Available online at www.sciencedirect.com
ScienceDirect
Pattern Recognition 40 (2007) 3509–3521
PATTERN RECOGNITION
THE JOURNAL OF THE PATTERN RECOGNITION SOCIETY
www.elsevier.com/locate/pr

Hybrid attribute reduction based on a novel fuzzy-rough model and information granulation

Qinghua Hu*, Zongxia Xie, Daren Yu

Harbin Institute of Technology, Harbin, Heilongjiang Province 150001, China

Received 2 April 2006; received in revised form 15 January 2007; accepted 15 March 2007

Abstract

Feature subset selection has become an important challenge in areas of pattern recognition, machine learning and data mining. As different semantics are hidden in numerical and categorical features, there are two strategies for selecting hybrid attributes: discretizing numerical variables or numericalize categorical features. In this paper, we introduce a simple and efficient hybrid attribute reduction algorithm based on a generalized fuzzy-rough model. A theoretic framework of fuzzy-rough model based on fuzzy relations is presented, which underlies a foundation for algorithm construction. We derive several attribute significance measures based on the proposed fuzzy-rough model and construct a forward greedy algorithm for hybrid attribute reduction. The experiments show that the technique of variable precision fuzzy inclusion in computing decision positive region can get the optimal classification performance. Number of the selected features is the least but accuracy is the best.

Keywords: Numerical feature; Categorical feature; Feature selection; Attribute reduction; Fuzzy set; Rough set; Inclusion degree

1. Introduction

Attribute reduction, also called feature subset selection is a common problem in pattern recognition, machine leaning and data mining as there usually are many candidate attributes collected to represent recognition problems. Databases expand quickly not only in the rows (objects) but also in the column (attributes) nowadays. Tens, hundreds even thousands of attributes are stored in databases in some real-world applications [1]. Some of attributes are irrelevant to the learning or recognition tasks. Experiments show irrelative attributes will deteriorate the performance of the learning algorithms for the curse of dimensionality, increase training and test times [2,3]. Feature subset selection can also facilitate data visualization and data understanding. What is more, measuring and storing all of the attributes relevant and irrelevant to the recognition problems is very expensive in practice. It is likely that the omission of some features will not seriously increase error probability. In such cases, the loss of optimality may not only be tolerable but even desirable relatively to the costs involved [4].

Roughly speaking, there are two strategies in attribute subset selection. One is called wrapper [5], which employs a learning algorithm to evaluate the selected attribute subsets. As evaluating the attributes subsets by training and test a classifier, wrapper is usually time consuming. The other is called filter, which selects attributes with a significance measure, independent of learning algorithms, such as distance [6], information gain [7], consistency [8], similarity [9,10] and dependency [11]. In essence, all the measures can be divided into two classes: the distance-based measures and consistency-based measures. Linear discriminant analysis (LDA), principle component analysis (PCA), neural networks [12] and SVM [13] are the representatives of algorithms based on distances. In this process, all attributes are considered as numerical. They are coded as integral numbers if there are some categorical features in data. However, methods based on consistency take all the attributes as symbolic values. The numerical attributes are discretized into several intervals and the intervals are assigned with a set of symbolic values [14–16]. Consistency measures do not attempt to maximize the class separability but try to retain the

* Corresponding author. Tel.: +86 45186413241252; fax: +86 45186413241221.
E-mail address: huqinghua@hcms.hit.edu.cn (Q. Hu).

0031-3203/$30.00 © 2007 Pattern Recognition Society. Published by Elsevier Ltd. All rights reserved.
doi:10.1016/j.patcog.2007.03.017

discriminating power of the data of original features. Rough set based attribute reduction presents systematic theoretic framework for consistency-based feature subset selection [17,18].

It is unreasonable to measure similarity or dissimilarity with distance metric as to categorical attributes. For example, as to outlook attribute, it takes values in set {sunny, rainy, overcast}. We can code the value set as 1, 2 and 3, respectively. However, we can also code them with 3, 2 and 1. It is nonsense to compute the distance between the coded values. On the other side, discretizing numerical attributes usually bring information loss because the degrees of membership of values to discretized values are not considered [19]. It is clear that some reduction algorithms for hybrid attributes should be developed. In order to deal with this problem, Tang and Mao [20] presented an error probability-based measure for mixed feature evaluation. For a mixed feature subset, the entire feature space is first divided into a set of homogeneous subspaces based on nominal features. The merit of the mixed feature subset is then measured based on sample distributions in the homogeneous subspaces spanned by continuous features. In Ref. [21] Pedrycz and Vukovich considered features to be granular rather than numerical. Shen and Jensen [18,22,23] generalized the dependency function defined in classical rough set model into the fuzzy case and presented a fuzzy-rough QUICKREDUCT algorithm. In Refs. [24,25] Bhatt and Gopal showed that QUICKREDUCT algorithm is not convergent on many real data sets due to its poorly designed stopping criteria; and the computational complexity of the algorithm increases exponentially with the number of input variables and in multiplication with the size of data patterns. They gave the concept of fuzzy-rough sets on compact computational domain, which is then utilized to improve computational efficiency. As Shannon's information entropy was introduced to search reducts in classical rough set model [26,27], Hu et al. extended the entropy to measure the information quantity in fuzzy sets [28] and applied the proposed measure to calculate the uncertainty in fuzzy-rough approximation spaces [29] and reduce hybrid data [30].

Granular computing has attracted much attention in the last decade. Both Pedrycz's work and fuzzy-rough set-based hybrid feature selection algorithms are involved with a basic idea, which is to generate a family of fuzzy information granules from numerical features and transform numerical attributes into fuzzy linguistic variables, which keep the semantics of the data and are easy to understand. Fuzzy information granulation and granular computing are important concepts in fuzzy set and rough set theories in recent years [31,32]. Fuzzy set, rough set and their combinations seem to be efficient tools for granular computing [33–36]. The approaches to generating fuzzy information granules from data [34,35,37,38], the models for granular computing [39–41], the applications of granular computing [21,42–44] were discussed, respectively. In this paper we will show a fuzzy-rough model for granular computing and hybrid data reduction.

The classical rough set model, proposed by Pawlak [45], is based on crisp equivalence relations and crisp equivalence classes. It is applicable to categorical attribute reduction and knowledge discovery. Categorical attributes partition the object set into some mutually exclusive crisp subsets, called elemental sets, or elemental information granules. Arbitrary subset X in the universe can be approximated by the union of the elemental information granules. The maximal union of the information granules, which objects are contained in X, is called the lower approximation of X. The minimal union of the information granules, which can contain the objects in X, is called the upper approximation of X. The categorical attributes yield a granularity of the universe, and then we approximate any concept in the specific granularity level. The finer the granulated space, the more precise the approximation. In order to deal with numerical and fuzzy attributes in information systems, rough set and fuzzy set are combined together [46–49]. Fuzzy information granules are the foundation stone of these models, fuzzy information granulation and fuzzy equivalence relations were introduced to form fuzzy granule systems. The fuzzy granules can be generated from numerical or fuzzy data by fuzzy partition [50], fuzzy clustering [21,51] and genetic algorithms [55].

In this paper, we do not discuss how to generate a family of fuzzy information granules from a hybrid data set. We will focus on studying the relations and structures of the fuzzy information granules. We present a novel fuzzy-rough model based on the inclusion degree [52,53] and show that the proposed model is a natural extension of Pawlak's model. Then we define the attribute significance measure based on the proposed model, which is applicable to both categorical and numeric attributes, and construct a greedy hybrid attribute reduction for classification analysis. The experiments show that the proposed method can keep or improve the classification power with very few features.

The rest of the paper is organized as follows. Section 2 shows some basic concepts on rough sets and fuzzy-rough sets. Section 3 presents the novel fuzzy-rough set model. The significance measure for hybrid features and reduction algorithm are introduced in Section 4. The experimental analysis is given in Section 5. Then conclusion comes in Section 6.

2. Fundamentals on Pawlak's rough sets and fuzzy-rough sets

Pawlak's definition on rough sets starts with an equivalence relation and a family of equivalence classes. A finite and nonempty universe of objects $U = \{x_1, x_2, \ldots, x_n\}$ is characterized with a collection of attributes. Each attribute generates an indiscernible relation R on U. Then $\langle U, R\rangle$ is called an approximation space. The equivalence classes $[x_i]_R$ are called elemental information granules in the approximation space. They form a family of concepts to approximate arbitrary subset of objects. Given an arbitrary subset $X \subseteq U$, one can define two unions of elemental information granules:

$$\begin{cases} \underline{R}X = \cup\{[x_i] | [x_i] \subseteq X\}, \\ \overline{R}X = \cup\{[x_i] | [x_i] \cap X \neq \emptyset\}. \end{cases}$$

Equivalently, they can also be written as

$$\begin{cases} \underline{R}X = \{x_i | [x_i] \subseteq X\}, \\ \overline{R}X = \{x_i | [x_i] \cap X \neq \emptyset\}, \end{cases} .$$

They are called lower and upper approximations of X in the approximation space. We say X is a definable set if $\underline{R}X = \overline{R}X$, otherwise, X is a rough set. Rough sets are the object subsets which cannot be precisely described by the corresponding elemental information granules.

There are two kinds of attributes as to a classification problem: condition A and decision D. Assume the objects are partitioned into N mutually exclusive crisp subsets $\{X_1, X_2, \ldots, X_N\}$ by decision D, where X_i corresponds the object subset with decision i. Given arbitrary subset $B \subseteq A$, then we can define the lower and upper approximations of the decision D as

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

where R is the equivalence relation induced by attributes B. $\underline{R}D$ is also called the positive region of D with respect to condition B, denoted as $POS_B(D)$.

A dependency function involving B and D is formulated as

$$\gamma = \frac{|POS_B(D)|}{|U|},$$

where $|\bullet|$ is the cardinality of a set. Dependency function reflects B's power to approximate D. $0 \leqslant \gamma \leqslant 1$. We say D completely depends on B if $\gamma = 1$. It means that the decision can be precisely described by the elemental information granules generated by attributes B. This function measures the significance of categorical attributes relative to the decision. In practice, the attributes may be numerical or fuzzy. Correspondingly, the relation and partition induced by these attributes are fuzzy. In this case, we are involved in approximating a fuzzy or crisp set with a family of fuzzy information granules.

Formally, a fuzzy classification problem can be described as follows. A set of fuzzy input and output attributes $\{P_1, P_2, \ldots, P_p\}$ and d are given to describe the objects $U = \{x_1, x_2, \ldots, x_n\}$. Each attribute is limited to a small set of fuzzy linguistic terms $A(P_i) = \{F_{ik} | k = 1, 2, \ldots, C_i\}$. Each object $x_i \in U$ is classified by a set of classes $A(Q) = \{F_l | l = 1, \ldots, C_Q\}$, where Q is a decision attribute and F_l can be a fuzzy set or a crisp set. One can generate a family of fuzzy information granules with P, where the fuzzy partition is defined as $U/P = \{F_{ik} | i = 1, \ldots, p;\ k = 1, \ldots, C_i\}$. Given arbitrary fuzzy set A in U, $u_A(x) : U \to [0, 1]$, $\forall x \in U$ and $F_{ik} \in U/P$, one define a tuple $\langle u_{\underline{A}}, u_{\overline{A}} \rangle$, where lower and upper approximation membership functions are defined as

$$u_{\underline{A}}(F_{ik}) = \inf_{x \in U} \max\{1 - u_{F_{ik}}(x),\ u_A(x)\},$$

$$u_{\underline{A}}(F_{ik}) = \sup_{x \in U} \min\{u_{F_{ik}}(x), u_A(x)\}.$$

The positive region of a fuzzy set F_l is the maximal membership degree with which a unique class can be classified by fuzzy set F_{ik}, written as

$$u_{POS}(F_{ik}) = \sup_{F_l \in A(Q)} \{u_{\underline{F_l}}(F_{ik})\}.$$

The membership of $x \in U$ to the fuzzy positive region is given by

$$u_{POS}(x) = \sup_{F_{ik} \in A(P_i)} \min\{u_{F_{ik}}(x), u_{POS}(F_{ik})\}.$$

With the definition of fuzzy positive regions, one can compute the dependence function as

$$u_P(Q) = \frac{\sum_{x \in U} u_{POS}(x)}{|U|}.$$

In [24,25], Bhatt proposed the concept of fuzzy-rough sets in the compact computational domain based on fuzzy t-norm and t-conorm, where the membership functions of lower and upper approximations are defined as

$$u_{\underline{A}}(F_{ik}) = \begin{cases} \inf\limits_{x \in D_{\underline{A}}(F_{ik})} \max\{u_{\overline{F_{ik}}}(x), u_A(x)\}; & D_{\underline{A}}(F_{ik}) \neq \emptyset\}, \\ 1, \quad D_{\underline{A}}(F_{ik}) \neq \emptyset, \end{cases}$$

$$u_{\overline{A}}(F_{ik}) = \begin{cases} \sup\limits_{x \in D_{\underline{A}}(F_{ik})} \min\{u_{F_{ik}}(x), u_A(x)\}; & D_{\overline{A}}(F_{ik}) \neq \emptyset\}, \\ 0, \quad D_{\overline{A}}(F_{ik}) = \emptyset, \end{cases}$$

where $D_{\underline{A}}(F_{ik})$ and $D_{\overline{A}}(F_{ik})$ are compact computational domains for lower and upper approximation membership functions, defined as

$$\begin{cases} D_{\underline{A}}(F_{ik}) = \{x \in U | u_{F_{ik}}(x) \neq 0 \wedge u_A(x) \neq 1\}, \\ D_{\overline{A}}(F_{ik}) = \{x \in U | u_{F_{ik}}(x) \neq 0 \wedge u_A(x) \neq 0\}. \end{cases}$$

3. A new fuzzy-rough model and attribute properties in fuzzy-rough approximation spaces

In most of the cases, a classification problem can be formulated as $\langle U, A^r \cup A^c, d \rangle$, where U is the universe of objects, A^r is the subset of condition attributes with numerical values and A^c is categorical attributes; d is the decision with N finite values. The learning task is to build a mapping $f : A^r \cup A^c \to d$ with the given objects. Here, the universe is partitioned into N crisp equivalence classes by the decision; whereas, $A^r \cup A^c$ generate some fuzzy information granules in the universe. The task is to approximate the crisp decision classes with the fuzzy information granules.

3.1. Fuzzy-rough set model

Definition 1. Assumed R is a fuzzy equivalence relation induced by a numerical attribute or fuzzy attribute. $\forall x, y, z \in U$, it satisfies:

(1) reflexivity: $R(x, x) = 1$;
(2) symmetry: $R(x, y) = R(y, x)$; and
(3) transitivity: $R(x, z) \geqslant \bigvee_y (R(x, y) \wedge R(y, z))$,

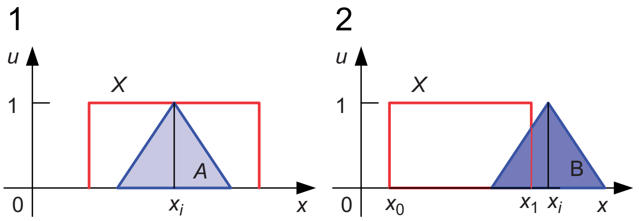

Fig. 1. The lower approximation and upper approximation of a crisp set.

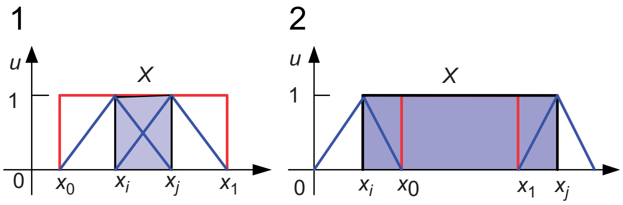

Fig. 2. Lower and upper approximation regions.

where $\wedge$ and $\vee$ mean the operations "min" and "max", respectively. The relation can be written as a matrix as

$$M(R) = \begin{bmatrix} r_{11} & r_{12} & \cdots & r_{1n} \\ r_{21} & r_{22} & \cdots & r_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ r_{n1} & r_{n2} & \cdots & r_{nn} \end{bmatrix}.$$

Definition 2. The fuzzy equivalence class $[x_i]_R$ of x_i induced by the relation R is defined as

$$[x_i]_R = \frac{r_{i1}}{x_1} + \frac{r_{i2}}{x_2} + \cdots + \frac{r_{in}}{x_n},$$

where "+" means the union. Obviously, $[x_i]_R$ is a fuzzy information granule. It is easy to find that the definition of fuzzy equivalence classes is a natural extension of crisp one. If the attribute is categorical, the relation, relation matrix and equivalence class will degrade to the classical case.

Definition 3. R is a fuzzy equivalence relation on the universe$U$; $X \subseteq U$ is a crisp subset of objects. Then the lower and upper approximations of X can be defined as

$$\underline{R}X = \{x_i | [x_i]_R \subseteq X, x_i \in U\},$$

$$\overline{R}X = \{x_i | [x_i]_R \cap X \neq \emptyset, x_i \in U\},$$

where $A \subseteq B$ means $\forall x \in U, u_A(x) \leqslant u_B(x)$.

Correspondingly, the boundary region of X is given as

$$BN(X) = \overline{R}X - \underline{R}X.$$

In Fig. 1, X is the crisp set. A and B are two fuzzy sets generated by a fuzzy relation. $A \subseteq X$, we say x_i is the object belonging to the lower approximation of X. $B \not\subset X$ and $B \cap X \neq \emptyset$, then we say x_i belongs to the upper approximation.

In Fig. 2, X is the crisp set of points in the interval $[x_0, x_1]$. We granulate the x space with a family of triangle fuzzy sets.

In Fig. 2(1), the fuzzy information granules with the center x_i and x_j is the limits of the lower approximation of X because two fuzzy sets are completely included in X and the fuzzy sets beyond them are not. Then the lower approximation is the shadow in Fig. 2(1). Similarly, the upper approximation is the shadow in Fig. 2(2).

It is easy to find that the lower and upper approximations of a crisp X are crisp, which will bring much convenience in computing them in real-world applications and overcome the problem proposed in Ref. [25].

The inclusion operator "$\subseteq$" for fuzzy sets was introduced by Zadeh [54], called Zadeh's inclusion. It is too strong in real-world applications. Some generalizations were proposed.

Given a fuzzy set A, we say x *α-belong-to* A, when and only when $x \in A_\alpha$ where A_α is the α-cut of A. $\forall x \in U$ if we have $x \in (\overline{A} \cup B)_\alpha$ we say that fuzzy set A *α-belong-to* B, denoted by $A - <_\alpha B$.

Equivalently, the weak inclusion can be defined as follows:

$$\forall x \in U, \quad \max(1 - u_A(x), u_B(x)) \geqslant \alpha.$$

On the other hand, we can introduce inclusion degree function to loosen Zadeh's inclusion.

Definition 4. Let A and B be two fuzzy sets in the universe U, the inclusion $I(A, B)$ is defined as

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

where $||A|| = |A|/|U|$, $|A| = \sum_{x \in U} u_A(x)$.

We denote $A \subset_\varepsilon B$, meaning $I(A, B) \geqslant \varepsilon$.

Definition 5. Based on the inclusion function, the lower, upper approximations and boundary region of $X \subseteq U$ can also be defined as

$$\underline{R}X = \{x_i | I([x_i]_R, X) = 1, x_i \in U\},$$

$$\overline{R}X = \{x_i | I([x_i]_R, X) > 0, x_i \in U\},$$

$$BN(X) = \{x_i | 1 > I([x_i]_R, X) > 0, x_i \in U\}.$$

Definition 6. The variable precision lower and upper approximations of a crisp subset X by a family of fuzzy information granules are defined as

$$\underline{R}_k X = \{x_i | I([x_i]_R, X) \geqslant k, x_i \in U\},$$

$$\overline{R}_l X = \{x_i | I([x_i]_R, X) > l, x_i \in U\},$$

where $1 \geqslant k \geqslant 0.5$, $0.5 > l \geqslant 0$.

And the variable precision boundary region is

$$BN_{kl} = \overline{R}_l X - \underline{R}_k X = \{x_i | k \geqslant I([x_i]_R, X) > l, x_i \in U\}.$$

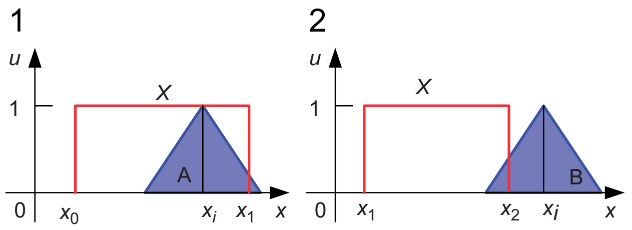

Fig. 3. Variable precision fuzzy-rough sets.

The variable precision fuzzy-rough model allows partial inclusion, partial precision, partial certainty, which is the coral advantage of fuzzy information granulation [31], and simulates the remarkable human ability to make rational decisions in an environment of imprecision (see also Fig. 3).

The forms of the proposed definitions of fuzzy-rough sets are quite similar to Pawlak's one. The unique difference is that the elemental sets or elemental information granules are generated with a fuzzy equivalence relation, therefore they are fuzzy. This generalization of rough sets make the theory applicable to deal with hybrid data learning and classification, which is most often the case that the values of attributes may be both crisp and real-valued. It the same time, the proposed model is easy to implement and understand compared with those in Refs. [18,24,46,49].

3.2. Hybrid information systems and hybrid decision tables

A hybrid information system can be written as $\langle U, A = A^r \cup A^c, V = V^r \cup V^c, f\rangle$, where U is the set of objects, V^r is the domain of real numbers for real-valued attributes A^r, V^c is the domain of categorical values for categorical attributes A^c, f is an information function $f : U \times A \to V$. As to classification problems, there is a decision variable in the information system. We called the information system as a decision table in this case.

A categorical attribute can induce a crisp equivalence relation on the universe and generate a family of crisp information granules, whereas a numerical attribute will give a fuzzy equivalence relation and form a set of fuzzy information granules. As crisp information granules are a special case of fuzzy ones, we will consider all of them as fuzzy ones in the following.

Given a hybrid information system $\langle U, A, V, f\rangle$, $B, B_1, B_2 \subseteq A$, we means R_B as the relation induced by the attribute subset B. Then we have

(1) $R_B = \cap_{a\in B} R_a$;
(2) $R_{B_1 \cup B_2} = R_{B_1} \cap R_{B_2}$;
(3) $[x]_B = \cap_{a\in B}[x]_a$;
(4) if $B_1 \subseteq B_2$, $R_{B_1} \supseteq R_{B_2}$; and
(5) if $B_1 \subseteq B_2$, $[x]_{B_1} \supseteq [x]_{B_2}$.

Definition 7. Given a hybrid decision table $\langle U, A \cup D, V, f\rangle$, $X_1, X_2, \ldots, X_N$ are the object set with decision 1 to N, $[x_i]_B$ is the fuzzy information granules including x_i and generated with attributes $B \subseteq A$. Then the lower and upper approximations of the decision D are defined as

$$\underline{B}D = \{\underline{B}X_1, \underline{B}X_2, \ldots, \underline{B}X_N\},$$

$$\overline{B}D = \{\overline{B}X_1, \overline{B}X_2, \ldots, \overline{B}X_N\},$$

where

$$\underline{B}X = \{x_i | I([x_i]_B, X) = 1, x_i \in U\},$$

$$\overline{B}X = \{x_i | I([x_i]_B, X) > 0, x_i \in U\}.$$

The decision boundary region of D with respect to attributes B is defined as

$$BN(D) = \overline{B}D - \underline{B}D.$$

Definition 8. The lower approximation of decision D also called positive region, denoted as $POS_B(D)$. As to a classification problem, one hopes the positive region is as great as possible and the boundary region is as little as possible. We define a dependency function as the ratio of positive region to the universe as follows:

$$\gamma = \frac{|POS_B(D)|}{|U|}.$$

Obviously, $0 \leqslant \gamma \leqslant 1$. The dependence function reflects the approximation power of a set of hybrid condition attributes and it can be used as the significance of attribute set.

Theorem 1. *$\langle U, A \cup D, V, f\rangle$ is a hybrid decision table; A is the set of hybrid condition attributes, D is the decision. $B_1 \subseteq B_2 \subseteq A$, then we have*

$$POS_{B_1}(D) \leqslant POS_{B_2}(D) \quad \textit{and} \quad \gamma_{B_1}(D) \leqslant \gamma_{B_2}(D).$$

Theorem 2. *$\langle U, A \cup D, V, f\rangle$ is a hybrid decision table. If the decision table is compatible, namely, $R_A \subseteq R_D$, we have $\gamma_A(D) = 1$.*

Definition 9. Given an hybrid decision table $\langle U, A \cup D, V, f\rangle$, $B \subseteq A$, we say attribute set B is a relative reduct if

(1) $\gamma_B(D) = \gamma_A(D)$ and
(2) $\forall a \in B, \gamma_B(D) > \gamma_{B-a}(D)$.

The first condition guarantees that the reduct has the same approximation power as the whole attribute set, and the second condition guarantees there is no redundant or superfluous attribute in the reduct.

As to the variable precision fuzzy-rough model, the lower and upper approximations, boundary and dependence function also can be defined as so.

Definition 10. Given a hybrid decision table $\langle U, A \cup D, V, f\rangle$, $X_1, X_2, \ldots, X_N$ are the objects with decision 1 to N, $[x_i]_B$ is the fuzzy information granules including x_i and generated with attributes $B \subseteq A$, Then the k-lower and l-upper approximations of the decision D are defined as

$$\underline{B}_k D = \{\underline{B}_k X_1, \underline{B}_k X_2, \ldots, \underline{B}_k X_N\},$$

$$\overline{B}_l D = \{\overline{B}_l X_1, \overline{B}_l X_2, \ldots, \overline{B}_l X_N\},$$

where

$$\underline{B}_k X = \{x_i \mid I([x_i]_B, X) \geqslant k, x_i \in U\}, \quad 1 \geqslant k \geqslant 0.5,$$

$$\overline{B}_l X = \{x_i \mid I([x_i]_B, X) > l, x_i \in U\}, \quad 0.5 > l \geqslant 0.$$

The k–l decision boundary region of D with respect to attributes B is defined as

$$BN_{kl}(D) = \overline{B}_l D - \underline{B}_k D.$$

Here $\underline{B}_k D$ are also called k–l decision positive region of D with respect to B, denoted as $POS_B^{kl}(D)$.

Definition 11. k–l dependency of D on B is defined as

$$\gamma_B^{kl}(D) = \frac{|POS_B^{kl}(D)|}{|U|}.$$

Correspondingly, we also have the following theorems.

Theorem 3. *$\langle U, A \cup D, V, f\rangle$ is a hybrid decision table; A is the set of hybrid condition attributes, D is the decision. $B_1 \subseteq B_2 \subseteq A$, then we have*

$$POS_{B_1}^{kl}(D) \leqslant POS_{B_2}^{kl}(D) \quad \text{and} \quad \gamma_{B_1}^{kl}(D) \leqslant \gamma_{B_2}^{kl}(D).$$

Theorem 4. *$\langle U, A \cup D, V, f\rangle$ is a hybrid decision table. If the decision table is consistent, namely, $R_A \subseteq R_D$, we have $\gamma_A^{kl}(D) = 1$.*

Definition 12. Giving a hybrid decision table $\langle U, A \cup D, V, f\rangle$, $B \subseteq A$, we say attribute set B is a k–l relative reduct if:

(1) $\gamma_B^{kl}(D) = \gamma_A^{kl}(D)$ and
(2) $\forall a \in B, \gamma_B^{kl}(D) > \gamma_{B-a}^{kl}(D)$.

Compared with Pawlak's rough set model, we can find that the forms of the proposed fuzzy-rough models are quite similar to the classical. However, the model is generalized and can deal with numeric and fuzzy attributes in the information systems.

4. Attribute reduction algorithms for hybrid data

As mentioned above, the dependency function measures the approximation power of a set of condition attributes. In rough set framework, attribute reduction is to find some attribute subsets which have the minimal attributes and the maximal approximation power. To construct an attribute reduction algorithm three problems should be made clear: attribute evaluating measure, search strategy and stop criterion. What is more, it is easy to induce a crisp equivalence relation with categorical attribute; however, as to numerical feature, it is not so straightforward. We should construct an algorithm to generate fuzzy information granules from the data with hybrid attributes. In this section we will deal with the problems.

4.1. Significance measures for hybrid attributes

The dependency function calculates the approximating power of an attribute set. It can be used as an attribute significance measure.

Definition 13. Given a hybrid decision table $\langle U, A \cup D, V, f\rangle$, $B \subseteq A$, $\forall a \in B$, one can define the significance of a in B as

$$Sig_1(a, B, D) = \gamma_B(D) - \gamma_{B-a}(D).$$

Note that an attribute's significance is the function of three variables: a, B and D. An attribute a may be of great significance in B_1 but of little significance in B_2. What is more, the attribute's significance will be different for each decision if they are multiple decision attributes in a decision table.

The above definition is applicable to backward feature selection. Similarly, a measure applicable to forward selection can be defined as follows.

Definition 14. Given a hybrid decision table $\langle U, A \cup D, V, f\rangle$, $B \subseteq A$, $\forall a \in A - B$, one can define the significance of a in B as

$$Sig_2(a, B, D) = \gamma_{B \cup a}(D) - \gamma_B(D).$$

As $0 \leqslant \gamma_B(D) \leqslant 1$ and $\forall a \in B$: $\gamma_B(D) \geqslant \gamma_{B-a}(D)$, we have

$$0 \leqslant Sig_1(a, B, D) \leqslant 1, \quad 0 \leqslant Sig_2(a, B, D) \leqslant 1.$$

We say attribute a is *superfluous* in B with respect to D if $Sig_1(a, B, D) = 0$, otherwise a is indispensable.

As pointed out before, the definitions of lower and upper approximations in fuzzy-rough model is too strict to tolerance noise in data. The variable precision fuzzy-rough model simulates the human ability making rational decision in the conditions of imprecision. This is just Zadeh's intention in proposing fuzzy information granulation [31] and computing with words [40]. Therefore, the variable precision fuzzy-rough model is an efficient implementation of granular computing. It will have good robustness and will work well in the environment with noise.

Two definitions of attribute significance can be defined based on the variable precision fuzzy-rough model.

Definition 15. Given a hybrid decision table $\langle U, A \cup D, V, f\rangle$, $B \subseteq A$, $\forall a \in B$, one can define the k–l significance of a in B as

$$Sig_3^{kl}(a, B, D) = \gamma_B^{kl}(D) - \gamma_{B-a}^{kl}(D).$$

Definition 16. Given a hybrid decision table $\langle U, A \cup D, V, f\rangle$, $B \subseteq A$, $\forall a \in A - B$, one can define the k–l significance of a in B as

$$Sig_4^{kl}(a, B, D) = \gamma_{B \cup a}^{kl}(D) - \gamma_B^{kl}(D),$$

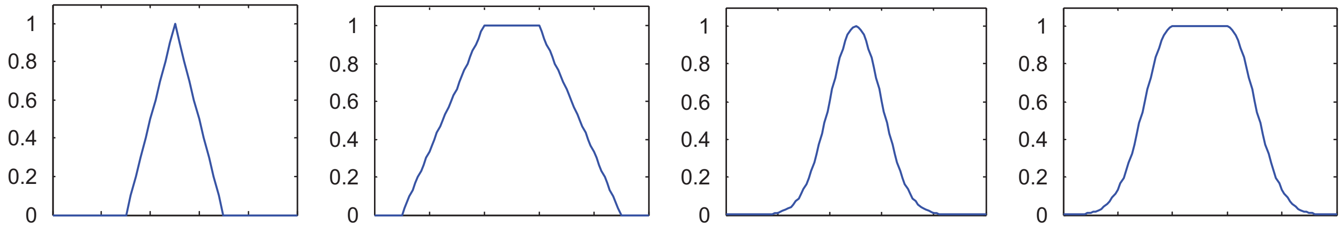

Fig. 4. Some similarity relation functions for numeric data.

where k and l are two thresholds, which reflect users' tolerance degrees of noise. The less the k and l, the more the users can tolerance noise.

This section presents four measures for computing the significance of hybrid attribute set. We also can construct some measures based on the size of boundary region or upper approximations.

4.2. Generating fuzzy information granules with hybrid attributes

In Ref. [31] Zadeh suggested that among the basic concepts which underlie human concept there are three that stand out in importance: granulation, organization and causation. Granulation of a set of objects U results in a collection of granules of U, with a granule being a clump of objects which are drawn together by indistinguishability, similarity, proximity or functionality. Models of information granulation in which the granules are crisp play important roles in a wide variety of methods and approaches, such as interval analysis, quantization and rough set theory. However, crisp information granulation fails to reflect the fact that in much of human reasoning and concept formation the granules are fuzzy rather than crisp. In human cognition, fuzziness of granules is a direct consequence if fuzziness of the concepts of indistinguishability, similarity, proximity and functionality. Fuzzy information granulation underlies the remarkable human ability to make rational decisions in conditions of imprecision, partial knowledge, partial certainty and partial truth.

Fuzzy information granulation plays an important role in human reasoning with uncertainty. As to categorical attributes, we can only generate a family of crisp equivalence relations and crisp equivalence information granules. However, we can produce a series of fuzzy information granules from numerical data with some granulating techniques. There have been some methods to find fuzzy information granules from data. In Ref. [33] several definitions of information granules are proposed based on equivalence relations, tolerance relations and so on. The concepts of elementary granules, sequences of granules, sets of granules, dynamic granules and labeled figure granules were given. Oh and Pedrycz introduced GA hybrid scheme to conduct fuzzy information granulation and to guarantee both global optimization and local convergence [55]. A series of work introduced fuzzy clustering for information granulation, where FCM based on 2-norm [21], FCM based on Tchebyschev distance [51], fuzzy-rough clustering [56] were studied. Here we will introduce a simple method to generate fuzzy equivalence relations and fuzzy information granules. There are two steps. First we generate a fuzzy similarity relation from the data with a symmetric function. Then transform the similarity relation into an equivalence one.

$\langle U, A \cup d, V, f \rangle$ is a hybrid decision table. No matter object set is described by nominal attributes or numeric features, the relations between the objects can be denoted by a relation matrix: $M(R) = (r_{ij})_{n\times n}$.

$B_1 \in A$ is a nominal attribute set, then

$$r_{ij} = \begin{cases} 1, & f(x_i, a) = f(x_j, a), \forall a \in B_1, \\ 0 & \text{otherwise.} \end{cases}$$

If attribute a is a numeric attribute, the value of the relation between x_i and x_j can be computed with a symmetric function:

$$r_{ij} = f(|x_i - x_j|),$$

where $|x_i - x_j|$ means the Euclidean distance between x_i and x_j, and function f should satisfy that:

(1) $f(0) = 1$, $f(\infty) = 0$ and $f(\bullet) \in [0, 1]$; and
(2) $x \geqslant y$: $f(x) \leqslant f(y)$.

A fuzzy similarity relation matrix will be produced by the function because relation R satisfies reflexivity and symmetry. Employing a max–min closure operation, we can get a fuzzy equivalence relation $M(R) = (r_{ij})_{n\times x}$ [57] (see also Fig. 4).

The fuzzy information granule induced by relation R and including x_i is

$$[x_i]_R = \frac{r_{1i}}{x_1} + \frac{r_{2i}}{x_2} + \cdots + \frac{r_{ni}}{x_n}.$$

4.3. Greedy algorithm for hybrid data reduction

The objective of rough set-based attribute reduction is to find a subset of attributes which has the same discriminating power as the original data and without redundancy. Although there are usually multiple reducts for a given decision table, in most of applications, it is enough to find one of them. With the measures of attributes, greedy search algorithms for attribute reduction can be constructed.

Two search strategies can be introduced. One is forward search and the other is backward search. The forward search starts with a nonempty set, and adds one or several attributes with great significances into a pool each time until the dependence does not increase. However, the backward search begins with the whole attribute set and deletes one or several features with the significance zero until the dependence decreases. Formally, a forward algorithm can be formulated as follows.

Algorithm. Forward attribute reduction based on variable precision fuzzy-rough model (FAR-VPFRS).

```
Input: Hybrid decision table $\langle U, A^c \cup A^r \cup d, V^c \cup V^r, f \rangle$ and Threshold $k // A^c$ and $A^r$ are categorical and numerical attributes
    $//k$ is the threshold for computing the lower approximations
    Output: One reduct red.
    Step 1: $\forall a \in A$ :compute the equivalence relation$R_a$;
    Step 2: $\phi \to red$; // red is the pool to contain the selected attributes
    Step 3: For each $a_i \in A - red$
    Compute $SIG(a_i, B, D) = \gamma^{kl}_{red \cup a}(D) - \gamma^{kl}_{red}(D)$, // Here we define $\gamma^{kl}_{\emptyset}(D) = 0$
    end
    Step 4: Select the attribute $a_k$ which satisfies:
$SIG(a_k, B, D) = \max_i(SIG(a_i, red, B))$
    Step 5: If $SIG(a_k, B, D) > 0$,
$red \cup a_k \to red$
go to step 3
    else
    return red
    Step 6: end
```

If there are N condition attributes, the time complexity for computing relation is N, the worst search time for a reduct is $N \times N$. The overall time complexity of the algorithm is $O(N^2)$.

5. Experimental analysis

The ability of classical rough set theory to categorical attribute reduction has been shown in other literatures [17]. The objective of these experiments is to show the power of the proposed method to select numerical or hybrid attributes. The data used in the experiments are outlined in Table 1. We can find that there are some numerical features in all of the databases. There are also some data sets with categorical and numerical attributes in the same time.

Two classical classification learning algorithms CART and RBF-SVM are introduced to evaluate the selected attributes. Firstly we normalize the numerical attribute x into the interval [0, 1] with

$$a' = \frac{a - a_{\min}}{a_{\max}}.$$

The value of the fuzzy similarity degree r_{ij} between objects x_i and x_j with respect to numerical attribute a is computed as

$$r_{ij} = \begin{cases} 1 - 4 \times |x_i - x_j|, & |x_i - x_j| \leqslant 0.25, \\ 0 & \text{otherwise.} \end{cases}$$

As $r_{ij} = r_{ji}$ and $r_{ii} = 0$, $0 \leqslant r_{ij} \leqslant 1$, the matrix $M = (r_{ij})_{n \times n}$ is a fuzzy similarity relation. We can get a fuzzy equivalence relation from M with max–min transitivity operation. In practice the operation cannot be conducted and we directly search reducts with a similarity relations.

As to the classical rough sets-based feature selection, numerical attributes should be discretized before selection. In order to compare these methods with the proposed one, fuzzy c-means clustering (FCM) is introduced to discretize numerical attributes. We conduct the reduction algorithm on the discretized decision tables and get a series of reducts, and then used the corresponding numerical attributes to construct classifiers. CART and SVM are introduced to evaluate the selected features. All of the results are obtained with 10-fold cross validation. Table 2 shows the comparisons of numbers of selected features and accuracies with the discretization method. Table 3 shows the comparison of the fuzzy information entropy-based method [30], where $N1$ and $N2$ are the numbers of attributes in the original data and reducts, respectively. Accuracy 1 and accuracy 2 are the classification accuracies with the original data and the reduced data, respectively.

It is easy to find from Table 2 that some of data sets obtain higher classification accuracies when some attributes are deleted. However, there are also some data sets where the accuracies greatly decrease. Especially as to data diab and heart, there are no attributes selected as all of the single attributes get dependency zero, no attribute can be selected in the

Table 1
Data description

	Data set	Abbreviation	Samples	Numerical features	Categorical features	Classes
1	Australian credit approval	crd	690	6	9	2
2	Pima indians diabetes	diab	768	8	0	2
3	Ecoli	ecoli	336	5	2	7
4	Heart disease	heart	270	7	6	2
5	Ionosphere	iono	351	34	0	2
6	Sonar, mines vs. rocks	sonar	208	60	0	2
7	Small soybean	soy	47	35	0	4
8	Wisconsin diagnostic breast cancer	wdbc	569	30	0	2
9	Wisconsin prognostic breast cancer	wpbc	198	33	0	2
10	Wine recognition	wine	178	13	0	3

Table 2
Feature selection based on classical rough set model where numerical attributes are discretized

Data	Feature		CART		SVM	
	$N1$	$N2$	Accuracy 1	Accuracy 2	Accuracy 1	Accuracy 2
crd	15	12	0.8217 ± 0.0459	0.8274 ± 0.1398	0.8144 ± 0.0718	0.8058 ± 0.0894
diab	8	0	0.7227 ± 0.0512	0.0000 ± 0.0000	0.7747 ± 0.0430	0.0000 ± 0.0000
ecoli	7	1	0.8197 ± 0.0444	0.4262 ± 0.0170	0.8512 ± 0.0591	0.4262 ± 0.0170
heart	13	0	0.7407 ± 0.0630	0.0000 ± 0.0000	0.8111 ± 0.0750	0.0000 ± 0.0000
iono	34	10	0.8755 ± 0.0693	0.9089 ± 0.0481	0.9379 ± 0.0507	0.9348 ± 0.0479
sonar	60	6	0.7207 ± 0.1394	0.6926 ± 0.0863	0.8510 ± 0.0948	0.7074 ± 0.1004
soy	35	2	0.9750 ± 0.0791	1.0000 ± 0.0000	0.9300 ± 0.1135	1.0000 ± 0.0000
wdbc	30	8	0.9050 ± 0.0455	0.9351 ± 0.0339	0.9808 ± 0.0225	0.9649 ± 0.0183
wpbc	33	7	0.6963 ± 0.0826	0.6955 ± 0.1018	0.7779 ± 0.0420	0.7837 ± 0.0506
wine	13	4	0.8986 ± 0.0635	0.8972 ± 0.0741	0.9889 ± 0.0234	0.9486 ± 0.0507
Average	24.80	5	0.8176	0.6383	0.8718	0.6571

Table 3
Feature selection based on fuzzy information entropy

Data	Feature		CART		SVM	
	$N1$	$N2$	Accuracy 1	Accuracy 2	Accuracy 1	Accuracy 2
crd	15	13	0.8217 ± 0.0459	0.8144 ± 0.1416	0.8144 ± 0.0718	0.8144 ± 0.0718
diab	8	8	0.7227 ± 0.0512	0.7213 ± 0.0404	0.7747 ± 0.0430	0.7747 ± 0.0430
ecoli	7	7	0.8197 ± 0.0444	0.8197 ± 0.0444	0.8512 ± 0.0591	0.8512 ± 0.0591
heart	13	9	0.7407 ± 0.0630	0.7593 ± 0.0766	0.8111 ± 0.0750	0.8074 ± 0.0488
iono	34	13	0.8755 ± 0.0693	0.9068 ± 0.0564	0.9379 ± 0.0507	0.9462 ± 0.0365
sonar	60	12	0.7207 ± 0.1394	0.7160 ± 0.0857	0.8510 ± 0.0948	0.8271 ± 0.0902
soy	35	2	0.9750 ± 0.0791	1.0000 ± 0.0000	0.9300 ± 0.1135	1.0000 ± 0.0000
wdbc	30	17	0.9050 ± 0.0455	0.9193 ± 0.0318	0.9808 ± 0.0225	0.9702 ± 0.0248
wpbc	33	17	0.6963 ± 0.0826	0.7103 ± 0.1092	0.7779 ± 0.0420	0.8087 ± 0.0601
wine	13	9	0.8986 ± 0.0635	0.9097 ± 0.0605	0.9889 ± 0.0234	0.9833 ± 0.0268
Average	24.80	10.70	0.8176	0.8277	0.8718	0.8783

first loop. The results shown in Table 3 are more robust than those in Table 2. Although there are more attributes selected in the reducts, the classifications have been improved compared with discretization methods.

As to variable precision neighborhood rough set model, we first specify the threshold $k=1$ and conduct the algorithm FAR-VPFRS. We get reducts of the data. The numbers of selected attributes and classification performances are shown in Table 4.

From Table 4, we can find that the features in *iono*, *sonar*, *soy* are greatly reduced. Especially for *soy*, only two attributes are preserved, and the accuracy is greatly improved at the same time. However, as to data *diab*, *ecoli*, heart and *wine*, almost all of the attributes cannot be reduced by the algorithm. As a whole, the numbers of the attributes cannot be greatly reduced when the algorithm takes 1 as the threshold.

According to the idea of granular computing, partial certainty and partial inclusion will be more robust to noise and uncertainty in the data. Therefore, we do not limit the threshold $k = 1$ because it is too strict for computing the fuzzy inclusion. We try $k = 0.5$–1 with step 0.05 and then we show the maximal classification accuracies, the corresponding thresholds and numbers of selected attributes are given in Tables 5 and 6.

Table 5 presents the classification results with CART algorithm, where k is the threshold with the maximal accuracy. We can find that the features are substantively deleted in most of data sets. As to data *crd*, *soy*, *wpbc*, *wdbc* and *wine*, only a few attributes are selected in the reducts, which allows the classification problem visualization. Surprisingly, the classification performances are improved for all the data sets except *diab*.

Table 6 shows the results with SVM. Comparing Tables 5 and 6, we can find SVM learning algorithm requires more features to get good performance than CART. Accordingly, the average optimal threshold is 0.77 for CART, whereas 0.83 for SVM.

Fig. 5 shows that classification accuracy varies with the specified threshold as to four data sets *crd*, *ecoli*, *sonar* and *wine*. We can find the performance does not monotonously increase with the threshold. There are optimal points for feature selec-

Table 4
Feature selection based on the fuzzy-rough model with threshold $k = 1$

Data	Feature		CART		SVM	
	$N1$	$N2$	Accuracy 1	Accuracy 2	Accuracy 1	Accuracy 2
crd	15	9	0.8217 ± 0.0459	0.8321 ± 0.0667	0.8392 ± 0.0356	0.8392 ± 0.0405
diab	8	7	0.7227 ± 0.0512	0.7253 ± 0.0485	0.7745 ± 0.0430	0.7747 ± 0.0430
ecoli	7	7	0.8197 ± 0.0444	0.8168 ± 0.0429	0.8512 ± 0.0591	0.8512 ± 0.0591
heart	13	12	0.7407 ± 0.0630	0.7407 ± 0.0630	0.8111 ± 0.0750	0.8074 ± 0.0694
iono	34	23	0.8755 ± 0.0693	0.8980 ± 0.0525	0.9379 ± 0.0507	0.9461 ± 0.0366
sonar	60	30	0.7207 ± 0.1394	0.7062 ± 0.1081	0.8510 ± 0.0948	0.8410 ± 0.0691
soy	35	2	0.9750 ± 0.0791	1.0000 ± 0.0000	0.9300 ± 0.1135	1.0000 ± 0.0000
wdbc	30	24	0.9050 ± 0.0455	0.9122 ± 0.0296	0.9808 ± 0.0225	0.9790 ± 0.0215
wpbc	33	26	0.6963 ± 0.0826	0.6547 ± 0.1093	0.7779 ± 0.0420	0.7934 ± 0.0479
wine	13	13	0.8986 ± 0.0635	0.8986 ± 0.0635	0.9889 ± 0.0234	0.9889 ± 0.0234
Average	24.80	15.30	0.8176	0.8185	0.8742	0.8821

Table 5
Comparison of the best accuracy with different thresholds (CART)

Data	$N1$	$N2$	Accuracy 1	Accuracy 2	k
crd	15	3	0.8217 ± 0.0459	0.8639 ± 0.0499	0.65
diab	8	8	0.7227 ± 0.0512	0.7253 ± 0.0485	1
ecoli	7	6	0.8197 ± 0.0444	0.8173 ± 0.0554	0.75
heart	13	6	0.7407 ± 0.0630	0.8259 ± 0.0742	0.6
iono	34	23	0.8755 ± 0.0693	0.9065 ± 0.0490	0.95
sonar	60	28	0.7207 ± 0.1394	0.7400 ± 0.1020	0.85
soy	35	2	0.9750 ± 0.0791	1.0000 ± 0.0000	0.5
wdbc	33	2	0.9050 ± 0.0455	0.9298 ± 0.0261	0.95
wpbc	30	1	0.6963 ± 0.0826	0.7484 ± 0.0862	0.75
wine	13	2	0.8986 ± 0.0635	0.9049 ± 0.0451	0.7
Average	24.80	8.10	0.8176	0.8462	0.77

Table 6
Comparison of the best accuracy with different thresholds (SVM)

Data	$N1$	$N2$	Accuracy 1	Accuracy 2	k
crd	15	3	0.8392 ± 0.0356	0.8639 ± 0.0499	0.65
diab	8	8	0.7745 ± 0.0430	0.7745 ± 0.0430	0.85
ecoli	7	7	0.8512 ± 0.0591	0.8512 ± 0.0591	0.80
heart	13	13	0.8111 ± 0.0750	0.8111 ± 0.0745	0.80
iono	34	21	0.9379 ± 0.0507	0.9518 ± 0.0332	0.95
sonar	60	30	0.8510 ± 0.0948	0.8410 ± 0.0691	1.0
soy	35	2	0.9300 ± 0.1135	1.0000 ± 0.0000	0.5
wdbc	33	24	0.9808 ± 0.0225	0.9790 ± 0.0215	1.0
wpbc	30	26	0.7779 ± 0.0420	0.8034 ± 0.0482	0.95
wine	13	13	0.9889 ± 0.0234	0.9889 ± 0.0234	0.8
Average	24.80	14.70	0.8742	0.8865	0.83

tion, where the number of the selected attributes is the least and the accuracy is the best. We should give an optimal threshold. For simpleness, we recommend the threshold should be in interval [0.75, 0.85].

We can find that 2 features are kept for data *wdbc* and *wine* from Table 5. We present the scatter plots of the data in the selected 2-dimensionality space, shown in Fig. 6. It is easy to find the data in the selected subspaces are easy to recognize.

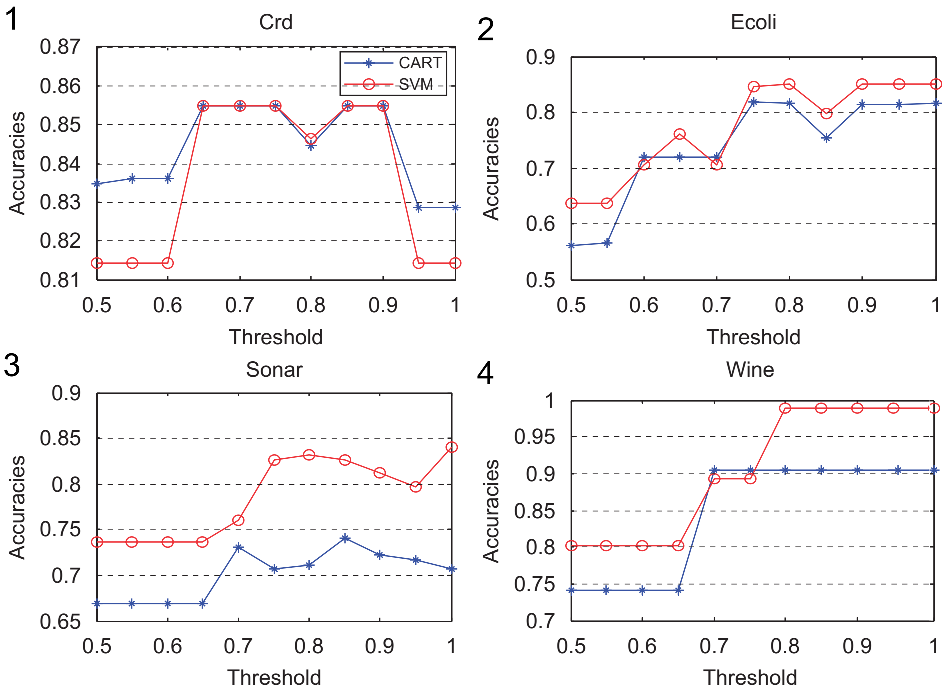

Fig. 5. Classification accuracy changes with the threshold.

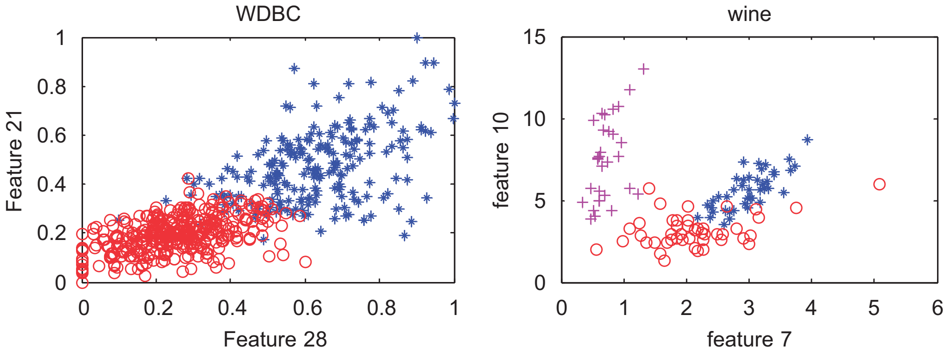

Fig. 6. Scatter plot of data wdbc and wine.

6. Conclusion

Selecting optimal numerical and categorical features is an important challenge for pattern recognition and machine learning. There are two strategies for selecting mixed attributes. One is to discretize numerical attributes into several intervals and take discretized attributes as categorical one. The other is to code a categorical attribute with some integral numbers and look it as a numerical variable. These techniques lose the original information in the data.

In this paper, we show a simple but efficient feature subset selection technique based on a proposed fuzzy-rough model. This approach does not require discretizing the numerical data, whereas classical rough set just work on categorical data. We introduce a symmetric function to compute fuzzy similarity relations between the objects with a numerical attribute and transform the similarity relation into a fuzzy equivalence one. We compute the positive region of the decision by fuzzy inclusion and variable precision fuzzy inclusion. Four attribute significance measures are defined. Based on the measure, we construct a forward hybrid attribute reduction algorithm, named FAR-VPFRS.

With 10 UCI data sets, a series of experiments are conducted for evaluating the proposed method. The results show that most of the features in raw data can be eliminated without decreasing classification performances. What is more, most of classification are improved. We also find that the optimal thresholds for computing variable precision positive regions depend on the learning algorithms. The experiments show a default threshold should be in the interval [0.75, 0.85].

References

[1] I. Guyon, A. Elisseeff, An introduction to variable feature selection, J. Mach. Learn. Res. 3 (2003) 1157–1182.

[2] N. Kwak, C.-H. Choi, Input feature selection for classification problems, IEEE Trans. on Neural Networks 13 (2002) 143–159.

[3] D.P. Muni, N.R. Das Pal, Genetic programming for simultaneous feature selection and classifier design, IEEE Trans. Syst. Man Cybern. Part B 36 (1) (2006) 106–117.

[4] T. Pavlenko, On feature selection, curse-of-dimensionality and error probability in discriminant analysis, J. Stat. Planning Inference 115 (2003) 565–584.

[5] R. Kohavi, G.H. John, Wrappers for feature subset selection, Artif. Intell. 97 (1–2) (1997) 273–324.

[6] K. Kira, L.A. Rendell, The feature selection problem: traditional methods and a new algorithm, in: Proceedings of AAAI-92, San Jose, CA, 1992, pp. 129–134.

[7] C.K. Lee, G.G. Lee, Information gain and divergence-based feature selection for machine learning-based text categorization, Inf. Process. Manage. 42 (2006) 155–165.

[8] M. Dash, H. Liu, Consistency-based search in feature selection, Artif. Intell. 151 (2003) 155–176.

[9] P. Mitra, C.A. Murthy, S.K. Pal, Unsupervised feature selection using feature similarity, IEEE Trans. Pattern Anal. Mach. Intell. 24 (3) (2002) 301–312.

[10] L. Yu, H. Liu, Feature selection for high-dimensional data: a fast correlation-based filter solution, in: Proceedings of the 20th International Conference on Machine Learning, 2003, pp. 856–863.

[11] M. Modrzejewski, Feature selection using rough sets theory, in: P.B. Brazdil (Ed.), Proceedings of the European Conference on Machine Learning, Vienna, Austria, 1993, pp. 213–226.

[12] R. Setiono, H. Liu, Neural-network feature selector, IEEE Trans. Neural Networks 8 (3) (1997) 654–662.

[13] J. Neumann, C. Schnorr, G. Steidl, Combined SVM-based feature selection and classification, Mach. Learn. 61 (2005) 129–150.

[14] H. Liu, R. Setiono, Feature selection via discretization, IEEE Trans. Knowl. Data Eng. 9 (4) (1997) 642–645.

[15] M.J. Beynon, An introduction of the condition class space with continuous value discretization and rough set theory, Int. J. Intell. Syst. 21 (2) (2006) 173–191.

[16] M.R. Chmielewski, J.W. GrzymalaBusse, Global discretization of continuous attributes as preprocessing for machine learning, Int. J. Approx. reasoning 15 (4) (1996) 319–331.

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

[18] R. Jensen, Q. Shen, Semantics-preserving dimensionality reduction: rough and fuzzy-rough-based approaches, IEEE Trans. Knowl. data Eng. 16 (12) (2004) 1457–1471.

[19] R. Jenson, Q. Shen, Fuzzy-rough sets for descriptive dimensionality reductions, Proceedings of IEEE International Conference on Fuzzy Systems, pp. 29–34.

[20] W.Y. Tang, K.Z. Mao, Feature selection algorithm for data with both nominal and continuous features, in: T.B. Ho, D. Cheung, H. Liu (Eds.), PAKDD 2005, Lecture Notes in Artificial Intelligence, vol. 3518, Springer, Berlin, Heidelberg, 2005, pp. 683–688.

[21] W. Pedrycz, G. Vukovich, Feature analysis through information granulation and fuzzy sets, Pattern Recognition 35 (2002) 825–834.

[22] Q. Shen, R. Jensen, Selecting informative features with fuzzy-rough sets and its application for complex systems monitoring, Pattern Recognition 37 (7) (2004) 1351–1363.

[23] R. Jensen, Q. Shen, Fuzzy-rough attribute reduction with application to web categorization, Fuzzy Sets and Systems 141 (3) (2004) 469–485.

[24] R.B. Bhatt, M. Gopal, On fuzzy-rough sets approach to feature selection, Pattern Recognition Lett. 26 (2005) 965–975.

[25] R.B. Bhatt, M. Gopal, On the compact computational domain of fuzzy-rough sets, Pattern Recognition Lett. 26 (2005) 1632–1640.

[26] D. Slezak, Approximate entropy reducts, Fundam. Inf. 53 (3–4) (2002) 365–390.

[27] G.Y. Wang, J. Zhao, J.J. An, et al., A comparative study of algebra viewpoint and information viewpoint in attribute reduction, Fundam. Inf. 68 (3) (2005) 289–301.

[28] Q.H. Hu, D.R. Yu, Entropies of fuzzy indiscernibility relation and its operations, Int. J. Uncertainty Fuzziness Knowl Based Syst. 12 (5) (2004) 575–589.

[29] Q.H. Hu, D.R. Yu, Z.X. Xie, J.F. Liu, Fuzzy probabilistic approximation spaces and their information measures, IEEE Trans. Fuzzy Syst. 14 (2) (2006) 191–201.

[30] Q.H. Hu, D.R. Yu, Z.X. Xie, Information-preserving hybrid data reduction based on fuzzy-rough techniques, Pattern Recognition Lett. 27 (5) (2006) 414–423.

[31] L. Zadeh, Toward a theory of fuzzy information granulation and its centrality in human reasoning and fuzzy logic, Fuzzy Sets and Systems 19 (1997) 111–127.

[32] Y.Y. Yao, Information granulation and rough set approximation, Int. J. Intell. Syst. 16 (1) (2001) 87–104.

[33] A. Skowron, J. Stepaniuk, Information granules: towards foundations of granular computing, Int. J. Intell. Syst. 16 (2001) 57–85.

[34] G. Bortolan, W. Pedrycz, Fuzzy descriptive models: an interactive framework of information granulation, IEEE Trans. Fuzzy Syst. 10 (6) (2002) 743–755.

[35] Y.-Q. Zhang, Constructive granular systems with universal approximation and fast knowledge discovery, IEEE Trans. Fuzzy Syst. 13 (1) (2005) 48–57.

[36] T.Y. Lin, Neighborhood systems and relational database, Abstract, Proceedings of CSC '88, February, 1988, p. 725.

[37] M.R. Berthold, M. Ortolani, D. Patterson, et al., Fuzzy information granules in time series data, Int. J. Intell. Syst. 19 (7) (2004) 607–618.

[38] A. Bargiela, W. Pedrycz, Recursive information granulation: aggregation and interpretation issues, IEEE Trans. Syst. Man Cybern. Part B 33 (1) (2003) 96–112.

[39] T.Y. Lin, Granular computing: fuzzy logic and rough sets, in: L.A. Zadeh, J. Kacprzyk (Eds.), Computing with Words in Information/Intelligent Systems, Physica-Verlag, Wurzburg, 1999, pp. 183–200.

[40] L.A. Zadeh, Fuzzy logic equals computing with words, IEEE Trans. Fuzzy Syst. 4 (2) (1996) 103–111.

[41] Y.Y. Yao, A partition model of granular computing, LNCS Trans. Rough Sets 1 (2004) 232–253.

[42] W. Pedrycz, A.V. Vasilakos, Linguistic models and linguistic modeling, IEEE Trans. Syst. Man Cybernet. Part B 29 (6) (1999) 745–757.

[43] T.Y. Lin, Data mining and machine oriented modeling: a granular computing approach, J. Appl. Intell. 13 (2) (2000) 113–124.

[44] Y.H. Chen, Y.Y. Yao, Multiview intelligent data analysis based on granular computing, Proceedings of 2006 IEEE International Conference on Granular Computing, 2006.

[45] Z. Pawlak, Rough Sets—Theoretical Aspects of Reasoning about Data, Kluwer Academic, Dordrecht, 1991.

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

[47] W. Wu, W. Zhang, Constructive and axiomatic approaches of fuzzy approximation operators, Inf. Sci. 159 (3–4) (2004) 233–254.

[48] T.Y. Lin, Granular data model: semantic data mining and computing with words, in: Proceeding of IEEE Conference on Fuzzy Systems, 2004, pp. 1141–1146.

[49] D.S. Yeung, D.G. Chen, E.C.C. Tsang, J.W.T. Lee, X.Z. Wang, On the generalization of fuzzy rough sets, IEEE Trans. Fuzzy Syst. 13 (3) (2005) 343–361.

[50] S. Guillaume, B. Charnomordic, Generating an interpretable family of fuzzy partitions from data, IEEE Trans. Fuzzy Syst. 12 (3) (2004) 324–335.

[51] A. Bargiela, W. Pedrycz, A model of granular data: a design problem with the Tchebyschev FCM, Soft. Comput. 9 (2005) 155–163.

[52] Z.M. Ma, W.J. Zhang, W.Y. Ma, Assessment of data redundancy in fuzzy relational databases based on semantic inclusion degree, Inf. Process. Lett. 72 (1999) 25–29.

[53] Z.B. Xu, J.Y. Liang, C.Y. Dang, K.S. Chin, Inclusion degree: a perspective on measures for rough set data analysis, Inf. Sci. 141 (3–4) (2002) 227–236.

[54] L.A. Zadeh, Fuzzy sets, Inf. Control 8 (1965) 338–353.

[55] S.-K. Oh, W. Pedrycz, H.-S. Park, Implicit rule-based fuzzy-neural networks using the identification algorithm of GA hybrid scheme based on information granulation, Adv. Eng. Inf. 16 (2002) 247–263.

[56] Q. Hu, D. Yu, An improved clustering algorithm for information granulation, Lecture Notes in Artificial Intelligence, vol. 3613, FSKD 2005, Proceedings, 2005, pp. 494–504.

[57] H.-S. Lee, An optimal algorithm for computing the max–min transitive closure of a fuzzy similarity matrix, Fuzzy Sets and Systems 123 (1) (2001) 129–136.

About the Author—QINGHUA HU received his master degree in power engineering from Harbin Institute of Technology, Harbin, China in 2002. Now he is a Ph.D. student with Harbin Institute of Technology. His research interests are focused on data mining, knowledge discovery with fuzzy and rough techniques. He has authored or coauthored more than 40 journal and conference papers in the areas of machine learning, data mining and rough set theory.

About the Author—ZONGXIA XIE received her B.Eng. in control engineering from Dalian Maritime University in 2003 and master degree from Harbin Institute of Technology in 2005, respectively. Now she is working for her Ph.D. in Harbin Institute of Technology. Her current research interests include feature selection, image recognition and SVM.

About the Author—DAREN YU was born in Datong, China, in 1966. He received his M.Sc. and D.Sc. degrees from Harbin Institute of Technology, Harbin, China in 1988 and 1996, respectively. Since 1988, he has been working at the School of Energy Science and Engineering, Harbin Institute of Technology. His main research interests are in modeling, simulation, and control of power systems. He has published more than one hundred conference and journal papers on power control and fault diagnosis.