Hybrid Attribute Reduction for Classification Based on A Fuzzy Rough Set Technique

Qinghua Hu* Daren Yu† Zongxia Xie‡

Abstract

Data usually exists with hybrid formats in real-world applications, and a unified data reduction for hybrid data is desirable. In this paper a unified information measure is proposed to computing discernibility power of a crisp equivalence relation and 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 traditional rough set approaches.

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 microarray data and text automatic categorization, where the number of features in the raw data ranges from hundreds to tens of thousands[1].Such high dimensionality brings great difficulty to pattern recognition, machine learning and data mining [2, 3]. Data reduction is a well-known data mining problem which is usually considered as an enhancement preprocessing technique for subsequent machining [4]. 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 [5, 6]. A lot of data reduction techniques are 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 [7, 8]. This paper will be focused on the second problem.

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 [9]. Feature extraction refers to constructing new features by 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 input space and delete the others. Principal component analysis (PCA) [10, 11, 12], Independent component analysis (ICA)[13, 14], 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 [4].

In last decade, much attention has been paid to feature subset selection. Two extensive reviews were published [7, 15] in artificial intelligence and a special issue of machine learning research was present in 2003 [1]. 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 subset is valid [16]. According to evaluation methods the feature subset selection can classified into two kinds: filtering and wrapper. Distance measures [17, 18], information measures [19, 20, 21], correlation coefficient [22] and consistency measures [6] are used for filtering methods. Wrapper refers to using a classifier as evaluation function in selection. KNN, neural network, SVM all can be employed. Isabelle Guyon [1] pointed that se-

*Harbin Institute of Technology, China.
†Harbin Institute of Technology, China.
‡Harbin Institute of Technology, China.

lecting 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 doesn't 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 which 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 in nature because there is no distance measure in the nominal domain. Data usually comes with a hybrid form in applications. For example, nominal attributes: sex, color, numeric attributes: age, temperature are coexist in hospital data. The above selection methods are suitable for a single format of features in nature. A feature subset selection for hybrid data is desirable in applications.

Rough set theory has proved to be a powerful tool for 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. So rough set methodology based dimensionality reduction will get a good feature subset. Some rough set based reduction and feature selection algorithms have been proposed. Consistency of data [24, 25], dependency of attributes [26], mutual information [27], discernibility matrix [28] and genetic algorithm are employed to find reducts of an information system [29]. And these techniques are applied to text classification [30], face recognition [3], texture analysis [31] and process monitoring [32]. An extensive review about rough set based feature selection was given in [33].

As we know, Pawlak's rough set model [26] works in case that only nominal attributes exist in information systems. However, data usually comes with a hybrid form. Nominal attributes, fuzzy attributes and numeric features coexist in real-world databases. Some generalizations of the model were proposed to deal with the problem. Rough set theory and fuzzy set theory were putted together and rough fuzzy sets and fuzzy rough sets were defined in [34]. The properties and axiomatization of fuzzy rough set theory [35, 36] were analyzed in detail. And the fuzzy rough set methods were applied to mining stock price [37], vocabulary for information retrieval [38] and fuzzy decision rules [39].

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 [40]. Qiang Shen etc pointed that this processing may lead to some information loss in the original data. A fuzzy-rough attribute reduction, called fuzzy-rough QUICKREDUCT algorithm, was given in [42] 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 §2. A novel information measure and its properties will be presented in §3. §4 gives another definition of dependency of attribute set and reduction algorithms for hybrid data. An extensive experimental analysis is described in §5. §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.

DEFINITION 2.1. Given a non-empty finite set X, R is a relation defined on X,denoted by a relation matrix $M(R)$:

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

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

R is a fuzzy equivalence relation,if $\forall x, y, z \in X$,R satisfies:

1) Reflextivity: $R(x,x) = 1, \forall x \in U$;

2) Symmetry: $R(x,y) = R(y,x), \forall x, y \in U$;

3) Transitivity: $R(x,z) \geq \min_y \{R(x,y), R(y,z)\}$.

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

A crisp equivalence relation will generate a crisp partition and a fuzzy equivalence relation generates a fuzzy partition.

DEFINITION 2.2. The fuzzy equivalence classes generated by a fuzzy equivalence relation R is defined as

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

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

THEOREM 2.1. *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) $R(x,y) = 0 \Leftrightarrow [x]_R \cap [y]_R = 0$;
2) $\underset{x \in X}{\vee} [x]_R = 1$;
3) $R(x,y) = 1 \Leftrightarrow [x]_R = [y]_R$;

DEFINITION 2.3. Given a fuzzy approximation space $< U, R >$, 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

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

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

$$\mu_{POS_B(d)} = \underset{X \subseteq U/d}{sup} \mu_{\underline{B}X}(x).$$

DEFINITION 2.4. Given a fuzzy information system $< U, A >$, B and d are two subset of attribute set A, the dependency degree of d to B is defined as

$$\gamma_B(d) = \textstyle\sum_{x \in U} \mu_{POS_B(d)}(x).$$

DEFINITION 2.5. Given a fuzzy information system $< U, A, V, f >$, $B \subseteq A$,$a \in B$, 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 2.6. Given a fuzzy information system $< U, A, V, f >$, $A = C \cup d$. 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 $a \in B$ 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) $\forall a \in B : \ \gamma_{B-a}(d) < \gamma_C(d)$.

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_A on U. The fuzzy relation matrix $M(R_A)$ is denoted by

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

where $r_{ij} \in [0,1]$ is the relation value of x_i and x_j. In fact, the nominal attribute is a special case, where $r_{ij} \in \{0,1\}$, which will generate a crisp equivalence relation.

DEFINITION 3.1. The quotient set generated by an equivalence relation is defined as

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

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

DEFINITION 3.2. The cardinality $|[x_i]_R|$ of $[x_i]_R$ is defined as

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

DEFINITION 3.3. Information quantity of the fuzzy attribute set or the fuzzy equivalence relation is defined as

$$H(R_A) = -\frac{1}{n} \sum_{i=1}^{n} log\lambda_i.$$

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

Property 1. If A is a nominal attribute,$M(R_A)$ is the relation matrix generated by A, $H(A)$ denotes the Shannon information quantity, and $H(R_A)$ is the information value computed according to definition 3.3, and then we have

$$H(A) = H(R_A) \quad .$$

According property 1, 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:$H : R \to R^+$, where R is a equivalence relation matrix, R^+ 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 3.4. Given a fuzzy information system $< U, A, V, f >$, A is the fuzzy or numeric 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 joint entropy of B and E is defined as

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

DEFINITION 3.5. Given a fuzzy information system $< U, A, V, f >$, 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

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

THEOREM 3.1. $H(E|B) = H(BE) - H(B)$

THEOREM 3.2. *Given a fuzzy information system $< U, A, V, f >$, 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 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) $\forall B \subseteq A : \ H(B) \geq 0$;
2) $H(BE) \geq \max\{H(B), H(E)\}$;
3) $B \supseteq E$ or $R_B \subseteq R_E : \ H(BE) = H(B)$;
4) $B \supseteq E$ or $R_B \subseteq R_E : \ H(E|B) = 0$;

The first item of theorem 3.1 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 discrnibility 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 has 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 3.3. *Given a fuzzy information system $< U, A, V, f >$,$B \subseteq A$,$a \in B$,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 3.4. *Given a fuzzy information system $< U, A, V, f >$,$A = C \cup d$. 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 3.3 and 3.4 give the definitions of dependency, reduct and relative reduct in terms of information theory, while definitions 2.5 and 2.6 are defined in terms of algebra. In fact two classes of definitions are equivalent. The proof was given in [50].

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 [1, 25, 27]. 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, 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,

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

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

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

where function f should satisfy:

1) $f(0) = 1, f(\propto) = 0$ *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.

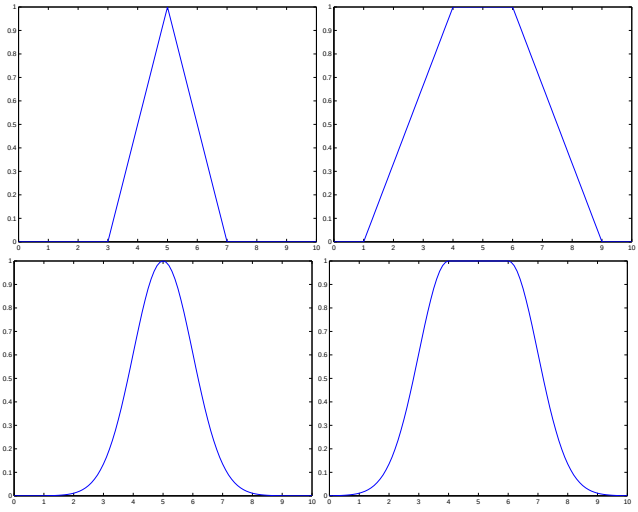

Figure 1: some similarity relation functions for numeric data.

As to fuzzy attributes, there are a great many candidate similarity measures [47]. 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:

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

Employing a max-min closure operation, we can get a fuzzy equivalence relation [48].

As has pointed in §3, 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 4.1. Given a fuzzy information system $< U, A, V, f >$, $B \subseteq A$,$a \in B$, the significance of attribute a in attribute set B is defined as

$$SIG(a, B) = H(B) - H(B - a)$$

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 4.2. Given a fuzzy information system $< U, A, V, f >$, $A = C \cup d$, 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

$$SIG(a, B, d) = H(d|B - a) - H(d|B)$$

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 $< U, A, V, f >$.

Output: One reduct of IS

Step 1: $\forall a \in A$:compute the equivalence relation;

Step 2: $\phi \rightarrow$ 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 $red \cup a \rightarrow red$ goto step3,Else return,End

Algorithm 2: Algorithm for calculating relative reduct.

Input: Information system IS $< U, A = C \cup d, V, f >$.

Output: One relative reduct D_red of IS

Step 1: $\forall a \in A$ compute the equivalence;

Step 2: $\phi \rightarrow D\_red$;

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 $D\_red \cup a \rightarrow D\_red$ goto step3,Else return D_red End

R. Jensen [42] 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 $SIG(a, red) \leq \lambda(SIG(a, red, d) \leq \lambda)$ 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 the following table. We find the attributes of data BC and BCW are nominal, and others are hybrid.

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 [1]. Ranking methods employ an evaluation function, such as inter-class distance, correlation criteria, mutual 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 [1].

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, 12 are selected one by one as a reduct, called subset 1.

In order to compare two feature subset selection, top six attributes 7, 13, 12, 10, 1, 11 are selected in ranking, called subset 2. Figures 2 and 3 show the distribution of data in 2-D feature space. Figure 2 is the distribution with attribute 7, 1, 1, 11, 11, 6, 6, 3, 3, 12, respectively. And Figure 3 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 SVM as a validation function for feature selection. 2/3 samples are randomly selected as training set, and the others are test set.

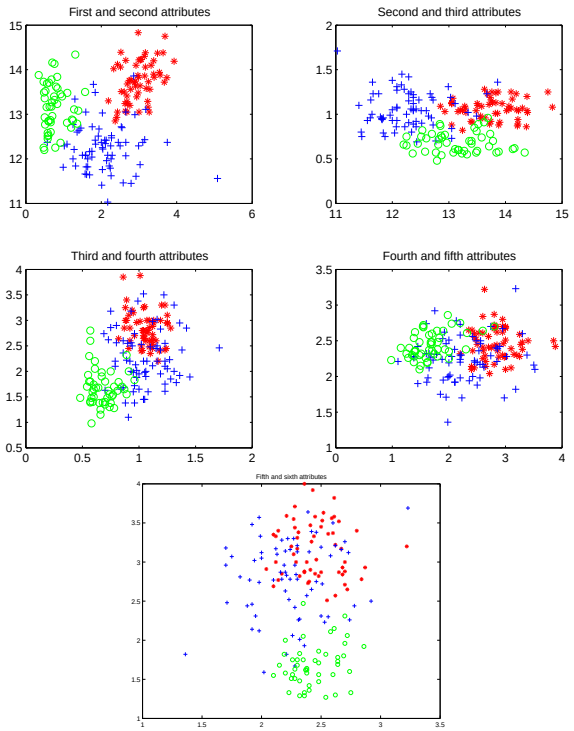

Figure 2: Distribution of wine samples with attributes 7, 1, 11, 6, 3, 12, Accuracy: 94.87%.

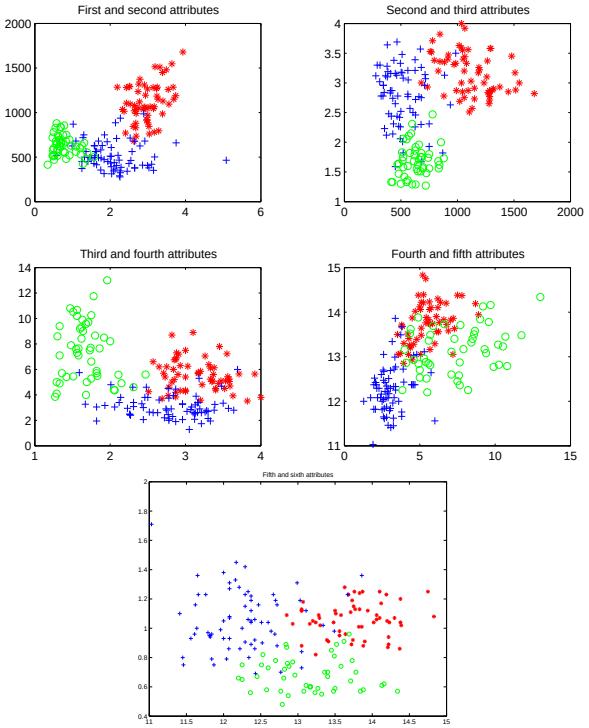

Figure 3: Distribution of wine samples with attributes 7, 13, 12, 10, 1, 11, Accuracy: 93.33%.

Table 1: Summary of the experiment data sets.

Data set			Class	Attribute number		
Abr.	Original name	Size	Number	Total	Numeric	Nominal
BC	Breast cancer	286	2	10	0	10
BCW	Breast-cancer-wisconsin1	699	2	10	0	10
WDBC	Breast-cancer-wisconsin2	569	2	31	30	1
WPBC	Breast-cancer-wisconsin3	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	61	60	1
Win	Wine Recognition	178	3	14	13	1
Vow	Vowel Database	990	11	11	10	1

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

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 speaking, ranking method only computes the dependency between condition attributes and decision attribute, while neglect the dependency among condition attributes. Let's analyze the correlation between the selected condition attributes. Correlation coefficients are showed in table 2 and 3. Wang [49] 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.

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 classifier as the validation function. The experiment data is shown in table 4.

The classical rough set theory works in nominal domain. We perform discretization on numeric data. The numeric attributes are discretized into three intervals by equal-width, equal-frequency and fuzzy c-means clustering. As to fuzzy-rough reduction algorithm, the relation matrices are computed with a triangle function. The numbers of selected attributes and accuracy of classification with SVM are shown in table 4.There is no numeric attribute in data sets BC and BCW. From the table we find the results of reduction and classification

Table 4: Comparisons of Fuzzy-rough technique vs. discritization with SVM classifiers.

	Original data		Reduct(Equi-width)		Reduct(Equi-Frequency)		Reduct(FCM)		Reduct(Fuzzy-rough)	
Data	n	Accuracy	n	Accuracy	n	Accuracy	n	Accuracy	n	Accuracy
BC	10	71.58%	8	72.63%	8	72.63%	8	72.63%	8	72.63%
BCW	10	98.28%	4	98.71%	4	98.71%	4	98.71%	4	98.71%
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	94.21%
Eco	8	70.18%	7	70.18%	7	70.18%	7	70.18%	7	70.18%
Gls	9	61.97%	7	64.79%	6	54.93%	8	63.38%	8	63.38%
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%
Vow	11	59.09%	10	63.94%	10	63.94%	10	63.94%	10	63.94%
Average		79.42%		78.58%		77.46%		79.30%		79.92%

with classical rough set method and the fuzzy one are identical, respectively, which shows that the method we proposed can degenerate to the classical case.

6 Conclusions.

Rough set theory has proven 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 the 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

[1] Isabelle Guyon and Andre Elisseeff,*An introduction to variable and feature selection*, Journal of machine learning research, 3 (2003), pp. 1157–1182.

[2] David Hand, Heikki Mannila and Padhraic Smyth,*Principles of data mining*, MIT publisher, 2001.

[3] H. Liu and R. Setiono,*Some issues on scalable feature selection*,Expert systems with applications, 15 (1998),pp. 333–339.

[4] E.C.C. Tsang, D.S. Yeung and X. Z. Wang,*OFFSS: Optimal fuzzy-valued feature subset selection*, IEEE transactions on fuzzy systems,2(2003), pp. 202–213.

[5] Kari Torkkola,*Feature extraction by non-parametric mutual information maximization*, Journal of machine learning research, 3 (2003),pp. 1415-1438.

[6] M. Dash and H. Liu,*Consistency-based search in feature selection*,AI 151(2003),pp. 155–176.

[7] Avrim L. Blum and Pat Langley,*Selection of relevant features and examples in machine learning*,Artificial intelligence 97(1997) ,pp. 245-271.

[8] H. Liu, H. Motoda and L. Yu, *Feature Selection with Selective Sampling*, Proceedings of the 19th ICML, July 8-12, 2002, Sydney, pp. 395–402

[9] H. X. Li and L. D. Xu,*Feature space theory-a mathematical foundation for data mining*, Knowledge-based systems 14 (2001) ,pp. 253-257.

[10] Hwang, Kuo-Feng, Chang and Chin-Chen,*A fast pixel mapping algorithm using principal component analysis*, Pattern Recognition Letters Volume: 23, Issue: 14, December, 2002, pp. 1747–1753.

[11] Gilmour, Justin and Wang Liuping,*Detection of process abnormality in food extruder using principle component analysis*, Chemical Engineering Science Volume: 57, Issue: 7, April, 2002, pp. 1091-1098.

[12] Chen Songcan and Zhu Yulian,*Subpattern-based principle component analysis*, Pattern Recognition Volume: 37, Issue: 5, May, 2004, pp. 1081–1083.

[13] Cheung, Y. and Xu, L,*Independent component ordering in ICA time series analysis*, Neurocomputing Volume: 41, Issue: 1-4, October, 2001, pp. 145–152.

[14] Wakako H,*Separation of independent components from data mixed by several mixing matrices*, Signal processing. Vol.82, No.12, 2002, pp. 1949–1961.

[15] Ron Kohavi and George H. John,Wrappers for feature subset selection, AI 97 (1997),pp. 73–324.

[16] Selwyn Piramuthu,Evaluating feature selection meth-

ods for learning in data mining applications, European journal of operational research,156 (2004),pp. 483–494.

[17] K. Kira and L.A. Rendell, *The feature selection problem: Traditional methods and a new algorithm*,Proceedings of AAAI-92, 1992, pp. 129–134.

[18] Kwak, N. Chong-Ho Choi,*Input feature selection for classification problems*, IEEE transaction on neural networks, Vol.13, No.1, 2002,pp. 143–159.

[19] Lei Yu, Huan Liu,*Efficiently handling feature redundancy in high dimensional data*,In Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, August 24 - 27(2003),pp. 685–690.

[20] W. Duch, et al,*Feature selection based on information theory, consistency and separability indices*, Proceeding on 9th neural information processing, vol.4(2002),pp. 1951–1955.

[21] L. Yu and H. Liu.*Feature Selection for High-Dimensional Data: A Fast Correlation-Based Filter Solution*, In Proceedings of The Twentieth International Conference on Machine Leaning (ICML-03), August 21-24(2003),pp. 856–863.

[22] P. Mitra. C.A. Murthy, S. K. Pal, *Unsupervised feature selection using feature similarity*, IEEE transactions on pattern analysis and machine intelligence,Vol. 24, No. 3(2002),pp. 301–312.

[23] Beynon, Malcolm,*Reducts within the variable precision rough sets model: A further investigation*, European Journal of Operational Research Volume: 134, Issue: 3, November 1(2001), pp. 592–605.

[24] Mi, Ju-Sheng; Wu, Wei-Zhi and Zhang, Wen-Xiu,*Approaches to knowledge reduction based on variable precision rough set model*, Information Sciences. Vol. 159, Issue: 3-4, 15(2004),pp. 255–272.

[25] Pawlak Z,*rough sets-theoretical aspects of reasoning about data. Kluwer academic publishers*, 1991.

[26] Wang G., Hu H., Yang D.,*Decision table reduction based on conditional information entropy*, Chinese journal of computers. Vol. 25, No. 7, 1-8(2002).

[27] Skowron A. Rauszer C.,*the discernibility matrices and functions in information systems*,Intelligent decision support: handbook of applications and advances of rough set theory, 1992,pp. 331–362.

[28] Wang Jue, Miao Duo-Qian,*Analysis on attribute reduction strategies of rough set*, Journal of computer science and technology. Vol. 13, No.2, 1998,pp. 189–193.

[29] Moradi, Hamid; Grzymala-Busse, Jerzy W.; Roberts, James A. ,*Entropy of English Text: Experiments with Humans and a Machine Learning System Based on Rough Sets*, Information Sciences Volume: 104, Issue: 1-2, January, 1998, pp. 31–47.

[30] Swiniarski, Roman W. Larry Hargis,*Rough sets as a front end of neural networks texture classifier*,Nurocomputating.,36(2001) pp. 85–102.

[31] Swiniarski, Roman W.; Skowron, Andrzej,*Rough set methods in feature selection and recognition*,Pattern Recognition Letters Volume: 24, Issue: 6, March, 2003, pp. 833–849.

[32] D. Dubois, H. Prade,*Putting fuzzy sets and rough sets together*,R. Slowiniski (Ed.), Ittelligent Decision support, Kluwer Academic, Dordrecht, 1992, pp. 203–232.

[33] Morsi, Nehad N.; Yakout, M.M.,*Axiomatics for fuzzy rough sets*,Fuzzy Sets and Systems Volume: 100, Issue: 1-3, November 16, 1998, pp. 327–342.

[34] Radzikowska, Anna Maria; Kerre, Etienne E.,*A comparative study of fuzzy rough sets*, Fuzzy Sets and Systems Vol.126, No.2, 2002, pp. 137–155.

[35] Wu, Wei-Zhi; Mi, Ju-Sheng; Zhang, Wen-Xiu,*Generalized fuzzy rough sets*. Information Sciences Volume: 151, May, 2003, pp. 263–282.

[36] Wu, Wei-Zhi; Zhang, Wen-Xiu,*Constructive and axiomatic approaches of fuzzy approximation operators*,Information Sciences Volume: 159, Issue: 3-4, February 15, 2004, pp. 233–254.

[37] Wang Yi-Fan,*Mining stock price using fuzzy rough set system*, Expert Systems with Applications Volume: 24, Issue: 1, January, 2003, pp. 13–23.

[38] Srinivasan, Padmini; Ruiz, Miguel E.; Kraft, Donald H.; Chen, Jianhua,*Vocabulary mining for information retrieval: rough sets and fuzzy sets*, Information Processing and Management Volume: 37, Issue: 1, January 1, 2001, pp. 15–38.

[39] Q. Shen and A. Chouchoulas,*A rough-fuzzy approach for generating classification rules*,Pattern Recognition, 35(11)(2002)pp. 2425–2438.

[40] Chmielewski, Michal R.; Grzymala-Busse, Jerzy W.,*Global Discretization of Continuous Attributes as Preprocessing for Machine Learning*,International Journal of Approximate Reasoning Volume: 15, Issue: 4, November, 1996, pp. 319–331.

[41] Roy, Amitava; Pal, Sankar K.,*Fuzzy discretization of feature space for a rough set classifier*,Pattern Recognition Letters V. 24, No.6(2003), pp. 895–902.

[42] R. Jensen, Q. Shen,*Fuzzy-rough attribute reduction with application to web categorization*,Fuzzy sets and systems,141 (2004),pp. 469–485.

[43] L. Zadeh,*Probability measures of fuzzy events*,J. Math. Anal. Appl.23(1965),pp. 421–427.

[44] Yager, Ronald R.,*Measures of Entropy and Fuzziness Related to Aggregation Operators*,Information Sciences Volume: 82, Issue: 3-4, January, 1995, pp. 147–166.

[45] Bertoluzza, Carlo; Doldi, Viviana; Naval, Gloria.,*Uncertainty measure on fuzzy partitions*,Fuzzy Sets and Systems Vol.142, No.1, 2004, pp. 105–116.

[46] Guo, Caimei and Zhang, Deli,*On set-valued fuzzy measures*, Information Sciences Volume: 160, Issue: 1-4, March 22, 2004, pp. 13–25.

[47] Dengfeng, Li; Chuntian, Cheng.,*New similarity measures of intuitionistic fuzzy sets and application to pattern recognitions*,Pattern Recognition Letters Volume: 23, Issue: 1-3, January, 2002, pp. 221–225.

[48] Lee, Hsuan-Shih. *An optimal algorithm for computing the max-min transitive closure of a fuzzy similarity matrix*, Fuzzy Sets and Systems Vol.123, No.1(2001), pp. 129–136.

[49] Qiang Wang, Yi Shen, Ye Zhang,*A fast method to evaluate the performance of image fusion techniques and its error analysis*, Instrumentation and measurement technology conference, 2003.

[50] Qinghua Hu and Daren Yu,*Entropies of fuzzy indiscerniblity relation and its operations*, International Journal of uncertainty, fuzziness and knowledge-based systems. Vol. 12, No. 5,pp. 575–589.

[51] Qinghua Hu,Daren Yu and Zongxia Xie,*Reduction algorithms for hybrid data based on fuzzy rough set approaches*,International Conference on Machine Learning and Cybernetics(2004),pp. 1469–1474.