulation and nearest neighborhoods:Rough Set Approach, in: Granular Computing: An Emerging Paradigm, Physica-Verlag, Heidelberg, Germany, 2001, pp. 125–142.
[16] J.S. Mi, Y. Leung, H.Y. Zhao, T. Feng, Generalized fuzzy rough sets determined by a triangular norm, Inf. Sci. 178 (16) (2008) 3203–3213.
[17] Z. Pawlak, Rough sets, Int. J. Comput. Information Sci. 11 (5) (1982) 341–356.
[18] Y. Qian, J. Liang, D. Li, F. Wang, N. Ma, Approximation reduction in inconsistent incomplete decision tables, Knowl. Based Syst. 23 (5) (2010) 427–433.
[19] A. Mieszkowicz-Rolka, L. Rolka, Variable precision fuzzy rough sets, in: Transactions on Rough sets 1, LNCS-3100, Springer, Berlin, Cermany, 2004, pp. 144–160.
[20] C. Wang, Y. Qi, M. Shao, Q. Hu, D. Chen, Y. Qian, Y. Lin, A fitting model for feature selection with fuzzy rough sets, IEEE Trans. Fuzzy Syst. (2016), doi:10.1109/TFUZZ.2016.2574918.
[21] C. Wang, Q. He, D. Chen, Q. Hu, A novel method for attribute reduction of covering decision systems, Inf. Sci. 254 (2014) 181–196.
[22] W.Z. Wu, W.X. Zhang, Neighborhood operator systems and approximations, Inf. Sci. 144 (2002) 201–217.
[23] Y.Y. Yao, Relational interpretations of neighborhood operators and rough set approximation operators, Inf. Sci. 101 (1998) 239–259.
[24] X. Zhang, B. Zhou, P. Li, A general frame for intuitionistic fuzzy rough sets, Inf. Sci. 216 (2012) 34–49.
[25] H.Y. Zhang, S.Y. Yang, Ranking interval sets based on inclusion measures and application to three-way decisions, Knowl. Based Syst. 91 (2016) 62–70.
[26] S.Y. Zhao, E.C.C. Tsang, D.G. Chen, The model of fuzzy variable precision rough sets, IEEE Trans. Fuzzy Syst. 17 (2) (2009) 451–467.