Available online at www.sciencedirect.com

SCIENCE DIRECT®

Pattern Recognition Letters 27 (2006) 414–423

Pattern Recognition Letters

www.elsevier.com/locate/patrec

Information-preserving hybrid data reduction based on fuzzy-rough techniques

Qinghua Hu *, Daren Yu, Zongxia Xie

Harbin Institute of Technology, Power Engineering, 150001, Heilongjiang Province, PR China

Received 1 January 2005; received in revised form 10 July 2005
Available online 25 October 2005

Communicated by Prof. F. Roli

Abstract

Data reduction plays an important role in machine learning and pattern recognition with a high-dimensional data. In real-world applications data usually exists with hybrid formats, and a unified data reducing technique for hybrid data is desirable. In this paper, an information measure is proposed to computing discernibility power of a crisp equivalence relation or a fuzzy one, which is the key concept in classical rough set model and fuzzy-rough set model. Based on the information measure, a general definition of significance of nominal, numeric and fuzzy attributes is presented. We redefine the independence of hybrid attribute subset, reduct, and relative reduct. Then two greedy reduction algorithms for unsupervised and supervised data dimensionality reduction based on the proposed information measure are constructed. Experiments show the reducts found by the proposed algorithms get a better performance compared with classical rough set approaches.

Keywords: Attribute reduction; Hybrid data; Fuzzy-rough set; Information measure

1. Introduction

In recent years, data has become increasingly larger not only in rows (i.e. number of instances) but also in columns (i.e. number of features) in many applications, such as gene selection from micro-array data and text automatic categorization, where the number of features in the raw data ranges from hundreds to tens of thousands (Guyon and Elisseeff, 2003). High dimensionality brings great difficulty in pattern recognition, machine learning and data mining (Hand et al., 2001; Liu and Setiono, 1998). Data reduction is a well-known data mining problem, which is usually considered as an enhancement preprocessing technique for subsequent machining (Tsang et al., 2003). It will bring many potential benefits: reducing the measurement, storage and transmission, reducing training and utilization times, defying the curse of dimensionality to improve prediction performance in terms of speed, accuracy and simplicity, facilitating data visualization and data understanding (Torkkola, 2003; Dash and Liu, 2003). A lot of data reduction techniques were proposed to deal with these challenging tasks. Due to the complexity of data and classification in real-world applications, it seems not an easy task to build a general data reduction technique, so researches on data reduction have been conducted for last several decades and are still extracting much attention from pattern recognition and data mining society. Data reduction can begin with two aspects: reducing the number of samples or reducing the number of features. The first one will be implemented by resample techniques and the second is done with dimensionality reduction techniques (Blum and Langley, 1997; Liu et al., 2002). This paper will be focused on the second problem.

* Corresponding author. Tel.: +86 45186413241; fax: +86 45186413241221.
E-mail address: huqinghua@hcms.hit.edu.cn (Q. Hu).

0167-8655/$ - see front matter © 2005 Elsevier B.V. All rights reserved.
doi:10.1016/j.patrec.2005.09.004

An extensive amount of researches have been conducted over last decades to get reliable approaches for dimensionality reduction, which roughly falls into two types of paradigms: feature extraction and feature subset selection (Li and Xu, 2001). Feature extraction refers to constructing new features with a linear or nonlinear transformation from the original input space to a feature space, while feature subset selection is to find some informative features from the original set and delete the others. Principal component analysis (PCA) (Hwang and Chang, 2002; Chen and Zhu, 2004), independent component analysis (ICA) (Cheung and Xu, 2001; Wakako, 2002), linear discriminant analysis (LDA) are to find a linear transformation and projection pursuit regression constructs a nonlinear mapping from input space to feature space. A main drawback of these methods is that the constructed features do not have true meaning, and complex computation may be required (Tsang et al., 2003).

In last decade, much attention has been paid to feature subset selection. Two extensive reviews were published (Blum and Langley, 1997; Kohavi and John, 1997) in *artificial intelligence* and a special issue of machine learning research was present in 2003 (Guyon and Elisseeff, 2003). Generally speaking, there are four basic components in all feature subset selections: an evaluation function of feature subset, a search strategy to find the best feature subset as defined by the corresponding evaluation function, a stopping criterion to decide when to stop and a validation procedure to check whether the selected subset is valid (Piramuthu, 2004). According to evaluation methods the feature subset selection can classified into two kinds: filtering and wrapper. Distance measures (Kira and Rendell, 1992; Kwak and Choi, 2002), information measures (Yu and Liu, 2003a,b; Duch et al., 2002), correlation coefficient (Mitra et al., 2002) and consistency measures (Dash and Liu, 2003) are used for filtering methods. Wrapper refers to using a classifier as the evaluation function in selection. KNN, neural network, SVM all can be employed. Isabelle Guyon and Elisseeff (2003) pointed that selecting the most relevant features is usually suboptimal for building a good predictor in filtering because the performance of the trained predictor depends on not only feature subset, but also the learner used. In other words, a best feature subset in terms of an evaluation function does not mean a best prediction performance. An optimal feature subset selection should be conducted by the corresponding classifier employed, which leads to wrapper methods. However wrapper methods will take high time-complexity. It is may be infeasible in real-world applications. Filtering as an efficient feature selection is widely used in practice. In filtering methods, information measures and consistency measures work effectively when data are nominal. Compared with these measures, distance measures and correlation coefficient are proposed for numeric data. However, data usually comes with a hybrid form in applications. For example, nominal attributes: sex, color, numeric attributes: age, temperature coexist in hospital data set. The above selection methods are suitable for a single format of features in nature. A feature subset selection for hybrid data is desirable.

Rough set theory has proved to be a powerful tool to deal with uncertainty and has been applied to data reduction, rule extraction, data mining and granularity computation. Reduct is a minimal attribute subset of the original data which is independent and has the same discernibility power as all of the attributes in rough set framework. Obviously reduction is a feature subset selection process, where the selected feature subset not only retains the representational power, but also has minimal redundancy. Some rough set based reduction and feature selection algorithms have been proposed. Consistency of data (Mi et al., 2004; Pawlak, 1991), dependency of attributes (Wang and Miao, 1998), mutual information (Wang et al., 2002), discernibility matrix (Skowron and Rauszer, 1992) are employed to find reducts of an information system. And these techniques are applied to data reduction (Beynon, 2001; Li et al., 2004) text classification (Moradi et al., 1998), texture analysis (Swiniarski and Hargis, 2001). An extensive review about rough set based feature selection was given in (Swiniarski and Skowron, 2003).

As we know, Pawlak's rough set model works in case that only nominal attributes exist in information systems. However, data usually comes with a hybrid form. Nominal, fuzzy and numeric features coexist in real world databases. Some generalizations of the model were proposed to deal with this problem. Rough set theory and fuzzy set theory were putted together and rough-fuzzy sets and fuzzy-rough sets were defined in (Dubois and Prade, 1992). The properties and axiomatization of fuzzy rough set theory (Morsi and Yakout, 1998; Wu and Zhang, 2004; Wu et al., 2004) were analyzed in detail. And the fuzzy-rough set methods were applied to data reduction (Hu et al., 2005; Hu et al., in press) mining stock price (Wang, 2003), vocabulary for information retrieval (Srinivasan et al., 2001) and fuzzy decision rules (Shen and Chouchoulas, 2002).

Just as reduction plays an important role in classical rough set theory, a reduction algorithm for fuzzy information systems is desirable. In traditional processing, discretization is performed on numeric data as a preprocessing for machine learning (Chmielewski and Grzymala-Busse, 1996). Qiang Shen et al. pointed that this processing may lead to some information loss from the original data. A fuzzy-rough attribute reduction, called fuzzy-rough QUICKREDUCT algorithm, was given in (Jensen and Shen, 2004; Shen and Jensen, 2004) based on fuzzy dependency function. Fuzzy dependency function has the power to measure the discernibility power of nominal attributes and fuzzy attributes.

In this paper, we will introduce an information measure for fuzzy equivalence relations. Then we will redefine the dependency of a hybrid attribute set and give unsupervised and supervised reduction algorithms for hybrid data based on the measure. The rest of the paper is organized as follows: some preliminary knowledge about rough set and fuzzy-rough set theory is present in Section 2. A novel information measure and its properties will be presented in Section 3. Section 4 gives a new definition of dependency of attribute set and reduction algorithms for hybrid data. An extensive experimental analysis is described in Section 5. Section 6 concludes the paper.

2. Some primary definitions on fuzzy-rough set model

Pawlak's rough set model can only deal with data containing nominal values. As we know the real-world applications usually contain real-valued or fuzzy attributes. A fuzzy equivalence relation would be generated by a real-valued attribute or a fuzzy attribute, instead of crisp equivalence relation. The fuzzy-rough set model is fitted for the case where both the relation and the object subset to be approximated are fuzzy.

Given a non-empty finite set X,R is a binary relation defined on X, denoted by a relation matrix $M(R)$:

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

where $r_{ij} \in [0,1]$ is the relation value of x_i and x_j.

R is a fuzzy equivalence relation if R satisfies

(1) Reflectivity: $R(x,x) = 1\ \forall x \in X$;
(2) Symmetry: $R(x,y) = R(y,x)$, $\forall x,y \in X$;
(3) Transitivity: $R(x,Z) \geqslant \min_y\{R(x,y), R(y,z)\}$.

Given an arbitrary set X, R is a fuzzy equivalence relation defined on X. $\forall x,y \in X$, some operations on relation matrices are defined as

(1) $R_1 = R_2 \iff R_1(x,y) = R_2(x,y)\ \forall x,y \in X$;
(2) $R = R_1 \cup R_2 \iff R(x,y) = \max\{R_1(x,y), R_2(x,y)\}$;
(3) $R = R_1 \cap R_2 \iff R(x,y) = \min\{R_1(x,y), R_2(x,y)\}$;
(4) $R_1 \subseteq R_2 \iff R_1(x,y) \leqslant R_2(x,y)$.

A crisp equivalence relation will generate a crisp partition of the universe, whereas a fuzzy equivalence relation induces a fuzzy partition.

Definition 1. U is the universe and R is a fuzzy binary relation over U. The fuzzy partition of the universe U, generated by a fuzzy equivalence relation R, is defined as

$$U/R = \{[x_i]_R\}_{i=1}^{n}, \tag{1}$$

where $[x_i]_R = \frac{r_{i1}}{x_1} + \frac{r_{i2}}{x_2} + \cdots + \frac{r_{in}}{x_n}$ is the fuzzy equivalence class generated by x_i and R.

Here U/R means the partition of U induced by relation R. Due to the fuzzy equivalence relation, U/R is a fuzzy partition and then $[x_i]_R$ is a fuzzy set. This is a main difference of fuzzy-rough sets with Pawlak's rough sets.

Theorem 1. *Given arbitrary set X, R is a fuzzy equivalence relation defined on X. The fuzzy quotient set of X by relation R is denoted by X. $\forall x,y \in X$, we have*

(1) $R(x,y) = 0 \iff [x]_R \cap [y]_R = \phi$;
(2) $[x]_R = [y]_R \Rightarrow R(x,y) = 1$.

Definition 2. Given a fuzzy probability approximation space $\langle U,R\rangle$, X is a fuzzy subset of U. The *lower approximation* and *upper approximation*, denoted by $\underline{R}X$ and $\overline{R}X$, are defined as

$$\begin{cases} \mu_{\underline{R}X}(x) = \wedge\{\mu_X(y) \vee (1 - R(x,y)) : y \in U\}, & x \in U \\ \mu_{\underline{R}X}(x) = \vee\{\mu_X(y) \wedge R(x,y) : y \in U\}, & x \in U \end{cases}. \tag{2}$$

The membership of an object $x \in U$, belonging to the fuzzy positive region is defined as

$$\mu_{\mathrm{POS}_{\tilde{B}}(d)} = \sup_{X \subseteq U/d} \mu_{\underline{B}X}(x). \tag{3}$$

Definition 3. Given a fuzzy information system $\langle U,A\rangle$, B and d are two subset of attribute set A, the dependency degree of d to B is defined as

$$\gamma_B(d) = \sum_{x \in U} \mu_{\mathrm{POS}_B(d)}(x). \tag{4}$$

Definition 4. Given a fuzzy information system $\langle U,A,V,f\rangle$, $B \subseteq A$, $a \in B$, if $U/B = U/(B-a)$, we say knowledge a is *redundant* or *superfluous* in B. otherwise, we say knowledge a is *indispensable*. If any a belonging to B is *indispensable*, we say B is *independent*. If attribute subset $B \subseteq A$ is *independent* and $U/B = U/A$, we say B is a *reduct* of A.

Definition 5. Given a fuzzy information system $\langle U,A,V,f\rangle$, $A = C \cup d$. B is a subset of C. $\forall a \in B$, a is redundant in B relative to d if $\gamma_{B-a}(d) = \gamma_B(d)$, otherwise a is indispensable. B is independent if $\forall a \in B$ is indispensable, otherwise B is dependent. B is a subset of C. B is a reduct of C if B satisfies:

(1) $\gamma_B(d) = \gamma_C(d)$;
(2) $\forall a \in B: \gamma_{B-a}(d) < \gamma_B(d)$.

The fuzzy-rough set model is the generalization of classical rough set model and rough-fuzzy set model. When the relations between objects are crisp equivalence relations and the object subset to be approximated is a fuzzy set then the model will degrade to rough-fuzzy set model. Furthermore, if object subset to be approximated is crisp, the model is the classical one.

3. Information measure for fuzzy-rough set model

In this section we will propose a new entropy to measure the discernibility power of a fuzzy equivalence relation.

Given a finite set U, A is a fuzzy or real-valued attribute set, which generates a fuzzy equivalence relation R on U. The fuzzy relation matrix is $M(R)$.

The fuzzy equivalence class generated by x_i and R is

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

Definition 6. The cardinality $[x_i]_R$ is defined as

$$|[x_i]_R| = \sum_{j=1}^{n} r_{ij}. \tag{5}$$

Definition 7. Information quantity of the fuzzy attribute set or the fuzzy equivalence relation is defined as

$$H(R) = -\frac{1}{n}\sum_{i=1}^{n} \log \lambda_i, \tag{6}$$

where $\lambda_i = \frac{|[x_i]_R|}{n}$.

If the relation R is a crisp equivalence relation, the proposed information measure is identical to Shannon's one. The following definitions of joint entropy and conditional entropy have the same property. In the follows, we will denote two information measures indiscriminatingly.

The formula of information measure forms a map: $H : R \to \Re^+$, where R is a equivalence relation matrix, $\Re^+$ is the non-negative real-number set. This map builds a foundation on that we can compare the discernibility power, partition power or approximating power of multiple fuzzy equivalence relations. Entropy value increases monotonously with the discernibility power or the knowledge's fineness. So the finer partition is, the greater entropy is, and the more significant attribute set is.

Definition 8. Given a fuzzy information system $\langle U, A, V, f \rangle$, A is the fuzzy or numeric attribute set. B, E are two subsets of A. $[x_i]_B$ and $[x_i]_E$ are fuzzy equivalence classes containing x_i generated by B and E, respectively. The joint entropy of B and E is defined as

$$H(BE) = H(R_E R_B) = -\frac{1}{n}\sum_{i=1}^{n} \log \frac{|[x_i]_B \cap [x_i]_E|}{n} \tag{7}$$

Definition 9. Given a fuzzy information system $\langle U, A, V, f \rangle$, A is the attribute set. B and E are two subsets of A. $[x_i]_B$ and $[x_i]_E$ are fuzzy equivalence classes containing x_i generated by B and E, respectively. The conditional entropy of E conditioned to B is defined as

$$H(E|B) = -\frac{1}{n}\sum_{i=1}^{n} \log \frac{|[x_i]_E \cap [x_i]_B|}{|[x_i]_B|} \tag{8}$$

Theorem 2. $H(E|B) = H(BE) - H(B)$.

Theorem 3. *Given a fuzzy information system* $\langle U, A, V, f \rangle$, A *is the fuzzy attribute set.* B *and* E *are two subsets of* A. $[x_i]_B$ *and* $[x_i]_E$ *are fuzzy equivalence classes containing* x_i *generated by* B *and* E*, respectively. The fuzzy equivalence relations induced by* B *and* E *are denoted by* R *and* S*, respectively. Then we have*:

(1) $\forall B \subseteq A{:}H(B) \geqslant 0$;
(2) $H(BE) \geqslant \max\{H(B), H(E)\}$;
(3) $B \subseteq E$ or $R_B \subseteq R_E{:}H(BE, P) = H(B)$;
(4) $B \subseteq E$ or $R_B \subseteq R_E{:}H(E|B) = 0$.

The first item of Theorem 3 shows the information introduced by any attribute subset is non-negative, the second shows the discernibilty power of the union of two attribute subset will be no less than that of any single subset, which means introducing a new attribute or attribute subset at least will not decrease the discernibility power. The last two items show attribute subset won't introduce information relative B if E is contained by B. The properties of the information measure have a same observation of classification as the Boolean logic methodology, which is a class of paradigm of classifier, such as ID3, CART, C4.5 and rough set theory.

Theorem 4. *Given a fuzzy information system* $\langle U, A, V, f \rangle$, $B \subseteq A$, $a \in B$, $H(B) = H(B - a)$ *if* a *is redundant*; $H(B) > H(B - a)$ *if* B *is independent.* B *is a reduct if* B *satisfies*:

(1) $H(B) = H(A)$;
(2) $\forall a \in B{:}H(B) > H(B - a)$.

Theorem 5. *Given a fuzzy information system* $\langle U, A, V, f \rangle$, $A = C \cup d$. B *is a subset of* C. $\forall a \in B$, $H(d|B - a) = H(d|B)$ *if* a *is redundant in* B *relative to* d; $H(d|B - a) > H(d|B)$ *if* B *is independent.* B *is a reduct of* C *relative to* d *if* B *satisfies*:

(1) $H(d|B) = H(d|C)$;
(2) $\forall a \in B{:}H(d|B - a) > H(d|B)$.

Theorems 4 and 5 give the definitions of dependency, reduct and relative reduct in terms of information theory, in fact two classes of definitions are equivalent. The proof was given in (Hu and Yu, 2004, in press).

Example 1. Given a set $X = \{x_1, x_2, x_3\}$. R_1, R_2, R_3 are fuzzy equivalence relation matrices on X, induced by attributes a_1, a_2 and a_3, as follows:

$$R_1 = \begin{bmatrix} 1 & 0.9 & 0 \\ 0.9 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix},$$

$$R_2 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0.8 \\ 0 & 0.8 & 1 \end{bmatrix},$$

$$R_3 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0.9 \\ 0 & 0.9 & 1 \end{bmatrix}.$$

We have $[x_1]_{R_1} = \frac{1}{x_1} + \frac{0.9}{x_2} + \frac{0}{x_3}$, $|[x_1]_{R_1}| = 1.9$

$$R_4 = R_1 \cap R_2 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix},$$

$$R_5 = R_1 \cap R_3 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}.$$

Let us compute the following entropy $H(R_1)$, $H(R_2)$.

$H(R_1) = 0.9676, \quad H(R_2) = 1.0196.$

The joint entropy of relations R_1 and R_2 is

$H(R_1R_2) = H(R_1 \cap R_2) = 1.5850.$

Then the conditional entropies $H(R_1|R_2)$ and $H(R_2|R_1)$ are

$$\begin{aligned} H(R_1|R_2) &= H(R_1R_2) - H(R_2) \\ &= 1.5850 - 1.0196 \\ &= 0.5654, \\ H(R_2|R_1) &= H(R_1R_2) - H(R_1) \\ &= 1.5850 - 0.9676 \\ &= 0.6174. \end{aligned}$$

Given an equivalence relation R_{d} induced by a decision attribute:

$$R_{\mathrm{d}} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix},$$

we find

$H(R_{\mathrm{d}}|R_1R_2) = H(R_{\mathrm{d}}|R_1R_3) = H(R_{\mathrm{d}}|R_1R_2R_3) = 0,$

which shows the joint relation of R_1 and R_2 has the same partition as the relation R_{d}, and so does the joint relation of R_1 and R_3. So $\{a_1, a_2\}$ and $\{a_1, a_3\}$ are two relative reducts.

4. Reduction algorithms for unsupervised and supervised hybrid data

Reduct is an important concept in rough set theory and data reduction is a main application of rough set theory in pattern recognition and data mining. As it has been proven that finding the minimal reduct of an information system is a NP hard problem. Some heuristic algorithms have been invented based on significance measures of attributes. These algorithms get a suboptimal result but relatively low time-consuming (Guyon and Elisseeff, 2003). Shannon's entropy was used as a significance measure in some classical machine learning algorithm, such as the famous ID3 algorithm series, and proven to be a good measure. In the above section, we propose a novel information measure for fuzzy indiscernibility or equivalence relation and show that the entropy can be degraded to Shannon's one when the relation measured is a crisp equivalence one. It shows that the proposed measure can be used as a measure of discernibility power of a crisp equivalence relation and a fuzzy one. So unified reduction algorithms for hybrid data are feasible.

Data dimensionality reduction will be divided into three steps: relation computation, reduction and reduct validation. Relation computation is to generate relation matrices using a relation function with attributes. Then reduction algorithms are performed on the matrices and find some reduct of the original data. Finally employing a validation function, which may be a classifier or a discriminability criterion, we test the reduct and find a best one. The procedure is shown as follows. No matter cases $\{x_i\}_{i=1}^{n}$ are described by nominal attributes or numeric features or fuzzy variables (Fig. 1), the relations between the cases can all be denoted by a relation matrix: $M(R) = (r_{ij})_{n\times n}$.

If A is a nominal attribute set,

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

If attribute a is a numeric attribute, the value the relation can mapped by a symmetric function

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

where function f should satisfy

(1) $f(0) = 1$, $f(\infty) = 0$ and $f(\bullet) \in [0, 1]$;
(2) $r_{ij} = r_{ji}$ and $r_{ii} = 1$.

According to (2), relation R will satisfies reflexivity and symmetry. So a similarity relation matrix will be produced by the functions.

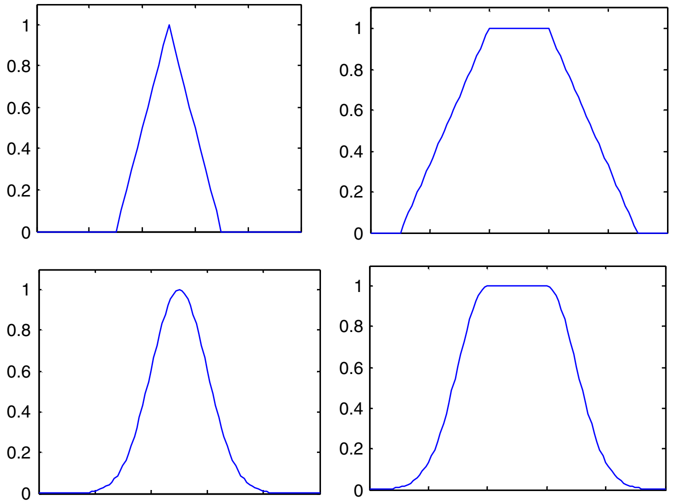

Fig. 1. Some similarity relation functions for numeric data.

As to fuzzy attributes, there are a great many candidate similarity measures (Li and Cheng, 2002). For example

(1) Hamming similarity measure:

$$S(x_i, x_j) = \frac{1}{m}\sum_{k=1}^{m}(1 - |\mu_{A_k}(x_i) - \mu_{A_k}(x_j)|);$$

(2) Max–Min similarity measure:

$$S(x_i, x_j) = \frac{1}{m}\left\{\sum_{k=1}^{m}\frac{\min(\mu_{A_k}(x_i), \mu_{A_k}(x_j))}{\max(\mu_{A_k}(x_i), \mu_{A_k}(x_j))}\right\}.$$

Employing a max–min closure operation, we can get a fuzzy equivalence relation (Lee, 2001).

As has pointed in Section 2, the proposed entropy can be used as measure of the discernibility power of a relation or an attribute. The greater the entropy value is, the stronger the discernibility is and the more significant the attribute is. According to the properties of proposed entropy, adding a novel condition attribute into the information system, the entropy value will increase monotonously, which reflexes that adding information will lead to enhancement of the discernibility power. The increment of information by an attribute reflexes the increment of discernibility of the system. So the significance of an attribute can be defined as follows.

Definition 10. Given a fuzzy information system $\langle U, A, V, f\rangle$, $B \subseteq A$, $a \in B$, the significance of attribute a in attribute set B is defined as

$$SIG(a, B) = H(B) - H(B - a) \tag{9}$$

The above definition works in unsupervised feature selection. $SIG(a, B)$, called Significance of attribute a in B, measures the increment of discernibility power introduced by attribute a.

Definition 11. Given a fuzzy information system $\langle U, A, V, f\rangle$, $A = C \cup d$, where C is the condition attribute set and d is the decision attribute. $B \subseteq C$ $\forall a \in B$, the significance of attribute a in attribute set B relative to d is defined as

$$SIG(a, B, d) = H(d|B - a) - H(d|B) \tag{10}$$

This definition computes the increment of discernibility power relative to the decision introducing by attribute a. So it may be used as a supervised measure for feature selection.

Based on the above measures, two greedy algorithms for computing reduct and relative reduct can be constructed, respectively.

Algorithm 1. Algorithm for calculating reduct

```
    Input: Information system IS $\langle U, A, V, f\rangle$.
    Output: One reduct of IS
Step 1: $\forall a \in A$: compute the equivalence relation;
Step 2: $\phi \to red$;
Step 3: For each $a_i \in A - red$
            Compute $H_i = H(a_i, red)$
        End
Step 4: Choose attribute which satisfies:
        $H(a|red) = \max_i(SIG(a_i, red))$
Step 5: If $H(a|red) > 0$, then $red \cup a \to red$ goto step 3
        Else return $red$
        End
```

Algorithm 2. Algorithm for calculating relative reduct

```
    Input: Information system IS $\langle U, A = C \cup d, V, f\rangle$.
    Output: One relative reduct $D\_red$ of IS
Step 1: $\forall a \in A$: compute the equivalence;
Step 2: $\phi \to D\_red$;
Step 3: For each $a_i \in C - D\_red$
            Compute $H_i = SIG(a_i, D\_red, d)$
        End
Step 4: Choose attribute which satisfies:
        $SIG(a, red, d) = \max_i(H_i)$
Step 5: If $SIG(a, red, d) > 0$, then $D\_red \cup a \to D\_red$ goto
        step 3
        Else return, $D\_red$
        End
```

Jensen and Shen (2004) proposed that a problem may arise when this approach is compared to the crisp attribute reduction. In classical rough set attribute reduction, a reduct is defined as a subset of attributes which has the same information quantity as the full attribute set, which means that the value $H(B)(H(d|B))$ should be identical to $H(A)(H(d|A))$. However, in the fuzzy-rough approaches, it is not necessarily the case. We can specify the degree threshold λ. So that the algorithms will stop if the condition $SIG(a, red) \leqslant \lambda(SIG(a, red, d) \leqslant \lambda)$ is satisfied.

5. Experiments and analysis

A series of experiments have been conducted to test the proposed significance measure of attributes and feature selection based on UCI data. In this section, we will show some experimental results and analysis. All experiments have been performed on data set shown in Table 1. We find the attributes of data BC and BCW are nominal, and others are hybrid.

5.1. Experiment 1: Ranking based feature selection vs. the proposed dimensionality reduction

In feature subset selection, many algorithms include ranking as a principal or auxiliary selection mechanism because of its simplicity, scalability and good empirical success. Ranking methods employ an evaluation function, such as inter-class distance, correlation criteria, mutual

Table 1
Summary of the experiment data sets

Data set		Size	Class number	Attribute number		
Abbreviation	Original name			Total	Numeric	Nominal
WDBC	Breast-cancer-wisconsin 2	569	2	31	30	1
WPBC	Breast-cancer-wisconsin 3	198	2	33	32	1
Cre	Credit Approval	690	2	16	6	10
Cle	Cleve Database	303	5	14	5	9
Der	Dermatology	366	6	34	33	1
Eco	Protein localization	336	8	8	7	1
Gls	Glass identification	214	6	9	8	1
Heart	Heart disease	270	2	14	6	8
Ion	Ionosphere	351	2	35	34	1
Son	Sonar mines	1389	3	1	60	1
Win	Wine recognition	178	3	14	13	1

information and accuracy of a classifier to sort the candidate features. Some top features are selected. The main drawback of ranking is it can not detect the redundancy or correlation among condition set. So although they are the greatest discernible feature individually, their combination may have weak discernible power. Only under certain independence or orthogonality, ranking may be optimal with respect to a given classifier (Guyon and Elisseeff, 2003).

In the follows, an experiment is shown based on data *wine*. The order of significance of attribute set is $\{7,13,12,10,1,11,6,2,8,4,9,5,3\}$. With reduction Algorithm 2, attribute subset $\{7,1,11,6,3,13\}$ are selected one by one as a reduct, called subset 1.

In order to compare two feature subset selection methods, top six attributes $\{7,13,12,10,1,11\}$ are selected in ranking, called subset 2. Fig. 2 shows the distribution of data in two-dimension feature space. Fig. 2(a) is the distribution with attribute $\{7,1\}$, $\{1,11\}$, $\{11,6\}$, $\{6,3\}$, $\{3,13\}$, respectively. And Fig. 2(b) is the distribution with attribute $\{7,13\}$, $\{13,12\}$, $\{12,10\}$, $\{10,1\}$, $\{1,11\}$. From the two-dimension feature space, we find that the attributes by ranking have even better discernibility power than the attributes selected by the fuzzy-rough reduction algorithm. Here we choose support vector machine (SVM) as a validation function for feature subsets. 2/3 samples are randomly selected as training set, and the others are test set. The accuracy with attribute subset 1 is 94.87%, while the accuracy with attribute subset 2 is 93.33%.

Why the attributes with better discriminability in two-dimensional space get an even worse classification performance? As we have pointed, selecting the most relevant features is usually suboptimal for building a classifier if the features are redundant or dependent. Generally

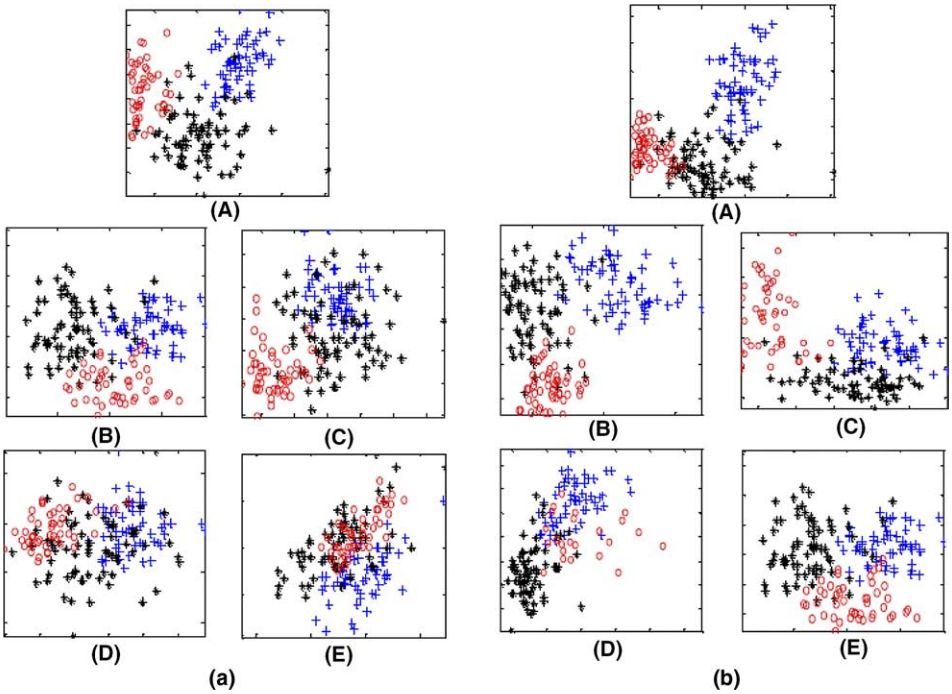

Fig. 2. Distribution of *wine* samples with attributes: (a) $\{7,1,11,6,3,13\}$, accuracy: 94.87%; (b) $\{7,13,12,10,1,11\}$, accuracy: 93.33%.

speaking, ranking method only computes the dependency between condition attributes and decision attribute, while neglect the dependency among condition attributes.

Let us analyze the correlation between the selected condition attributes. Correlation coefficients are showed in Tables 2 and 3.

Wang et al. (2003) introduced correlation entropy to measure the correlation of a variable set. The entropy is defined as

$$H_R = -\sum_{i=1}^{N} \frac{\lambda_i}{N} \log_N \frac{\lambda_i}{N},$$

where λ_i is ith eigenvalue of correlation coefficient matrix. The greater the entropy value is, the weaker the correlation of attribute set is. If all attributes are linear correlation, the correlation entropy is 0, and if all the correlation coefficient are zero, then the entropy is 1. Wang called the dependency of attributes overlap information. We employ the measure to compute the correlation degree of the selected attributes. The correlation entropy of subset 1 is 0.8110, while entropy of subset 2 is 0.7364, which shows the correlation degree of subset 1 is lower than that of subset 2.

5.2. Experiment 2: Comparison of reduction methods

In order to test the performance of the proposed reduction algorithm, some contrastive experiments are conducted based on UCI data set. We compare the classical rough set reduction with the proposed one and employ SVM and CART classifier as the validation function. The experiment data is shown in Tables 4 and 5.

As we know, the classical rough set theory just works in nominal domain. So discretization is performed on numeric data before reduction. The numeric attributes are discretized into three intervals by equal-width, equal-frequency and fuzzy c-means clustering techniques. As to fuzzy-rough reduction algorithm, the relation matrices are computed with a triangle function.

We can find that the average accuracies with the proposed algorithms are higher than those with classical rough set based reduction. In the same time we can find the proposed algorithms keep more attributes in the reducts, which show there are some useful attributes in the reduced subsets. In the classical rough set based reduction the discretization is usually performed before reduction and learning with some crisp cuts. The selection of cuts is crucial to the performance of the sequent learning. The cuts should reflex the structure of the data and patterns. Generally speaking, the boundary of the patterns is fuzzy and indistinguishable, a crisp cut point can capture the actual semantic in the data. Therefore, discretization enhances the discernibility power of the original training data. As to fuzzy data or numeric features, fuzzy equivalence relations are capable of modeling the uncertainty in the data sets. So the learned models with the proposed technique may get good performances. Especially to the data with few attributes, such as data *wine*, the classical methods only keep 4 attributes, whereas the proposed method retains 6 attributes. As WDBC and WPDC the proposed technique withholds much more features in the final reducts. Accordingly, good performances are observed. The results of the numeric experiments maybe suggest the data are over reduced with the classical techniques.

It's certain that we do not believe the more the features are, the higher the accuracies we will get. Bhatt and Gopal (2005) showed the fuzzy-rough set reduction algorithm was not convergent on many real datasets due to the poorly designed termination criteria. In fact, the convergence depends on not only the termination criteria, but also the computation of the fuzzy sets and fuzzy relations.

Table 2
Correlation coefficient matrix of attribute set {7, 1, 11, 6, 3, 4} with correlation entropy 0.8110

	A1	A2	A3	A4	A5	A6
A1	1.0000	0.2368	0.5435	0.8646	0.1151	−0.3514
A2	0.2368	1.0000	−0.0717	0.2891	0.2115	−0.3102
A3	0.5435	−0.0717	1.0000	0.4337	−0.0747	−0.2740
A4	0.8646	0.2891	0.4337	1.0000	0.1290	−0.3211
A5	0.1151	0.2115	−0.0747	0.1290	1.0000	0.4434
A6	−0.3514	−0.3102	−0.2740	−0.3211	0.4434	1.0000

Table 3
Correlation coefficient matrix of attributes {7, 13, 12, 10, 1, 11} with correlation entropy 0.7364

	A1	A2	A3	A4	A5	A6
A1	1.0000	0.4942	0.7872	−0.1724	0.2368	0.5435
A2	0.4942	1.0000	0.3128	0.3161	0.6437	0.2362
A3	0.7872	0.3128	1.0000	−0.4288	0.0723	0.5655
A4	−0.1724	0.3161	−0.4288	1.0000	0.5464	−0.5218
A5	0.2368	0.6437	0.0723	0.5464	1.0000	−0.0717
A6	0.5435	0.2362	0.5655	−0.5218	−0.0717	1.0000

Table 4
Comparisons of fuzzy-rough technique vs. discritization with linear SVM classifiers

Data	Original data		Reduct (equi-width)		Reduct (equi-fre.)		Reduct (FCM)		Reduct (fuzzy-rough)	
	n	Accuracy (%)	*n*	Accuracy (%)	*n*	Accuracy (%)	*n*	Accuracy (%)	*n*	Accuracy (%)
WDBC	31	93.16	8	94.21	12	93.68	6	95.26	17	95.26
WPBC	33	74.24	8	71.21	6	75.76	6	68.18	17	81.82
Cre	16	82.17	11	81.74	9	83.04	11	81.74	12	81.74
Cle	14	59.41	10	57.43	8	60.4	9	59.41	12	56.44%
Der	34	90.91	12	93.39	11	99.17	11	99.17	11	99.17
Eco	8	70.18	7	70.18	7	70.18	7	70.18	7	70.18
Heart	14	83.33	9	83.33	8	82.22	8	84.44	9	83.33
Ion	35	92.31	7	85.47	7	85.47	8	87.18	12	88.03
Son	61	78.57	6	71.43	6	52.86	8	74.29	9	74.29
Win	14	96.67	4	91.67	4	91.67	4	91.67	6	94.87
Average		82.10		80.01		79.45		81.15%		82.52

Table 5
Comparisons of fuzzy-rough technique vs. discritization with decision tree

Data	Original data		Reduct (equi-width)		Reduct (equi-fre.)		Reduct (FCM)		Reduct (fuzzy-rough)	
	n	Accuracy (%)	*n*	Accuracy (%)	*n*	Accuracy (%)	*n*	Accuracy (%)	*n*	Accuracy (%)
WDBC	31	91.05	8	90.00	12	94.74	6	91.58	17	96.32
WPBC	33	59.09	8	59.09	6	60.61	6	57.58	17	62.12
Cre	16	82.17	11	81.30	9	80.43	11	80.87	12	81.30
Cle	14	58.42	10	49.50	8	52.48	9	56.44	12	58.42
Der	34	95.04	12	97.52	11	98.35	11	98.35	11	97.52
Eco	8	80.70	7	80.70	7	80.70	7	80.70	7	80.70
Heart	14	74.44	9	73.33	8	75.56	8	74.44	9	75.56
Ion	35	94.02	7	92.31	7	92.31	8	90.60	12	89.74
Son	61	61.43	6	68.57	6	58.57	8	75.71	9	74.29
Win	14	91.67	4	86.67	4	88.33	4	86.67	6	90.00
Average		78.80		77.90		78.21		79.29		80.60

Sometimes the number of the attributes finally used in training or learning is determined in advance. In some cases, the users have some prior knowledge about the structure of the data; they are able to select an appropriate similarity function, whereas the experiments shown above were conducted without any prior knowledge. The convergence and good accuracies are observed in the results.

6. Conclusion

Rough set theory has proven to be a powerful tool for feature subset selection and rule extraction. The classical rough set model just works in nominal domain. In this paper, we propose a novel information measure, which can measure the discernibility power of a crisp equivalence relation and fuzzy one. And it is proven that when relation matrix is a crisp equivalence one, the proposed entropy will be degraded to Shannon's entropy. Based on the proposed entropy, some basic definitions in fuzzy-rough set model are presented. Two reduction algorithms for unsupervised and supervised dimensionality reduction are given. Experiments show the algorithms get the same results as that of the classical rough set approaches when the attributes of data are all nominal. However, the performance of the proposed reduction is better than the classical methods with respect to hybrid data.

References

Beynon, M., 2001. Reducts within the variable precision rough sets model: a further investigation. Euro. J. Oper. Res. 134 (3), 592–605.

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

Blum, A.L., Langley, P., 1997. Selection of relevant features and examples in machine learning. Artificial Intell. 97, 245–271.

Chen, S., Zhu, Y., 2004. Subpattern-based principle component analysis. Pattern Recognit. 37 (5), 1081–1083.

Cheung, Y., Xu, L., 2001. Independent component ordering in ICA time series analysis. Neurocomputing 41 (1–4), 145–152.

Chmielewski, M., Grzymala-Busse, J., 1996. Global discretization of continuous attributes as preprocessing for machine learning. Internat. J. Approx. Reason. 15 (4), 319–331.

Dash, M., Liu, H., 2003. Consistency-based search in feature selection. Artificial Intell. 151, 155–176.

Dubois, D., Prade, H., 1992. Putting fuzzy sets and rough sets together. In: Slowiniski, R. (Ed.), Intelligent Decision Support. Kluwer Academic, Dordrecht, pp. 203–232.

Duch, W. et al., 2002. Feature selection based on information theory, consistency and separability indices. Proc. 9th Neural Information Processing 4, 1951–1955.

Guyon, I., Elisseeff, A., 2003. An introduction to variable and feature selection. J. Machine Learning Res. 3, 1157–1182.

Hand, D., Mannila, H., Smyth, P., 2001. Principles of Data Mining. MIT publisher.

Hu, Q., Yu, D., 2004. Entropies of fuzzy indiscerniblity relation and its operations. Int. J. Uncertainty, Fuzziness and Knowledge-Based Systems 12 (5), 575–589.

Hu, Q., Yu, D. (in press). Information measures on approximation spaces. Int. J. Uncertainty, Fuzziness and Knowledge Based Systems.

Hu, Q., Yu, D., Xie, Z., Liu, J. (in press). Fuzzy probabilistic approximations and their information measures. IEEE Trans. Fuzzy Systems.

Hu, Q., Yu, D., Xie, Z., 2005. Hybrid data reduction for classification with a fuzzy-rough set technique. In: 5th SIAM Conf. on Data Mining.

Hwang, K., Chang, C., 2002. A fast pixel mapping algorithm using principal component analysis. Pattern Recognition Lett. 23 (14), 1747–1753.

Jensen, R., Shen, Q., 2004. Fuzzy-rough attribute reduction with application to web categorization. Fuzzy Sets Syst. 141, 469–485.

Kira, K., Rendell, L.A., 1992. The feature selection problem: traditional methods and a new algorithm, In: Proc. AAAI-92, pp. 129–134.

Kohavi, R., John, G.H., 1997. Wrappers for feature subset selection. Artificial Intell. 97, 73–324.

Kwak, N., Choi, C.-H., 2002. Input feature selection for classification problems. IEEE Trans. Neural Networks 13 (1), 143–159.

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

Li, D., Cheng, C., 2002. New similarity measures of intuitionistic fuzzy sets and application to pattern recognitions. Pattern Recognition Lett. 23 (1–3), 221–225.

Li, H., Xu, L.D., 2001. Feature space theory—a mathematical foundation for data mining. Knowledge-Based Syst. 14, 253–257.

Li, D., Zhang, B., Leung, Y., 2004. On knowledge reduction in inconsistent decision information systems. Int. J. Uncertainty, Fuzziness and Knowledge Based Systems 12 (5), 651–672.

Liu, H., Setiono, R., 1998. Some issues on scalable feature selection. Expert Syst. Appl. 15, 333–339.

Liu, H., Motoda, H., Yu, L. , 2002. Feature selection with selective sampling. In: Proc. 19th ICML. Sydney, pp. 395–402.

Mi, J., Wu, W., Zhang, W., 2004. Approaches to knowledge reduction based on variable precision rough set model. Informat. Sci. 159 (3–4), 255–272, 15.

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

Moradi, H., Grzymala-Busse, J.W., Roberts, J.A., 1998. Entropy of english text: experiments with humans and a machine learning system based on rough sets. Informat. Sci. 104 (1–2), 31–47.

Morsi, N.N., Yakout, M.M., 1998. Axiomatics for fuzzy-rough sets. Fuzzy Sets Syst. 100 (1–3), 327–342.

Pawlak, Z., 1991. Rough Sets—Theoretical Aspects of Reasoning about Data. Kluwer Academic Publishers.

Piramuthu, S., 2004. Evaluating feature selection methods for learning in data mining applications. Euro. J. Oper. Res. 156, 483–494.

Shen, Q., Chouchoulas, A., 2002. A rough-fuzzy approach for generating classification rules. Pattern Recognit. 35 (11), 2425–2438.

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

Skowron, A., Rauszer, C., 1992. The discernibility matrices and functions in information systems. Intelligent Decision Support: Handbook of Applications and Advances of Rough Set Theory, 331–362.

Srinivasan, P., Ruiz, M.E., Kraft, D.H., Chen, J., 2001. Vocabulary mining for information retrieval: rough sets and fuzzy sets. Informat. Process. Mgmt. 37 (1), 15–38.

Swiniarski, R.W., Larry, H., 2001. Rough sets as a front end of neural networks texture classifier. Neurocomputating 36, 85–102.

Swiniarski, R.W., Skowron, A., 2003. Rough set methods in feature selection and recognition. Pattern Recognition Lett. 24 (6), 833–849.

Torkkola, K., 2003. Feature extraction by non-parametric mutual information maximization. J. Mach. Learn. Res. 3, 1415–1438.

Tsang, E.C.C., Yeung, D.S., Wang, X.Z., 2003. OFFSS: Optimal fuzzy-valued feature subset selection. IEEE Trans. Fuzzy Syst. 11 (2), 202–213.

Wakako, H., 2002. Separation of independent components from data mixed by several mixing matrices. Signal Process. 82 (12), 1949–1961.

Wang, Y., 2003. Mining stock price using fuzzy-rough set system. Expert Syst. Appl. 24 (1), 13–23.

Wang, J., Miao, D., 1998. Analysis on attribute reduction strategies of rough set. J. Comput. Sci. Technol. 13 (2), 189–193.

Wang, G., Hu, H., Yang, D., 2002. Decision table reduction based on conditional information entropy. Chin. J. Comput. 25 (7), 1–8.

Wang, Q., Shen, Y., Zhang, Y., et al., 2003. A quantitative method for evaluating the performances of hyperspectral image fusion. IEEE Trans. Instrument. Measur. 52 (4), 1041–1047.

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

Wu, W., Mi, J., Zhang, W., 2004. Generalized fuzzy-rough sets. Informat. Sci. 151, 263–282.

Yu, L., Liu, H., 2003a. Feature selection for high-dimensional data: a fast correlation-based filter solution. In: Proc. 20th Internat. Conf. on Machine Leaning, pp. 856–863.

Yu, L., Liu H., 2003b. Efficiently handling feature redundancy in high-dimensional data. In: Proc. 9th ACM SIGKDD Internat. Conf. on Knowledge Discovery and Data Mining, August 24–27, pp. 685–690.