Fuzzy Mutual Information Based min-Redundancy and Max-Relevance Heterogeneous Feature Selection

Daren Yu, Shuang An, Qinghua Hu

Harbin Institute of Technology, Harbin 150001, China
E-mail: huqinghua@hit.edu.cn

Abstract

Feature selection is an important preprocessing step in pattern classification and machine learning, and mutual information is widely used to measure relevance between features and decision. However, it is difficult to directly calculate relevance between continuous or fuzzy features using mutual information. In this paper we introduce the fuzzy information entropy and fuzzy mutual information for computing relevance between numerical or fuzzy features and decision. The relationship between fuzzy information entropy and differential entropy is also discussed. Moreover, we combine fuzzy mutual information with "min-Redundancy-Max-Relevance", "Max-Dependency" and "min-Redundancy-Max-Dependency" algorithms. The performance and stability of the proposed algorithms are tested on benchmark data sets. Experimental results show the proposed algorithms are effective and stable.

Keywords: Feature selection; fuzzy mutual information; redundancy; relevance; stability

1. Introduction

As the capability of acquiring and storing information increases, more and more candidate features are gathered in pattern recognition and machine learning. Unfortunately, most of these features are usually irrelevant or redundant for a given learning task. These irrelevant or redundant features may confuse learning algorithms and deteriorate learning performance. So it is useful to select a subset of relevant and indispensable features for designing effective classification systems.

A great number of feature selection algorithms based on mutual information have been developed in recent years [1,7,11,18,19,20,24,37,40,42]. In constructing a feature selection algorithm, there are two key issues: evaluation measure and search strategy. An evaluation measure is used to measure the significance of features. A number of measures have been developed so far, such as dependency [28,41,46], consistency [6,32], fuzzy dependency [12] and mutual information [1]. Mutual information was firstly introduced to measure relevance between discrete variables. Subsequently, it was widely used to measure feature quality in feature selection [11,24,47]. As to search strategy, it can be divided into two categories. One category could guarantee to find the optimal feature subset, such as the exhaustive search method [23] and the branch-and-bound algorithm [26,39], and the other one is to find the suboptimal feature subset. The second category covers a wide range of heuristic search strategies, such as sequential forward selection [16,45], sequential backward elimination [23], floating search [31,38], hill-climbing [10,30], best-first or beam search [34] and min-Redundancy-Max-Relevance (mRMR) [29]. Especially, mRMR is considered as an effective one.

Published by Atlantis Press

It just requires estimating binary probability density for computing mutual information between a feature and decision instead of multivariable probability density. Moreover, mRMR method removes redundant features by considering the mutual information between features.

As we know, in Shannon's mutual information, probabilities are unknown in practice and should be estimated with a finite set of samples. As to symbol variables, we can use the frequency of samples to estimate the probabilities. As to continuous variables, there are two methods for obtaining probabilities. One is to discretize real variables [2], and the other one is to estimate probability density with Parzen window [18]. It is observed that discretization may lead to information loss [36] and it is time consuming to estimate probability density with Parzen window. Furthermore, we can not get accurate estimate in high-dimensional spaces with sparse samples. By considering the above problems, Hu et al. proposed fuzzy information entropy to directly compute mutual information between numerical or fuzzy variables based on relation matrixes [13]. As to discrete variables, a crisp equivalence relation matrix can be generated from data. In this case, the fuzzy information entropy is identical to Shannon's one. And for continuous variables, fuzzy binary relations are used to compute relation matrix. By this way, fuzzy information entropy is directly calculated from discrete and continuous variables without discretization. Consequently, the fuzzy mutual information derived from fuzzy information entropy can directly compute relevance between numerical features. Hu et al. combined this measure with a greedy forward search strategy [12]. As we know greedy algorithms are suboptimal and may not produce good performance in applications [5], we integrate fuzzy mutual information with some other search strategies to construct better algorithms.

In this work, we first discuss the relationship between fuzzy information entropy and differential entropy and find they are identical in some cases. Secondly, we integrate fuzzy mutual information with mRMR algorithm denoted by FMI_mRMR and maximal dependency algorithm denoted by FMI_MD. Considering the redundancy between features for FMI_MD algorithm, we combine minimal redundancy with FMI_MD denoted by FMI_mRMD. Then we test the three algorithms with experiments on 14 data sets and do some comparison analysis between them in terms of classification accuracies of feature subsets. Moreover, we compare the three algorithms with MI_mRMR (mRMR algorithm based on Shannon's mutual information), CFS, FCBF and RELIEF. Finally, we analyze the stability of FMI_mRMR, FMI_MD and FMI_mRMD algorithms.

The paper is organized as follows. Section 2 introduces fuzzy information entropy and fuzzy mutual information. Section 3 discusses the relationship between fuzzy information entropy and differential entropy. Next, we introduce three feature selection algorithms with fuzzy mutual information in Section 4 and give several evaluation measures of stability for algorithms in Section 5. Finally, experimental analysis and conclusions are given in Sections 6 and 7, respectively.

2. Fuzzy information entropy and fuzzy mutual information

Information entropy was originally introduced for measuring the uncertainty of random variables [35]. As to discrete variables, the probability densities of variables in information entropy are computed with frequency. But for high-dimensional continuous variables, it is very difficult to estimate $p(\cdot)$ in practice. There are usually two ways to address this problem. One is to discretize the variables, and the other one is to estimate $p(\cdot)$ of the variables with Parzen window.

Considering the above problem, fuzzy information entropy was introduced to measure the uncertainty of random variables [12,13]. Now we first introduce this definition.

A fuzzy binary relation R, used to measure relationship between two variables, is a fuzzy equivalence relation if it satisfies

$$\begin{array}{l}(1)\text{Reflectivity}: R(x,y)=1, \forall x \in X;\\ (2)\text{Symmetry}: R(x,y)=R(y,x), \forall x,y \in X;\\ (3)\text{Transitivity}: R(x,z) \geqslant \min_y\{R(x,y),R(y,z)\}.\end{array} \tag{1}$$

Given a finite set $U = \{x_1, x_2, ..., x_n\}$, F is a fuzzy or real-valued attribute set, which generates a fuzzy equivalence relation R_F on U, denoted by a relation matrix $M(R_F)$

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

Fuzzy equivalence class associated with x_i and R_F is a fuzzy set which can be written as

$$[x_i]_{R_F} = \frac{r_{i1}}{x_1} + \frac{r_{i2}}{x_2} + \cdots + \frac{r_{in}}{x_n}, \tag{3}$$

where $r_{ij} = R_F(x_i, x_j) \in [0,1] (j = 1, 2, ..., n)$ is the fuzzy equivalence relation between x_i and x_j.

For a crisp equivalence relation R_c, the equivalence class of x_i also can be described as

$$[x_i]_{R_F} = \frac{r_{i1}}{x_1} + \frac{r_{i2}}{x_2} + \cdots + \frac{r_{in}}{x_n}, \tag{4}$$

where $r_{ij} = R_c(x_i, x_j) \in \{0, 1\} (j = 1, 2, ..., n)$ is the relation between x_i and x_j. This is because $R_c(x_i, x_j)$ satisfies

$$R_c(x_i, x_j) = \begin{cases} 1, \text{if} & x_i = x_j, \\ 0, \text{if} & x_i \neq x_j. \end{cases} \tag{5}$$

Example $X = \{x_1, x_2, x_3, x_4\}$, $f \in F$, $X_f = \{0.1, 0.1, 0.3, 0.3\}$. If we take

$$R_f(x, y) = \exp(\frac{-\parallel x - y \parallel}{2}),$$

as fuzzy equivalence relation, then

$[x_1]_{R_f} = \frac{1}{x_1} + \frac{1}{x_2} + \frac{0.90}{x_3} + \frac{0.90}{x_4}$;
$[x_2]_{R_f} = \frac{1}{x_1} + \frac{1}{x_2} + \frac{0.90}{x_3} + \frac{0.90}{x_4}$;
$[x_3]_{R_f} = \frac{0.90}{x_1} + \frac{0.90}{x_2} + \frac{1}{x_3} + \frac{1}{x_4}$;
$[x_4]_{R_f} = \frac{0.90}{x_1} + \frac{0.90}{x_2} + \frac{1}{x_3} + \frac{1}{x_4}$.

If R_f is a crisp equivalence relation,
$[x_1]_{R_f} = \{x_1, x_2\} = \frac{1}{x_1} + \frac{1}{x_2} + \frac{0}{x_3} + \frac{0}{x_4}$;
$[x_2]_{R_f} = \{x_1, x_2\} = \frac{1}{x_1} + \frac{1}{x_2} + \frac{0}{x_3} + \frac{0}{x_4}$;
$[x_3]_{R_f} = \{x_3, x_4\} = \frac{0}{x_1} + \frac{0}{x_2} + \frac{1}{x_3} + \frac{1}{x_4}$;
$[x_4]_{R_f} = \{x_3, x_4\} = \frac{0}{x_1} + \frac{0}{x_2} + \frac{1}{x_3} + \frac{1}{x_4}$.

Based on fuzzy equivalence relation, fuzzy information entropy is defined as

$$FH(R_F) = -\frac{1}{n}\sum_{i=1}^{n} \log \frac{|[x_i]_{R_F}|}{n}, \tag{6}$$

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

If relation R_F is a crisp equivalence relation, namely $r_{ij} \in \{0, 1\}$, the fuzzy information entropy can be educed from Shannon's information entropy. This is proven as follows.

Proof: If relation R_F is a crisp equivalence relation, $r_{ij} = 0$ means $x_i \neq x_j$, and $r_{ij} = 1$ means $x_i = x_j$. The equivalence class of x_i can be written as

$$[x_i]_F = [x_i]_{R_F} = \frac{r_{i1}}{x_1} + \frac{r_{i2}}{x_2} + \cdots + \frac{r_{in}}{x_n}. \tag{7}$$

Then we compute the probability of equivalence class using $P(X_i) = |X_i|/n$, where X_i is an equivalence class. Let probability distribution of equivalence classes be

$$\begin{bmatrix} X_1 & X_2 & \cdots & X_m \\ p(X_1) & p(X_2) & \cdots & p(X_m) \end{bmatrix}, \tag{8}$$

where $k_j = |X_j|$ $(j = 1, 2, ..., m)$. Shannon's information entropy equals

$$\begin{aligned} H(F) &= -\sum_{j=1}^{m} p(X_j) \log p(X_j) \\ &= -\sum_{j=1}^{m} \frac{k_j}{n} \log \frac{k_j}{n} \\ &= -\frac{k_1}{n} \log \frac{k_1}{n} - \frac{k_2}{n} \log \frac{k_2}{n} - \cdots - \frac{k_m}{n} \log \frac{k_m}{n} \\ &= -\frac{1}{n} \sum_{i=1}^{k_1} \log \frac{k_1}{n} - \frac{1}{n} \sum_{i=1}^{k_2} \log \frac{k_2}{n} - \cdots - \frac{1}{n} \sum_{i=1}^{k_m} \log \frac{k_m}{n} \\ &= (-\frac{1}{n}) \sum_{i=1}^{k_1} \log \frac{|[x_i]_{R_F}|}{n} + (-\frac{1}{n}) \sum_{i=k_1+1}^{k_1+k_2} \log \frac{|[x_i]_{R_F}|}{n} \\ &\quad + \cdots + (-\frac{1}{n}) \sum_{i=n-k_m+1}^{n} \log \frac{|[x_i]_{R_F}|}{n} \\ &= (-\frac{1}{n}) \sum_{i=1}^{n} \log \frac{|[x_i]_{R_F}|}{n} \\ &= FH(R_F), \end{aligned} \tag{9}$$

where $\sum_{j=1}^{m} k_j = n$. $|[x_i]_{R_F}|$ is the size of equivalence class $[x_i]_{R_F}$ $(i = 1, 2, ..., n)$.

We can see that fuzzy information entropy is identical to Shannon's one from (9) for crisp equivalence relation. Therefore, fuzzy information entropy also can be used to address discrete variables.

Let F_1 and F_2 be two subsets of F, fuzzy joint information entropy is defined as

$$\begin{aligned} FH(F_1,F_2) &= FH(R_{F_1},R_{F_2}) \\ &= -\frac{1}{n}\sum_{i=1}^{n}\log\frac{|[x_i]_{F_1}\cap[x_i]_{F_2}|}{n}, \end{aligned}\tag{10}$$

and fuzzy information entropy of $F = \{F_1,F_2,...,F_m\}$ is

$$\begin{aligned} FH(F) &= FH(F_1,...,F_m) = FH(R_{F_1},...,R_{F_m}) \\ &= -\frac{1}{n}\sum_{i=1}^{n}\log\frac{|[x_i]_{F_1}\cap...\cap[x_i]_{F_m}|}{n}. \end{aligned}\tag{11}$$

Given F_1, the fuzzy conditional information entropy of F_2 is defined as

$$\begin{aligned} FH(F_2|F_1) &= FH(R_{F_2}|R_{F_1}) \\ &= -\frac{1}{n}\sum_{i=1}^{n}\log\frac{|[x_i]_{F_1}\cap[x_i]_{F_2}|}{\left|[x_i]_{F_1}\right|}. \end{aligned}\tag{12}$$

It is proven that following properties hold [13].

Proposition 1 *Given a fuzzy information system $< U,F,V,f >$, F is the fuzzy attribute set, and $F_1,F_2 \subseteq F$. $[x_i]_{F_1}$ and $[x_i]_{F_2}$ are fuzzy equivalence classes of x_i generated by fuzzy equivalence relations R and S induced from F_1 and F_2, respectively. Then the following statements hold.*

$$\begin{aligned} &(1)\forall F_1 \subseteq F : FH(F_1) \geqslant 0; \\ &(2)FH(F_1,F_2) \geqslant \max\{FH(F_1),FH(F_2)\}; \\ &(3)F_1 \subseteq F_2 \quad or \quad R_{F_1} \subseteq R_{F_2} : FH(F_1,F_2) = FH(F_1); \\ &(4)F_1 \subseteq F_2 \quad or \quad R_{F_1} \subseteq R_{F_2} : FH(F_2|F_1) = 0; \\ &(5)FH(F_2|F_1) = FH(F_1|F_2) - FH(F_1); \\ &(6)FH(F_1|F_2) = FH(F_2|F_1) - FH(F_2); \end{aligned}\tag{13}$$

As the above properties of fuzzy information entropy and fuzzy mutual information are summarized and discussed by Hu et al. [13], we here do not present detailed analysis and discussion.

Once given the definition of fuzzy information entropy, we calculate mutual information using the following equations

$$FMI(F_1;F_2) = FH(F_1) - FH(F_1|F_2), \tag{14}$$

$$FMI(F_2;F_1) = FH(F_2) - FH(F_2|F_1), \tag{15}$$

$$FMI(F_1;F_2) = FH(F_1) + FH(F_2) - FH(F_1,F_2). \tag{16}$$

It is easy to know

$$FMI(F_1;F_2) = FMI(F_2;F_1). \tag{17}$$

By introducing (6) into (16), the fuzzy mutual information between F_1 and F_2 equals

$$\begin{aligned} FMI(F_1;F_2) &= -\frac{1}{n}\sum_{i=1}^{n}\log\frac{|[x_i]_{F_1}|}{n} - \frac{1}{n}\sum_{i=1}^{n}\log\frac{|[x_i]_{F_2}|}{n} \\ &\quad +\frac{1}{n}\sum_{i=1}^{n}\log\frac{|[x_i]_{F_1}\cap[x_i]_{F_2}|}{n} \\ &= -\frac{1}{n}\sum_{i=1}^{n}\log\frac{|[x_i]_{F_1}|\cdot|[x_i]_{F_2}|}{n\cdot|[x_i]_{F_1}\cap[x_i]_{F_2}|}. \end{aligned}\tag{18}$$

From above formula we can see fuzzy mutual information could be computed for both discrete and continuous variables. It overcomes the limitation of Shannon's mutual information.

Now, we use an example to illustrate the computation of fuzzy mutual information.

Given two continuous variables $X_1 = \{0.1,0.3,0.5,0.6\}$ and $X_2 = \{0.2,0.4,0.7,0.9\}$, we use

$$S(x,y) = \exp(-\parallel x-y \parallel) \tag{19}$$

to measure similarity. Relation matrices $M(R_{X_1})$ and $M(R_{X_2})$ are

$$M(R_{X_1}) = \begin{pmatrix} 1 & 0.82 & 0.67 & 0.61 \\ 0.82 & 1 & 0.82 & 0.74 \\ 0.67 & 0.82 & 1 & 0.90 \\ 0.61 & 0.74 & 0.90 & 1 \end{pmatrix}$$

and

$$M(R_{X_2}) = \begin{pmatrix} 1 & 0.82 & 0.61 & 0.50 \\ 0.82 & 1 & 0.74 & 0.61 \\ 0.61 & 0.74 & 1 & 0.82 \\ 0.50 & 0.61 & 0.82 & 1 \end{pmatrix}.$$

The fuzzy information entropy of X_1 and X_2 are

$$\begin{aligned} FH(X_1) &= -\frac{1}{4}(\log\frac{3.10}{4}+\log\frac{3.38}{4}+\log\frac{3.39}{4} \\ &\quad +\log\frac{3.25}{4}) \\ &= 0.29 \end{aligned}$$

and

$$\begin{aligned} FH(X_2) &= -\frac{1}{4}(\log\frac{2.93}{4}+\log\frac{3.17}{4}+\log\frac{3.17}{4} \\ &\quad +\log\frac{2.93}{4}) \\ &= 0.39. \end{aligned}$$

For the fuzzy joint information entropy $FH(X_1,X_2)$, we first compute the intersection of $M(R_{X_1})$ and $M(R_{X_2})$ i.e.

$$M(R_{X_1})\cap M(R_{X_2}) = \begin{pmatrix} 1 & 0.82 & 0.61 & 0.50 \\ 0.82 & 1 & 0.74 & 0.61 \\ 0.61 & 0.74 & 1 & 0.82 \\ 0.50 & 0.61 & 0.82 & 1 \end{pmatrix}.$$

And then

$$\begin{aligned} FH(X_1,X_2) &= -\frac{1}{4}(\log\frac{2.93}{4}+\log\frac{3.17}{4}+\log\frac{3.17}{4} \\ &\quad +\log\frac{2.93}{4}) \\ &= 0.39. \end{aligned}$$

In this way,

$$\begin{aligned} FMI(X_1;X_2) &= FH(X_1)+FH(X_2)-FH(X_1,X_2) \\ &= 0.29+0.39-0.39=0.29. \end{aligned}$$

3. Relationship between fuzzy information entropy and differential entropy

Shannon's information entropy of continuous variables can not be directly computed, so some algorithms were proposed to estimate the probability density function with a set of samples[18,29]. Here, we discuss the relationship between fuzzy information entropy and differential entropy in which probability density is estimated with Parzen window.

Given a set of samples $U=\{x_1,x_2,...,x_n\}$, the probability density estimated with Parzen window is

$$\widehat{p}_{Pw}(x)=\frac{1}{n}\sum_{i=1}^{n}\varphi(x-x_i,h), \qquad (20)$$

where $\varphi(\cdot)$ is window function and h is the window width.

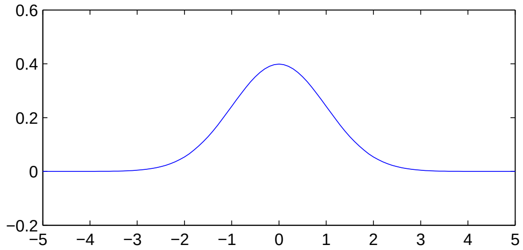

Fig. 1. 1-D Gaussian window

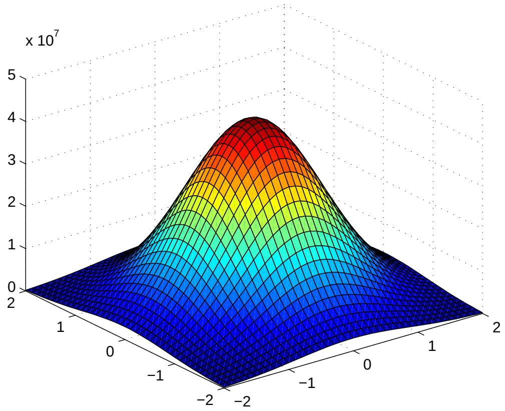

Fig. 2. 2-D Gaussian window

The Gaussian window function is defined as

$$\varphi(z,h)=\frac{1}{(2\pi)^{d/2}h^d|\Sigma|^{1/2}}\exp\left(-\frac{z^T\Sigma^{-1}z}{2h^2}\right), \quad (21)$$

where $z=x-x_i$, Σ is covariance matrices of z. For example, Fig.1 and Fig.2 show one dimensional Gaussian window ($d=1$) and two dimensional Gaussian window ($d=2$), respectively.

Probability density estimated with one dimensional Gaussian window is

$$\begin{aligned} \widehat{p}_{Pw}(x) &= \frac{1}{n}\sum_{i=1}^{n}\varphi(x-x_i,1) \\ &= \frac{1}{n}\sum_{i=1}^{n}\left(\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{(x-x_i)^2}{2}\right)\right). \quad (22) \end{aligned}$$

This is the average value of n Gaussian function values with each sample as center. Next, we use an example to illustrate how to estimate probability density.

Example: Given a set X with five samples $x_1 = 2$, $x_2 = 2.5, x_3 = 3, x_4 = 1$ and $x_5 = 6$, we take the formula (22) with $h = 1$ as window function to estimate probability density of $x = 3$.

$$\begin{aligned}
&p_1 = \tfrac{1}{\sqrt{2\pi}}\exp(-\tfrac{(x_1-x)^2}{2}) = 0.242\\
&p_2 = \tfrac{1}{\sqrt{2\pi}}\exp(-\tfrac{(x_2-x)^2}{2}) = 0.352\\
&p_3 = \tfrac{1}{\sqrt{2\pi}}\exp(-\tfrac{(x_3-x)^2}{2}) = 0.399\\
&p_4 = \tfrac{1}{\sqrt{2\pi}}\exp(-\tfrac{(x_4-x)^2}{2}) = 0.054\\
&p_5 = \tfrac{1}{\sqrt{2\pi}}\exp(-\tfrac{(x_5-x)^2}{2}) = 0.004\\
&p(x=3) = (p_1+p_2+p_3+p_4+p_5)/5 = 0.210
\end{aligned}$$

From (9) we can see

$$\frac{|[x_i]_{R_F}|}{n} = \frac{1}{n}\sum_{j=1}^{n} r_{ij} \tag{23}$$

is identical to the probability density of Shannon's information entropy. If we use one dimensional Gaussian membership function $\psi(\cdot)$ to compute similarity r_{ij} between x_i and x_j, the above formula is

$$\frac{|[x_i]_{R_F}|}{n} = \frac{1}{n}\sum_{j=1}^{n} r_{ij} = \frac{1}{n}\sum_{j=1}^{n} \psi(x_i - x_j), \tag{24}$$

denoted by $\widehat{p}_{R_F}(x)$. $\widehat{p}_{R_F}(x)$ is similar to $\widehat{p}_{Pw}(x)$.

If we compute differential entropy

$$H(F) = -\int\limits_{f\in F} p(f)\log p(f)df \tag{25}$$

with

$$H(F) = -\sum_{f\in F} p(f)\log p(f), \tag{26}$$

the computation cost of Shannon's entropy is the same as that of fuzzy information entropy. But if we use (25) to compute entropy, we should estimate a probability density function. That is to say, Parzen window method not only estimates probability density of given samples, but also uses samples to estimate the probability density of unknown points. While as for fuzzy information entropy we only compute the membership degree of a sample belonging to others. Moreover, in computing fuzzy joint entropy of multiple variables, we use the intersection of fuzzy sets $[x_i]_{F_1} \cap [x_i]_{F_2}$, instead of joint probability density. Obviously, the estimation using Parzen window is more complex than computing membership degree. And in the case of high-dimensional space, it is very difficult to precisely estimate probability density functions with finite samples.

4. Fuzzy mutual information based feature selection algorithms

Feature selection is to select a set of features which has the maximal relevance with decision. The usual way of feature selection is to select a feature singly which has the maximal relevance with decision, which is called Max-Relevance [1]. That is to say the feature is the most important for decision. Let S be a feature subset selected and c be decision. The Max-Relevance is defined as

$$\max D(S,c), D = \frac{1}{|S|}\sum_{f_i\in S} I(f_i;c). \tag{27}$$

However, feature selection according to Max-Relevance may produce redundancies i.e. the new feature selected f_i is strongly relevant to some features selected previously. Therefore, min-Redundancy

$$\min R(S), R = \frac{1}{|S|^2}\sum_{f_i,f_j\in S} I(f_i,f_j) \tag{28}$$

was combined with Max-Relevance [8]. That equals

$$\max \Phi(D,R), \Phi = D - R, \tag{29}$$

called min-Redundancy-Max-Relevance denoted by mRMR.

Given the set S_{k-1} with $k-1$ features selected, the k'th feature can be determined by

$$\max_{f_j\in F-S_{k-1}} [I(f_j;c) - \frac{1}{k-1}\sum_{f_i\in S_{k-1}} I(f_j;f_i)]. \tag{30}$$

Here, we replace mutual information with fuzzy mutual information. The above formula equals

$$\max_{f_j \in F - S_{k-1}} (FMI(f_j;c) - \frac{1}{k-1} \sum_{f_i \in S_{k-1}} FMI(f_j;f_i)). \quad (31)$$

Fuzzy mutual information based mRMR, denoted by FMI_mRMR, is able to directly address continuous features.

The purpose of feature selection is to find a feature subset S_k, which has the maximal relevance to decision c. This is called Max-Dependency (MD) defined as

$$\max D'(S_k,c), D' = I(S_k;c). \quad (32)$$

That means the k'th feature can be determined as the one that makes $I(S_k;c)$ largest, where $I(S_k;c)$ takes the form

$$\begin{aligned} & I(S_k;c) \\ & = \iint p(S_k,c) \log \frac{p(S_k,c)}{p(S_k)p(c)} dS_k dc \\ & = \iint p(S_{k-1},f_k,c) \log \frac{p(S_{k-1},f_k,c)}{p(S_{k-1},f_k,)p(c)} dS_{k-1} df_k dc \\ & = \int \ldots \int p(f_1,\ldots,f_k,c) \log \frac{p(f_1,\ldots,f_k,c)}{p(f_1,\ldots,f_k)p(c)} df_1 \ldots df_k dc. \end{aligned} \quad (33)$$

Similarly, we integrate MD with fuzzy mutual information, denoted by FMI_MD. That equals

$$\begin{aligned} & FMI(S_k;c) = FH(S_k) + FH(c) - FH(S_k,c) \\ & = -\frac{1}{n} \sum_{i=1}^{n} \log \frac{|[x_i]_{f_1} \cap [x_i]_{f_2} \cap \cdots \cap [x_i]_{f_k}| \cdot |[x_i]_c|}{n \cdot |[x_i]_{f_1} \cap [x_i]_{f_2} \cap \cdots \cap [x_i]_{f_k} \cap [x_i]_c|}. \end{aligned} \quad (34)$$

Similarly, when we are selecting features with MD, redundancy might have been produced because the new selected feature may have some relevance to the features that have been selected in advance. In this sense we combine Max-Dependency with min-Redundancy, which is called min-Redundancy-Max-Dependency (mRMD) expressed as

$$\max \Phi(D',R), \Phi = D' - R, \quad (35)$$

which equals

$$\max_{f_j \in F - S_{k-1}} (I(S_{k-1} \cup f_j;c) - \frac{1}{k-1} \sum_{f_i \in S_{k-1}} I(f_j;f_i)). \quad (36)$$

Combined with fuzzy mutual information the above formula equals

$$\max_{f_j \in F - S_{k-1}} (FMI(S_{k-1} \cup f_j;c) - \frac{1}{k-1} \sum_{f_i \in S_{k-1}} FMI(f_j;f_i)). \quad (37)$$

We denote this method FMI_mRMD.

The pseudocode for the three feature selection algorithms, FMI_mRMR, FMI_MD and FMI_mRMD, are as follows.

```
Input: $X,F,c$     $X$ is a sample set,
                   $F$ is a feature set and $c$ is decision.
Output: $S$        $S$ is a feature ranking.
    begin
      initialize $S = \phi$
      while $F \neq \phi$
        find $f \in F$ satisfying (1),(2) or (3)
          $S = S \bigcup \{f\}$
          $F = F - \{f\}$
      end
      return $S$
    end
```

Remarks: (1) $f = \arg\max_{f \in F - S_{k-1}} (FMI(f;c) - \frac{1}{k-1} \sum_{f_i \in S_{k-1}} FMI(f;f_i))$;
(2) $f = \arg\max_{f \in F - S_{k-1}} FMI(S_{k-1} \cup \{f\},c)$;
(3) $f = \arg\max_{f \in F - S_{k-1}} (FMI(S_{k-1} \cup \{f\};c) - \frac{1}{k-1} \sum_{f_i \in S_{k-1}} FMI(f;f_i))$.
If f satisfies (1), this is FMI_mRMR feature selection algorithm; if f satisfies (2), this is FMI_MD feature selection algorithm; and if f satisfies (3), this is FMI_mRMD feature selection algorithm.

Output of each algorithm is a feature ranking. Take FMI_mRMR algorithm as an example.

Step 1: we compute the fuzzy mutual information between each feature and decision, and select the feature f'_1 with the maximum value as the first member of feature ranking S. Then $S = \{f'_1\}$, and $F = F - \{f'_1\}$.

Step 2: $\forall f \in F$, we compute $FMI(f;c) - \frac{1}{|S|} \sum_{f'_k \in S} FMI(f;f'_k) = FMI(f;c) - FMI(f;f'_1)$. By this way, we get $|F|$ values. And the feature f'_2 with the maximum value is selected. Then $S = \{f'_1, f'_2\}$, and $F = F - \{f'_2\}$.

Step 3: $\forall f \in F$, we compute $FMI(f;c) - \frac{1}{|S|} \sum_{f'_k \in S} FMI(f;f'_k) = FMI(f;c) - \frac{1}{2}(FMI(f;f'_1) + FMI(f;f'_2))$. By this way, we get $|F|$ values, and the feature f'_3 with the maximum value is selected. Then $S = \{f'_1, f'_2, f'_3\}$, and $F = F - \{f'_3\}$.

Step 4: repeat Step 3 until $F = \emptyset$.

By this way, we can get a feature ranking with FMI_mRMR algorithm. The computational complexity of this incremental search method is $O(|S| \cdot$

$|F|$, where $|S|$ is the number of features selected, and $|F|$ is the number of features being not selected. Similarly, we can get two feature rankings with FMI_mRMD and FMI_MD algorithms. The computational complexity of the incremental search methods for FMI_mRMD and FMI_MD are $O(|S| \cdot |F|)$ and $O(|F|)$, respectively.

5. Evaluation measures of stability

We can evaluate the performance of feature selection algorithms with the size and classification performance of selected features [22,32,33]. Moreover, stability is also an aspect for evaluating feature selection algorithms [15]. This section we give evaluation measures of stability for the algorithms.

In this work, we evaluate stability of feature selection algorithms with the similarity of feature rankings and that of feature subsets. A technique like cross-validation is introduced. We divide the samples into k subsets and use k-1 subsets to rank features using feature selection algorithms. We get k feature rankings after k rounds using a certain algorithm. Accordingly, we get k feature subsets. The larger the similarity of k feature rankings or feature subsets is, the more stable the algorithm is.

To measure the similarity between two feature rankings $R = \{r_1, r_2, ..., r_N\}$ and $R' = \{r'_1, r'_2, ..., r'_N\}$, we use Spearman's rank correlation coefficient [25]

$$S_R(R,R') = 1 - 6\sum_{i=1}^{N} \frac{(r_i - r'_i)^2}{N(N^2-1)}. \qquad (38)$$

Here, $S_R \in [-1,1]$. $S_R = 1$ means that the two rankings are identical; $S_R = 0$ means that there is no correlation between the two ranks; $S_R = -1$ means that they have exactly inverse orders.

We measure the similarity between two feature subsets F_1 and F_2 with Tanimoto distance [9]

$$\begin{aligned} S_F(F_1,F_2) &= 1 - \frac{|F_1| + |F_2| - 2|F_1 \cap F_2|}{|F_1| + |F_2| - |F_1 \cap F_2|} \\ &= \frac{|F_1 \cap F_2|}{|F_1| + |F_2| - |F_1 \cap F_2|}. \end{aligned} \qquad (39)$$

After calculating the similarity of feature rankings and subsets, we can get a similarity matrix

$$S = \begin{bmatrix} s_{11} & s_{12} & \cdots & s_{1k} \\ s_{21} & s_{22} & \cdots & s_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ s_{k1} & s_{k2} & \cdots & s_{kk} \end{bmatrix}. \qquad (40)$$

In order to measure the similarity of all the rankings or subsets, the Kalousis et al. [15] used

$$TS_1 = \sum_{i=1}^{k}\sum_{j=1}^{k} s_{ij} \qquad (41)$$

to measure the similarity matrix. Wang et al. [43,44] introduced another way

$$TS_2 = -\sum_{i=1}^{k} \frac{\lambda_i}{k} \log_k \frac{\lambda_i}{k} \qquad (42)$$

to measure similarity matrix, where $TS_2 \in [0,1]$. λ_i $(i = 1,2,...,k)$ are eigenvalues of similarity matrix. As the similarity matrix is real symmetry, $0 \leqslant \lambda_i \leqslant k$ $(i = 1,2,...,k)$. If k results are the same, we get $\lambda_1 = k$, $\lambda_i = 0$ $(i > 1)$, $TS_2 = 0$. Then we consider the feature selection algorithm is the most stable. When $\forall\, i \neq j$, $s_{ij} = 0$, S is identity matrix and $\lambda_i = 1$ $(i = 1,2,...,k)$, $TS_2 = 1$. Then we consider feature selection algorithm is the least stable. The smaller TS_2 is, the stronger the stability is. So we can use TS_2 to measure the stability of feature selection algorithms.

Moreover, Hu et al. [14] used another entropy

$$TS_3 = -\frac{1}{k}\sum_{j=1}^{k} \log \sum_{i=1}^{k} \frac{s_{ij}}{k} \qquad (43)$$

to measure the similarity matrix, where $TS_3 \in [0, \log k]$. If $\forall i,j$, $s_{ij} = 1$, which means the k results are the same, $TS_3 = 0$. In this case, the feature selection algorithm is the most stable. If $\forall i \neq j$, $s_{ij} = 0$, S is an identity matrix, then $TS_3 = \log k$. We consider the feature selection algorithm is the least stable. The smaller TS_3 is, the stronger the stability is. In this work we use TS_3 to measure the stability of feature selection algorithms.

6. Experiments

In this section, FMI_mRMR, FMI_MD and FMI_mRMD are tested on 14 benchmark tasks from UCI[3]. The summary of data sets is shown in Table 1, where ”Size” is the number of samples, ”Feature” is the number of all the features, ”N” stands for the number of numerical features, ”C” for the number of nominal features and ”Class” for the number of the classes.

Table 1: Summary of data sets

Data	Size	Feature	N	C	Class
heart	270	13	7	6	2
hepatitis	155	19	6	13	2
horse	368	22	7	15	2
iono	351	34	34	0	2
sonar	208	60	60	0	2
WDBC	569	30	31	0	2
wine	178	13	13	0	3
zoo	101	16	0	16	7
segmentation	2310	19	3	16	7
yeast	1484	7	6	1	10
breast	84	9216	9216	0	5
DLBCL	88	4026	4026	0	6
lung	96	7129	7129	0	3
SRBCT	88	2308	2308	0	5

First, we rank features with FMI_mRMR, FMI_MD and FMI_mRMD algorithms, respectively. Feature ranking leads to n sequential feature subsets $S_1 \subset S_2 \subset \cdots \subset S_{n-1} \subset S_n$, where n is the number of features, $S_1 = \{f_1\}$, $S_2 = \{f_1, f_2\}$,..., $S_n = \{f_1, f_2, ..., f_n\}$. Then we use 10-fold cross-validation to calculate the classification accuracies of $S_1, S_2, ..., S_n$ with LSVM[4], RBFSVM[4], CART[21] and KNN[27] classifiers, respectively. We select the subset S_i $(i = 1, 2, ..., n)$ with the highest classification accuracy as the final feature subset.

6.1. Accuracy comparison

We first test the effectiveness of feature selection algorithms. We take linear SVM (LSVM) as classifiers to illustrate the effectiveness of feature selection algorithms. The results are shown in Table 2, where ’All features’ means classification accuracies obtained without feature selection, ’n’ is features selected, and ’Acc’ is classification accuracy using LSVM with selected features.

Table 2 shows LSVM produces a good performance for classification without feature selection. From the results we can see that features selected by FMI_mRMR algorithm can produce higher classification accuracy than that produced with all features, which can adequately show efficiency of feature selection algorithm. It also shows, taking LSVM as classifier, features selected by FMI_mRMR are better than that selected by FMI_MD and FMI_mRMD.

Now, we conduct experiments to test FMI_mRMR, FMI_MD and FMI_mRMD, and compare their performance with some state-of-the-art techniques, such as MI_mRMR[29], CFS[11], FCBF[47] and RELIEF[17] algorithms.

MI_mRMR is min-Redundancy-Max-Relevancy based on Shannon’s mutual information, where continuous data should be discretized in preprocessing. CFS, ”correlation based feature selection”, is a simple filter algorithm that selects feature subset in terms of a correlation-based heuristic evaluation function. FCBF, ”Fast Correlation-Based Filter”, is a fast filter method which can identify relevant features as well as redundant ones among relevant features without pairwise correlation analysis. RELIEF is considered as one of the most successful technique due to its simplicity and effectiveness. It is to iteratively estimate weights of features according to their ability to discriminate neighboring patterns.

As different classifiers may produce different accuracies with the same feature subset, we use LSVM, RBFSVM, CART and KNN to classify data sets in this work. The classification accuracy comparison of data sets described by feature subsets selected using the four classifiers are shown in Fig.3. Twelve sub figures are results for twelve data sets. In each sub figure, there are four groups of bars, denoting four classifiers i.e. LSVM, RBFSVM, CART and KNN. Each bar presents the classification accuracy of a data set described by features selected with a classifier.

Table 2: Classification accuracy (%) with LSVM for different algorithms

Data	All features		FMI_mRMR		FMI_MD		FMI_mRMD	
	Acc	*n*	Acc	*n*	Acc	*n*	Acc	*n*
heart	83.3	13	84.1	7	82.6	7	82.6	7
hepatitis	86.2	19	89.2	3	91.0	10	91.0	10
horse	92.4	22	91.6	3	90.2	2	90.2	2
iono	87.6	34	89.0	2	88.7	22	88.7	22
sonar	77.9	60	78.9	11	77.5	15	77.5	15
WDBC	97.7	30	98.1	20	97.2	12	97.2	12
wine	98.9	13	98.9	5	98.3	7	98.3	7
zoo	93.4	16	95.4	12	93.4	6	93.4	6
segmentation	92.9	19	92.1	10	91.5	15	92.0	15
yeast	56.4	7	56.4	7	55.6	7	56.8	7
breast	95.4	9216	100.0	20	82.5	16	76.2	17
DLBCL	97.3	4026	97.3	9	80.2	15	67.1	5
lung	82.6	7129	80.3	6	77.4	11	75.0	2
SRBCT	82.1	2308	81.3	10	61.7	20	33.4	5
Average	87.5	1636	88.0	8	83.3	11	80.0	11

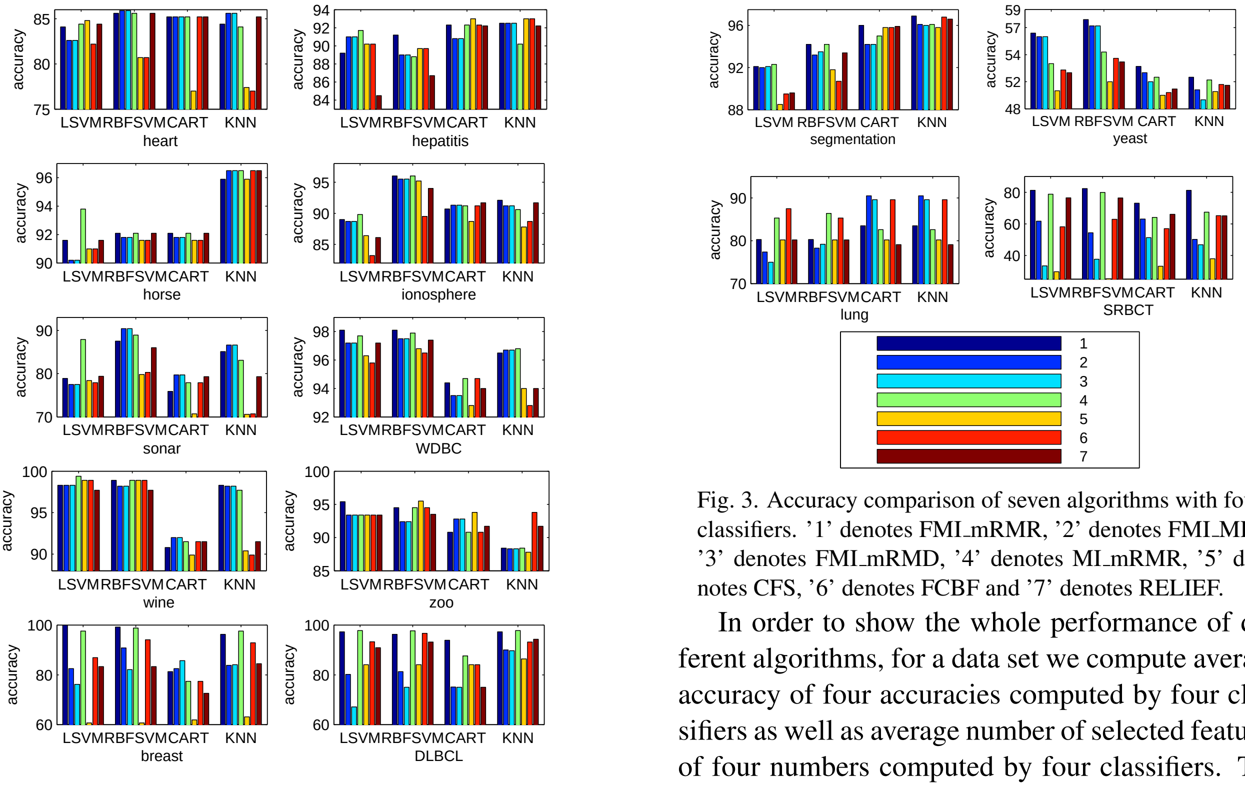

Fig. 3. Accuracy comparison of seven algorithms with four classifiers. '1' denotes FMI_mRMR, '2' denotes FMI_MD, '3' denotes FMI_mRMD, '4' denotes MI_mRMR, '5' denotes CFS, '6' denotes FCBF and '7' denotes RELIEF.

In order to show the whole performance of different algorithms, for a data set we compute average accuracy of four accuracies computed by four classifiers as well as average number of selected features of four numbers computed by four classifiers. The results are shown in Table 3. "*n*" is the average number of features selected, and "Acc" is average classification accuracy.

From the whole average ”TotalAverage”, we can get the following conclusions. Features selected by FMI_mRMR can produce the highest accuracy of all. Numbers of features selected by FMI_mRMR, FMI_MD and FMI_mRMD are less than or equal that by CFS, and the accuracies produced with the three algorithms are higher than that with CFS. The whole average ”TotalAverage” of accuracies presents that FMI_mRMR is the best of all the algorithms. The total average accuracy got by FMI_MD is close to that by FCBF and RELIEF, and the number of features selected by FMI_MD is less than or equal that by FCBF and RELIEF. Although the accuracy produced by FMI_mRMD is lower than that by FCBF and RELIEF, the number of features selected with FMI_mRMD is less than that with FCBF and RELIEF.

6.2. Stability analysis of feature selection algorithms

Stability is another view point to evaluate an algorithm for feature selection. In this section we discuss the stability of FMI_mRMR, FMI_MD and FMI_mRMD algorithms. In order to measure stability we use measures introduced in Section 5. We use TS_3 to calculate the stability of feature rankings and subsets. Here, $k = 3$. Furthermore, we compare the stability of the above three algorithms with MI_mRMR, CFS, FCBF and RELIEF. The results are shown in Tables 4 and 5.

Remarks: breast, DLBCL, lung and SRBCT are small data sets. If we use the method mentioned in Section 5 to evaluate stability, data sets used to select features are much smaller. This may make feature rankings and feature subsets selected inaccurate. So we do not consider these data sets in this work.

Table 4 shows evaluation results for the stability of feature rankings for different feature selection algorithms. It shows that FMI_mRMR has the smallest stability evaluation value. Section 5 analyzes that the smaller the value is, the more stable the algorithm is. So FMI_mRMR is the most stable of all the algorithms. FMI_MD and FMI_mRMD are less stable than FMI_mRMR. Besides, we can see FMI_mRMR, FMI_MD and FMI_mRMD are more stable than other algorithms

Table 5 shows evaluation results for the stability of feature subsets. It shows that FMI_mRMR algorithm has the smallest stability evaluation value, which means FMI_mRMR algorithm is still the most stable. The evaluation values for FMI_MD and FMI_mRMD are smaller than MI_mRMR, CFS, FCBF and RELIEF. FMI_MD is more stable than FMI_mRMD for selecting features.

7. Conclusions

Mutual information is widely used to measure relevance between discrete features and decision. It plays an important role in feature selection algorithms. Considering the limitation of Shannon’s mutual information, we introduce fuzzy information entropy and fuzzy mutual information to calculate relevance between continuous or fuzzy features and decision. Furthermore, we combine this measure with mRMR, MD and mRMD algorithms to construct feature selection algorithms. We test the algorithms on UCI data sets in terms of classification performance and stability. The following conclusions are drawn from the analysis.

Firstly, fuzzy mutual information is a feasible and effective measure for computing relevance between numerical features and decision. Fuzzy mutual information computes the relevance of high-dimensional features using intersection of fuzzy relation induced with single features, instead of estimating probability density, so the computational complexity decreases.

Secondly, the feature selection algorithm by combining fuzzy mutual information with mRMR search strategy is effective. The proposed algorithm is comparable with the classical mRMR, fuzzy mutual information based MD, fuzzy mutual information based mRMD, CFS, FCBF and RELIEF algorithms.

Finally, the experiments on stability show that fuzzy mutual information based mRMR, MD and

Table 3: Average classification accuracy (%) with four classifiers for different algorithms

Data	FMI_mRMR		FMI_MD		FMI_mRMD		MI_mRMR		CFS		FCBF		RELIEF	
	n	Acc	*n*	Acc	*n*	Acc	*n*	Acc	*n*	Acc	*n*	Acc	*n*	Acc
heart	5	84.8	5	84.8	5	84.8	6	84.8	7	80.0	6	81.3	3	85.1
hepatitis	3	91.3	5	90.8	5	90.8	5	90.7	7	91.4	7	91.3	8	88.9
horse	5	93.3	4	92.2	4	92.2	3	94.7	8	93.6	7	93.7	7	94.2
iono	10	92.1	11	91.7	11	91.7	13	91.7	11	89.5	7	88.2	8	90.8
sonar	7	96.6	6	96.7	6	96.7	7	96.9	11	94.5	9	94.8	8	94.6
WDBC	17	96.8	8	96.2	8	96.2	10	96.8	10	94.9	8	94.9	8	95.6
wine	23	81.9	16	83.6	16	83.6	23	84.5	17	43.3	12	76.8	17	81.0
zoo	6	92.3	5	91.7	5	91.7	4	91.8	8	92.6	6	93.1	10	92.6
segmentation	10	94.8	14	93.9	12	94.0	11	94.4	7	93.0	6	93.2	13	93.9
yeast	6	54.6	7	53.8	7	53.3	7	52.5	7	50.1	6	51.6	6	51.5
breast	17	94.2	12	84.9	6	82.1	7	92.8	7	61.6	11	87.8	13	80.9
DLBCL	7	96.2	7	81.7	4	76.7	10	95.2	17	84.6	13	91.8	18	88.3
lung	5	80.4	10	83.8	3	81.0	6	83.1	7	79.2	12	88.2	4	78.9
SRBCT	10	80.0	7	57.4	4	42.3	11	72.5	7	31.6	7	60.8	7	71.0
TotalAverage	11	87.8	10	84.6	8	82.7	11	87.3	11	77.2	10	84.8	11	84.8

Table 4: Stability of feature rankings

Data	FMI_mRMR	FMI_MD	FMI_mRMD	MI_mRMR	CFS	FCBF	RELIEF
heart	0.76	0.91	1.07	0.84	0.86	0.86	1.56
hepatitis	0.66	0.37	0.67	0.66	0.80	0.80	0.83
horse	0.50	0.58	0.66	0.67	1.00	1.00	1.33
ionosphere	0.56	0.60	0.61	0.97	0.90	0.90	0.66
sonar	0.60	0.60	0.64	0.95	1.11	1.11	0.74
WDBC	0.94	1.17	1.23	1.00	1.11	1.11	0.88
wine	0.67	0.72	0.66	0.78	0.72	0.72	1.04
zoo	0.58	0.58	0.56	0.69	0.80	0.80	0.66
segmentation	0.63	0.65	0.63	0.65	0.70	0.69	0.81
yeast	0.56	0.62	0.61	0.60	0.65	0.71	0.66
Average	0.65	0.68	0.73	0.78	0.86	0.87	0.92

Table 5: Stability of feature subsets

Data	FMI_mRMR	FMI_MD	FMI_mRMD	MI_mRMR	CFS	FCBF	RELIEF
heart	0.54	0.55	0.54	0.70	0.56	0.60	0.41
hepatitis	0.44	0.46	0.51	0.63	0.59	0.62	0.49
horse	0.41	0.47	0.50	0.63	0.62	0.67	0.62
ionosphere	0.49	0.49	0.55	0.76	0.84	0.71	0.79
sonar	0.49	0.79	0.80	1.02	0.86	0.80	0.73
WDBC	0.42	0.53	0.53	0.56	0.52	0.60	0.47
wine	0.25	0.33	0.43	0.56	0.44	0.50	0.30
zoo	0.21	0.29	0.31	0.21	0.35	0.49	0.26
segmentation	0.38	0.36	0.42	0.39	0.54	0.49	0.67
yeast	0.32	0.29	0.37	0.34	0.71	0.55	0.63
Average	0.39	0.46	0.50	0.58	0.60	0.60	0.54

mRMD feature selection algorithms are more stable than some state-of-the-art algorithms.

Acknowledgments

This work is supported by National Natural Science Foundation of China under Grants 60703013 and 10978011 and National Science Fund for Distinguished Young Scholars under Grant 50925625.

References

1. R. Battiti, ”Using mutual information for selecting features in supervised neural net learning,” IEEE Transactions on Neural Networks, 5, 531-549(1994).
2. J.Y. Ching, A.K.C. Wong and K.C.C. Chan, ”Class-dependent discretization for inductive learning form continuous and mixed-mode data,” IEEE Transactions on Pattern Analysis and Machine Intelligence, 17, 641-651(1995).
3. C.L. Blake and C.J. Merz, ”UCI Repository of Machine Learning Databases,” Available: http://www.ics.uci.edu/mlearn/MLRepository.html, 1998.
4. C. Corts and V. Vapnik, ”Support vector networks,” Machine Learning, 20, 1-25(1995).
5. T.M. Cover, ”The best two independent measurements are not the two best,” IEEE Transactions on Systems, Man, and Cybernetics, 4, 116-117(1974).
6. M. Dash and H. Liu, ”Consistency-based search in feature selection,” Artificial Intelligence, 151, 155-176(2003).
7. A.B. David and H. Wang, ”A formalism for relevance and its application in feature subset selection,” Machine Learning, 41, 175-195(2000).
8. C. Ding and H.C. Peng, ”Minimum redundancy feature selection from microarray gene expression data,” In Proceeding of the 2003 IEEE Computational Systems Bioinformatics Conference, Stanford, California, 523-528(2003).
9. R. Duda, P. Hart and D. Stork, ”Pattern classification and scene analysis,” Wiley, New York, 2001.
10. M.E. Farmer, S. Bapna and A.K. Jain, ”Large scale feature selection using modified random mutation hill climbing,” In Proceedings of the 17th International Conference on Pattern Recognition, UK, 2, 287-290(2004).
11. M.A. Hall, ”Correlation-based feature selection for discrete and numeric class machine learning,” In Proceedings of the Seventeenth International Conference on Machine Learning, Hamilton, New Zealand, 359-366(2000).
12. Q.H. Hu, Z.X. Xie and D.R. Yu, ”Hybrid attribute reduction based on a novel fuzzy-rough model and information granulation,” Pattern Recognition, 40, 3509-3521(2007).
13. Q.H. Hu, D.R. Yu and Z.X. Xie, ”Information-preserving hybrid data reduction based on fuzzy-rough techniques,” Pattern Recognition Letters, 27, 414-423(2006).
14. Q.H. Hu, J.F. Liu and D.R. Yu, ”Stability Analysis on Rough Set Based Feature Evaluation,” Lecture Notes in Computer Science, Springer Berlin/Heidelberg, 5009, 88-96(2008).
15. A. Kalousis, J. Prados and M. Hilario, ”Stability of feature selection algorithms: a study on high-dimensional spaces,” Knowledge and Information Systems, 12, 95-116(2007).
16. L.J. Ke, Z.R. Feng and Z.G. Ren, ”An efficient ant

colony optimization approach to attribute reduction in rough set theory," Pattern Recognition Letters, 29, 1351-1357(2008).

17. I. Kononenko, "Estimating Attributes: Analysis and Extensions of RELIEF," European Conference on Machine Learning, 171-182(1994).
18. N. Kwak and C.-H. Choi, "Input Feature selection by mutual information based on Parzen window," IEEE Transactions on Pattern Analysis and Machine Intelligence, 24, 1667-1671(1994).
19. N. Kwak and C.-H. Choi, "Input feature selection for classification problems," IEEE Transaction on Neural Networks, 13, 143-159(2002).
20. H.-M. Lee, C.-M Chen, J.-M. Chen and Y.-L. Jou, "An efficient fuzzy classifier with feature selection based on fuzzy information entropy," IEEE Transactions on Systems, Man, and Cybernetics-Part B: Cybernetics, 31, 426-432(1999).
21. B. Leo, H.F. Jerome, A.O. Richard and J.S. Charls, "Classification and regression trees," Chapman and Hall, New York, 1993.
22. Y.H. Li, M. Dong and J. Hua, "Localized feature selection for clustering," Pattern Recognition Letters, 29, 10-18(2008).
23. H. Liu and H. Motoda, "Feature selection for knowledge discovery and data mining," Kluwer Academic Publishers, Boston, 1998.
24. X.X. Liu, A. Krishnan and A. Mondry, "An entropy-based gene selection method for cancer classification using microarray data," BMC Bioinformatics, 6, 1-14(2005).
25. J.S. Maritz, "Distribution-free statistical methods," Chapman Hall, 217, 1981.
26. P. Narendra and K. Fukunaga, "A branch and bound algorithm for feature subset selection," IEEE Transactions on Computers, 26, 917-922(1997).
27. E.A. Patrick and F.P. Fisher, "A generalized *k*-nearest neighbor rule," Information and Control, 16, 128-152(1970).
28. Z. Pawlak and C. Rauszer, "Dependency of attributes in information systems," Bull. Polish Acad. Sci. Math. 33, 551-559(1985).
29. H.C. Peng, F.H. Long and C. Ding, "Feature selection based on mutual information: criteria of Max-Dependency, Max-Relevance, and Min-Redundancy," IEEE Transactions on Pattern Analysis and Machine Intelligence, 27, 1226-1238(2005).
30. M. Prasad, "Online feature selection for classifying emphysema in HRCT images," International Journal of Computational Intelligence Systems, 1, 127-133(2008).
31. P. Pudil, J. Novovicova and J. Kittler, "Floating search methods in feature selection," Pattern Recognition Letters, 15, 1119-1125(1994).
32. Y.H. Qian, J.Y. Liang and C.G. Dang, "Consistency measure, inclusion degree and fuzzy measure in decision tables," Fuzzy Sets and Systems, 159, 2353-2377(2008).
33. Y.H. Qian, J.Y. Liang, C.G. Dang, H.Y. Zhang and J.M. Ma, "On the evaluation of the decision performance of an incomplete decision table," Data and Knowledge Engineering, 65, 373-400(2008).
34. J.C. Schlimmer, "Efficiently inducing determinations: a complete and systematic search algorithm that uses optimal pruning," Proceedings of Tenth International Conference on Machine Learning, Morgan Kaufmann, MA, 284-290(1993).
35. C.E. Shannon, "A mathematical theory of communication," The Bell System Technical Journal, 27, 379-423(1948).
36. Q. Shen and R. Jensen, "Selecting informative features with fuzzy-rough sets and its application for complex systems monitoring," Pattern Recognition, 37, 1351-1363(2004).
37. C. Sima and E.R. Dougherty, "The peaking phenomenon in the presence of feature-selection," Pattern Recognition Letters, 29, 1667-1674(2008).
38. P. Somol, P. Pudil, J. Novoviova and P. Paclik, "Adaptive floating search methods in feature selection," Pattern Recognition Letters, 20, 1157-1163(1999).
39. P. Somol, P. Pudil and J. Kittler, "Fast branch and bound algorithms for optimal feature selection," IEEE Transactions on Pattern Analysis and Machine Intelligence, 26, 900-912(2004).
40. M.R. Suarez, J.R Vilar and J. Grande, "A feature selection method using a fuzzy mutual information measure," Advances in Soft Computing, Springer Berlin / Heidelberg, 44, 56-63(2007).
41. R.W. Swiniarski and A. Skowron, "Rough set methods in feature selection and recognition," Pattern Recognition Letters, 24, 833-849(2003).
42. W.Y. Tang and K.Z. Mao, "Feature selection algorithm for mixed data with both nominal and continuous features," Pattern Recognition Letters, 28, 563-571(2007).
43. Q. Wang, Y. Shen and Y. Zhang, "A quantitative method for evaluating the performances of hyperspectral image fusion," IEEE transactions on Instrumentation and Measurement, 52, 1041-1047(2003).
44. Q. Wang, Y. Shen and J.Q. Zhang, "Nonlinear correlation measure for multivariable data set," PhysicaD-Nonlinear Phenomena, 200, 287-295(2005).
45. Z.H. Wei, D.Q. Miao, J.-H. Chauchat and R.Z. Wen, "LiN-grams based feature selection and text representation for Chinese Text Classification," International Journal of Computational Intelligence Systems, 2, 365-374(2009).
46. W.Z. Wu, "Attribute reduction based on evidence theory in incomplete decision systems," Information Sciences, 178, 1355-1371(2008).

47. L. Yu and H. Liu, ”Efficient feature selection via analysis of relevance and redundancy,” Journal of Machine Learning Research, 5, 1205-1224(2004).