Feature selection based on neighborhood discrimination index

Changzhong Wang, Xizhao Wang, Degang Chen, Qinghua Hu, Yuhua Qian

***Abstract*—Feature selection is viewed as an important preprocessing step for pattern recognition, machine learning and data mining. Neighborhood is one of the most important concepts in classification learning and can be used to distinguish samples with different decisions. In this paper, a neighborhood discrimination index is proposed to characterize the distinguishment information of a neighborhood relation. It reflects the distinguishment ability of a feature subset. The proposed discrimination index is computed by considering the cardinality of a neighborhood relation rather than neighborhood similarity classes. Some variants of the discrimination index, that is, joint discrimination index, conditional discrimination index, mutual discrimination index, are introduced to compute the change of distinguishment information caused by combination of multiple feature subsets. They have the similar properties as Shannon entropy and its variants. A parameter, named neighborhood radius, is introduced in these discrimination measures to make them suitable for analysis of real-valued data. Based on the proposed discrimination measures, the significance measure of a candidate feature is defined and a greedy forward algorithm for feature selection is designed. The data sets selected from public data sources are used to compare the proposed algorithm with some existing algorithms, and the experimental results show that the discrimination index based algorithm yields better performance than some classical ones.**

***Index Terms* — Neighborhood relation, Discrimination index, Feature selection, Distinguishment information**

I. Introduction

With development of computer and database technology, the amount of data is greatly growing. Ideally, the information provided is useful, but in fact, the data often contains redundant information. Therefore, before using a data set, it is necessary to preprocess the data for removing redundant features. Feature selection is an important tool to reduce redundant features. Most researchers are dedicated to processing high dimensional data with feature selection. Its aim is to find a subset of optimal features with strong classification ability according to a evaluation criteria, and obtain high-dimensional characteristics by analyzing low dimensional data. Feature selection is an effective technique to simplify data analysis and acquire key features of data. Recently, it has attracted much attention in pattern recognition, machine learning and data mining [5]-[12], [22]-[26], [30]-[33], [41]-[48], [52].

Relations, produced by a subset of features, represent the similarity or dissimilarity between samples. Similar samples form a similarity class, dissimilar ones fall into different classes. A relation can be used to reflect the ability of features to distinguish samples. Relations have been applied to discretize real-valued data [33], [44], fuzzy clustering, attribute reduction [4], [6], [18], [19], [50], uncertainty reasoning and decision [27], [40], [59]. Furthermore, equivalence relations [29], [34], [38], [54], similarity relations [20], [39], [56]-[58], neighborhood relations [15], [35], [47], [49], [53], and dominance relation [9], [17], [51] are the foundations of a sequence of rough set models.

Entropy, as an uncertainty measure, is a very useful tool for characterizing the distinguishment information of a subset of features. The less likely a decision attribute has conditional entropy with respect to a feature subset, the more the feature subset has capability in distinguishing samples with different decisions. Entropy has played an important role in pattern recognition and feature selection. Since Shannon proposed information entropy to evaluate the uncertainty of discrete sample spaces, entropy has been applied in diverse fields [2], [6], [7], [14], [43]. The extension of entropy and its variants were adapted for feature selections in [1], [13], [21], [36]. In order to calculate the distinguishment information of fuzzy or numerical features, Yager introduced the concept of entropy into fuzzy similarity relations [55]. In fact, Yager's entropy is a generalization of Shannon entropy; It is defined by using equivalence classes or fuzzy similarity classes. In 2002, Hernandez and Recasens extended Yager's work and presented the formulae of joint entropy and conditional entropy based on Yager's entropy, and then they used these measures to learn fuzzy decision trees from a set of data samples [16]. Hu and Yu redefined the joint entropy and conditional entropy based on Yager's work, then used them to measure the uncertainty of distinguishment ability of a set of fuzzy similarity relations [14].

This work was supported by the National Natural Science Foundation of China under Grants 61572082, 61473111, 61303131, 61363056, 61173181and 61070242, the Program for Liaoning Excellent Talents in University under Grant (LR2012039), the natural science foundation of Liaoning Province (2014020142).

C.Z. Wang and Z. Dong are with the Department of Mathematics, Bohai University, Jinzhou 121000, China (e-mail: changzhongwang@ 126.com).

X. Z. Wang is with College of Computer Science and Software Engineering, Shenzhen University, Shenzhen 518060, China. E-mail: xizhaowang@ieee.org.

D. G. Chen is with the Department of Mathematics & Physics, North China Electric Power University, Beijing 102206 (e-mail: chengdegang@263.net).

Q. H. Hu is with School of Computer Science and Technology, Tianjin University, Tianjin 300072, China (e-mail: huqinghua@hit.edu.cn).

Y. H. Qian is with School of Computer and Information Technology, Shanxi University, Taiyuan 030006, P.R. China (e-mail: jinchengqyh@126.com).

In 2005, Mi et al. introduced a distinguishable measure of fuzzy equivalence relation based on fuzzy-rough set model [28]. In 2008, Qian and Liang proposed a combinational measure for evaluating the uncertainty of distinguishment ability of a subset of features [37]. In 2011, Hu and Zhang introduced the concepts of neighborhood entropy, neighborhood conditional entropy and neighborhood mutual information in numerical spaces for evaluating the relevance between continuous features and discrete decision attributes [13]. All the studies are focused on the extensions of Shannon entropy or Yager's entropy and their applications.

As we know, neighborhood is one of the most important concepts in classification learning [15], [23], [47], [49], [60]. Neighborhood can be used to generate similarity classes from the samples described by numerical features, and used to distinguish samples. The distinguishment information of a feature subset is related with the neighborhood relation induced by the feature subset. In this paper, we propose a new measure of distinguishment information based on neighborhood relations, which is called neighborhood discrimination index. Compared with Yager's entropy [55] and its varieties [13], [14], neighborhood discrimination index has the similar properties as Shannon entropy. But it is directly defined on neighborhood relation and acquired by computing the cardinality of neighborhood relation rather than neighborhood similarity classes. Thus the computational complexity of the proposed discrimination index is smaller. We define joint discrimination index, conditional discrimination index and mutual discrimination index and discuss their some basic properties. These measures are used to calculate the change of distinguishment formation caused by combination of multiple feature subsets. Just like Shannon conditional entropy, conditional discrimination index can be used to characterize the ability of a subset of features to distinguish samples with different decisions. The less the conditional discrimination index is, the more the feature subset has distinguishment ability. We also discuss the influence of neighborhood radius on neighborhood discrimination index. Then we define attribute importance and propose an algorithm of feature selection based on the proposed discrimination measures. Finally, we use some public standard data sets to verify the validity and stability of the proposed method, and compare the proposed algorithm with some existing methods. The experimental results show the proposed measures are efficient and effective for feature selections.

This paper is organized as follows. In Section Ⅱ, we recall some preliminaries on Shannon entropy in learning. In Section Ⅲ, we present the definitions of neighborhood discrimination index and its related discrimination measures, and discuss their properties. In Section Ⅳ, we define the significance of a candidate feature and design a heuristic algorithm of feature selection based on mutual discrimination index. In Section Ⅴ, we verify the feasibility and stability of the proposed algorithm. Section Ⅵ concludes the paper.

II. SHANNON ENTROPY IN LEANING

Suppose that U is a nonempty set of samples, A is a set of discrete attributes describing samples, and D is a decision attribute that partitions the sample space into r classes. Let $B \subseteq A$, then an equivalence relation R_B can be induced by attribute subset B as follows.

$$R_B = \{(x_i, x_j) \in U \times U \mid a(x_i) = a(x_j), \forall a \in B\}. \quad (1)$$

Suppose that the partition produced by R_B is denoted by $U/B = \{X_1, X_2, \cdots X_m\}$, where $a(x)$ is the attribute value of sample x on a. The elements in X_i are not distinguished by the attribute subset B as their feature values are the same. If we consider B is a random variable on U and the value space for B is $\{X_1, X_2, \cdots X_m\}$, then the probability distribution of B is described as follows:

$$B \sim \begin{bmatrix} X_1 & X_2 & \cdots & X_m \\ p(X_1) & p(X_2) & \cdots & p(X_m) \end{bmatrix}, \quad (2)$$

where $p(X_i) = |X_i|/|U|$ and $|X_i|$ is the cardinality of X_i, $i = 1, 2, \cdots, m$.

The Shannon entropy of attribute subset B is defined as follows:

$$H(B) = \sum_{i=1}^{m} -p(X_i) \log p(X_i). \quad (3)$$

Let C be another attribute subset of A and the partition induced by C be denoted by $U/C = \{Y_1, Y_2, \cdots, Y_n\}$, then the joint entropy of B and C is defined as:

$$H(B \cup C) = -\sum_{i=1}^{m}\sum_{j=1}^{n} p(X_i \cap Y_j) \log p(X_i \cap Y_j), \quad (4)$$

and the conditional entropy of B on C is computed by:

$$H(B|C) = -\sum_{i=1}^{m}\sum_{j=1}^{n} p(X_i \cap Y_j) \log p(X_i \mid Y_j), \quad (5)$$

where $p(X_i \mid Y_j) = |X_j \cap Y_i|/|Y_i|$.

$H(B|C)$ describes the uncertainty of B in the case that C is given. Obviously, $H(B|C) \geq 0$. If there exists $X_i \in U/B$ such that $p(X_i \mid Y_j) = 1$ for any $Y_j \in U/C$, then $H(B|C) = 0$. This means that the distinguishment ability of attribute subset B is completely contained in C in this case.

The mutual information of B and C is defined as:

$$I(B;C) = \sum_{i=1}^{m}\sum_{j=1}^{n} p(X_i \cap Y_j) \log \frac{p(X_i \cap Y_j)}{p(X_i)p(Y_j)}. \quad (6)$$

Mutual information describes the statistical correlation between B and C. It is easily proved that $I(B;C) \geq 0$. When B and C are independent, then $I(B;C) = 0$. In this case, B and C do not provide any forecast information. In addition,

we easily know that mutual information has the following properties.

1. $I(B;C)=I(C;B)$,
2. $I(B;C)=H(B)+H(C)-H(B\bigcup C)$,
3. $I(B;C)=H(B)-H(B\,|\,C)=H(C)-H(C\,|\,B)$. (7)

We consider decision attribute D as a random variable on U and suppose the value space for D is $\{\omega_1,\omega_2,\cdots,\omega_r\}$, where ω_i denotes the ith decision class. Then the conditional entropy of decision D on attribute subset B can be computed by:

$$H(D|B)=-\sum_{i=1}^{m}\sum_{j=1}^{r}p(\omega_j\cap X_i)\log p(\omega_j\,|\,X_i). \quad (8)$$

$H(D\,|\,B)$ is used to characterize the ability of B to distinguish samples with different class labels. The less $H(D\,|\,B)$ is, the greater the distinguishment ability of B is. When the attribute subset B completely divides all samples into their respective categories, then $H(D\,|\,B)=0$. According to the relationship between conditional entropy and mutual information, we can easily know that mutual information grows greater with the increase of the distinguishment ability of an attribute subset.

III. Neighborhood Discrimination Index and Its Variants

In the following discussions, a data set used for classification learning will be written as a decision table and denoted by $<U,A,D>$, where $U=\{x_1,x_2,\cdots,x_n\}$ is a nonempty set of samples, called a universe; $A=\{a_1,a_2,\cdots,a_m\}$ is a set of conditional attributes to characterize the samples, and D is a decision attribute and partitions the universe into r crisp equivalence classes $U/D=\{D_1,D_2,\cdots D_r\}$. The sign $|\cdot|$ is used to denote the cardinality of a set or relation.

In this section, a new measure, called neighborhood discrimination index, is proposed to compute the distinguishment ability of a feature subset. We begin with introducing the notion of neighborhood relations based on distance functions.

Given a feature subset $B\subseteq A$, R_B is a binary relation generated by B, We say R_B is a crisp similarity relation on U if R_B satisfy

1. Reflexivity: $(x,x)\in R_B$, $\forall x\in U$;
2. Symmetry: $(x,y)\in R_B\Rightarrow(y,x)\in R_B$ for any $x,y\in U$.

A crisp similarity relation R_B on the universe can be represented by a similarity matrix, generally denoted as $R_B=(r_{ij})_{n\times n}$, where $r_{ij}\in\{0,1\}$, $i,j=1,2,\cdots,n$. There are many ways to calculate r_{ij}, here we use the following measures:

$$r_{ij}=\begin{cases}1, & \Delta_p^B(x_i,x_j)\le\varepsilon,\\ 0, & \Delta_p^B(x_i,x_j)>\varepsilon;\end{cases} \quad (9)$$

where $x_l=[x_{l1},x_{l2},\cdots,x_{ls}]^T$, $l=i,j$ are two samples, T stands for the transpose operation of vector, B is a subset of attributes with $|B|=s$ and

$$\Delta_p^B(x_i,x_j)=\sqrt[p]{\sum_{k=1}^{s}\left\|x_{ik}-x_{jk}\right\|^p}, \quad (10)$$

here $\|\cdot\|$ stands for absolute value. Δ_p^B is called Manhattan distance if P=1, Euclidean distance if P=2, and Chebychev distance if $p=\infty$. ε is a threshold that is used to control sample similarity. We call threshold ε the radius of neighborhood. A similarity relation induced by distance function Δ_p^B and neighborhood radius ε is called a neighborhood similarity relation and denoted as R_B^ε. Let $R_{B_1}^{\varepsilon_1}$ and $R_{B_2}^{\varepsilon_2}$ be two neighborhood similarity relations, we say $R_{B_1}^{\varepsilon_1}$ is finer than $R_{B_2}^{\varepsilon_2}$ if $R_{B_1}^{\varepsilon_1}\subseteq R_{B_2}^{\varepsilon_2}$.

According to the above definition, we know that samples x_i and x_j are distinguishable if their distance is more than neighborhood radius ε with respect to feature subset B, i.e. $\Delta_p^B(x_i,x_j)>\varepsilon$; otherwise, they are indistinguishable. The finer a neighborhood similarity relation is, the greater its distinguishment ability is. There are two factors that impact on a neighborhood similarity relation. One is neighborhood radius ε, the other is feature subset B. For a given parameter ε, the neighborhood relation becomes finer as the number of features in B increases. The property can be formulated as follows.

Property 1. Let $B\subseteq A$, then $R_A^\varepsilon\subseteq R_B^\varepsilon$.

As we know, a neighborhood similarity relation characterizes the distinguishment ability of a feature subset. Property 1 shows that the more the number of features is, the finer its neighborhood relation is, and the greater the feature subset has distinguishment ability.

In the following, we introduce a new concept to measure the distinguishment ability of a feature subset.

Definition 1. Given a decision table $<U,A,D>$, where $U=\{x_1,x_2,\cdots,x_n\}$, $B\subseteq A$, ε is a neighborhood radius, and R_B^ε is the neighborhood similarity relation induced by B. The neighborhood discrimination index of B is defined as:

$$H^\varepsilon(B)=\log\frac{n^2}{|R_B^\varepsilon|}. \quad (11)$$

It is easily seen that $H^\varepsilon(B)\ge 0$ by the fact that $|R_B^\varepsilon|\le n^2$. It follows from the reflexivity of R_B^ε that $H^\varepsilon(B)\le\log n$. In particular, $H^\varepsilon(B)=\log n$ if $|R_B^\varepsilon|=n$, and $H^\varepsilon(B)=0$ if $|R_B^\varepsilon|=n^2$.

The neighborhood discrimination index measures the uncertainty quantity of distinguishment ability of a feature subset. It is a mapping from a feature space to the real space:

$H:(B,\varepsilon)\to R^+$, where R^+ is the domain of non-negative real numbers. With the mapping the distinguishment abilities of different feature subsets can be compared.

Compared with neighborhood entropy [13], neighborhood discrimination index has two main differences as follows.

1. The concept of neighborhood discrimination index is based on neighborhood relations, it can be directly obtained by computing the cardinality of neighborhood relation, while neighborhood entropy is defined on the neighborhood similarity classes and accumulatively obtained by considering the cardinality of similarity classes. Thus, the computational complexity of neighborhood discrimination index is somewhat less than neighborhood entropy.

2. Neighborhood entropy is a variant of Yager's entropy and is degenerated into Shannon entropy when a neighborhood relation degrades to an equivalence relation. So neighborhood entropy is a generalization of Shannon entropy, while neighborhood discrimination index is just a measure of the distinguishment ability of a feature subset. This is the essential difference between them.

Note that neighborhood discrimination index is not only a function of feature subset B, but also related to neighborhood radius ε. Next, we discuss the influence of neighborhood radius and feature subset on the discrimination index.

Proposition 1. If $\varepsilon_1 \le \varepsilon_2$, then $H^{\varepsilon_1}(B) \ge H^{\varepsilon_2}(B)$.

Proof. Let $(x_i,x_j)\in R_B^{\varepsilon_1}$, then $\Delta_p^B(x_i,x_j)\le\varepsilon_1$. From $\varepsilon_1\le\varepsilon_2$, we have that $\Delta_p^B(x_i,x_j)\le\varepsilon_2$, which implies $(x_i,x_j)\in R_B^{\varepsilon_2}$. Hence, $R_B^{\varepsilon_1}\subseteq R_B^{\varepsilon_2}$, and then $\left|R_B^{\varepsilon_1}\right|\le\left|R_B^{\varepsilon_2}\right|$. It follows $H^{\varepsilon_1}(B)\ge H^{\varepsilon_2}(B)$ by the definition of neighborhood discrimination index.

This property shows that the discrimination index of a feature subset becomes smaller as the radius of neighborhood increases. A small neighborhood radius means that the corresponding neighborhood relation is finer. Hence, the uncertainty quantity of distinguishment ability of the feature subset is greater.

Proposition 2. If $B_1\subseteq B_2$, then $H^\varepsilon(B_1)\le H^\varepsilon(B_2)$.

Proof. Let $(x_i,x_j)\in R_{B_2}^{\varepsilon}$, then $\Delta_p^{B_2}(x_i,x_j)\le\varepsilon$. From $B_1\subseteq B_2$, we have that $\Delta_p^{B_1}(x_i,x_j)\le\varepsilon$, which implies $(x_i,x_j)\in R_{B_1}^{\varepsilon}$. Hence, $R_{B_2}^{\varepsilon}\subseteq R_{B_1}^{\varepsilon}$, and then $\left|R_{B_2}^{\varepsilon}\right|\le\left|R_{B_1}^{\varepsilon}\right|$. It follows $H^\varepsilon(B_1)\le H^\varepsilon(B_2)$ by the definition of neighborhood discrimination index.

Proposition 2 shows that the neighborhood discrimination index is affected by the number of features. It increases monotonously with the size of a feature subset.

Definition 2. Let B_1, B_2 be two groups of features, ε be a neighborhood radius and $R_{B_1}^{\varepsilon}$, $R_{B_2}^{\varepsilon}$ be two neighborhood similarity relations induced by B_1, B_2, respectively. Then, the joint discrimination index of B_1 and B_2 is defined as:

$$H^\varepsilon(B_1,B_2)=\log\frac{n^2}{\left|R_{B_1}^{\varepsilon}\cap R_{B_2}^{\varepsilon}\right|} \tag{11}$$

The joint discrimination index represents the distinguishment ability of a joint feature subset. It increases with addition of some new features. Formally, the property can be expressed as follows.

Proposition 3. $H^\varepsilon(B_1,B_2)\ge H^\varepsilon(B_1)$, $H^\varepsilon(B_1,B_2)\ge H^\varepsilon(B_2)$.

It is easily to see that the joint discrimination index of B_1 and B_2 is greater than any individual discrimination index. It is interpreted that the distinguishment ability of the joint features gets stronger with the addition of new features. This is because we can get a finer neighborhood relation by introducing new features.

Proposition 4. If $B_1\subseteq B_2$, then $H^\varepsilon(B_1,B_2)=H^\varepsilon(B_2)$.

This property shows that addition of some new features will not bring increment of discrimination index if these features are contained in other existing features. In this case, the distinguishment information has been implied in the existing feature subset.

Definition 3. Let B_1, B_2 be two groups of features, ε be a neighborhood radius and $R_{B_1}^{\varepsilon}$, $R_{B_2}^{\varepsilon}$ be two neighborhood similarity relations induced by B_1, B_2, respectively. Then, the conditional discrimination index of B_1 on B_2 is defined as

$$H^\varepsilon(B_1\mid B_2)=\log\frac{\left|R_{B_2}^{\varepsilon}\right|}{\left|R_{B_1}^{\varepsilon}\cap R_{B_2}^{\varepsilon}\right|} \tag{12}$$

Since $\left|R_{B_1}^{\varepsilon}\cap R_{B_2}^{\varepsilon}\right|\le\left|R_{B_2}^{\varepsilon}\right|$, it is easily seen that $H^\varepsilon(B_1\mid B_2)\ge 0$. When $B_1\subseteq B_2$, then $R_{B_1}^{\varepsilon}\supseteq R_{B_2}^{\varepsilon}$. This means $H^\varepsilon(B_1\mid B_2)=0$. When $\left|R_{B_2}^{\varepsilon}\right|=n^2$ and $R_{B_1}^{\varepsilon}$ is an identity matrix, the conditional discrimination index reaches the maximum value. That is, $H^\varepsilon(B_1\mid B_2)=\log n$.

According the above discussion, we easily get the following property.

Proposition 5. Let B_1, B_2 be two groups of features. Then

(1) $H^\varepsilon(B_1\cup B_2)\ge\max\left\{H^\varepsilon(B_1),H^\varepsilon(B_2)\right\}$;

(2) $H^\varepsilon(B_1\mid B_2)=0$ if $B_1\subseteq B_2$.

The first item indicates that the discrimination index of the union of two feature subsets will be no smaller than that of any single subset. The last item shows feature subset B_1 won't introduce distinguishment information with respect to B_2 if B_1 is contained in B_2.

Proposition 6. Let B_1, B_2 be two groups of features. Then

$$H^{\varepsilon}(B_1|B_2)=H^{\varepsilon}(B_1,B_2)-H^{\varepsilon}(B_2). \quad (13)$$

Proof:

$$H^{\varepsilon}(B_1,B_2)-H^{\varepsilon}(B_2)=\log\frac{n^2}{\left|R^{\varepsilon}_{B_1}\cap R^{\varepsilon}_{B_2}\right|}-\log\frac{n^2}{\left|R^{\varepsilon}_{B_2}\right|}$$
$$=\log\frac{n^2}{\left|R^{\varepsilon}_{B_1}\cap R^{\varepsilon}_{B_2}\right|}\cdot\frac{\left|R^{\varepsilon}_{B_2}\right|}{n^2}=\log\frac{\left|R^{\varepsilon}_{B_2}\right|}{\left|R^{\varepsilon}_{B_1}\cap R^{\varepsilon}_{B_2}\right|}.$$

It is easily observed that the conditional discrimination index is the increment of distinguishment information by introducing a new feature subset after one feature subset has been known. It reflects the increment of distinguishment ability under the addition of a new feature subset.

Remark 1. Conditional discrimination index $H^{\varepsilon}(B_1|B_2)$ is not monotonic with the size of attribute subset B_2.

Definition 4. Let B_1, B_2 be two groups of features, ε be a neighborhood radius and $R^{\varepsilon}_{B_1}$, $R^{\varepsilon}_{B_2}$ be two neighborhood similarity relations induced by B_1, B_2, respectively. Then, the mutual discrimination index of B_1 and B_2 is defined as:

$$I^{\varepsilon}(B_1;B_2)=\log\frac{n^2\left|R^{\varepsilon}_{B_1}\cap R^{\varepsilon}_{B_2}\right|}{\left|R^{\varepsilon}_{B_1}\right|\cdot\left|R^{\varepsilon}_{B_2}\right|} \quad (14)$$

Proposition 7. Let B_1, B_2 be two groups of features, then we have the following properties.

1. $I^{\varepsilon}(B_1;B_2)=I^{\varepsilon}(B_2;B_1)$;
2. $I^{\varepsilon}(B_1;B_2)=H^{\varepsilon}(B_1)+H^{\varepsilon}(B_2)-H^{\varepsilon}(B_1,B_2)$;
3. $I^{\varepsilon}(B_1;B_2)=H^{\varepsilon}(B_1)-H^{\varepsilon}(B_1|B_2)=H^{\varepsilon}(B_2)-H^{\varepsilon}(B_2|B_1)$. (15)

Proof: (1) Straightforward.

(2) $H^{\varepsilon}(B_1)+H^{\varepsilon}(B_2)-H^{\varepsilon}(B_1,B_2)$

$$=\log\frac{\left|n^2\right|}{\left|R^{\varepsilon}_{B_1}\right|}+\log\frac{\left|n^2\right|}{\left|R^{\varepsilon}_{B_2}\right|}-\log\frac{\left|n^2\right|}{\left|R^{\varepsilon}_{B_1}\cap R^{\varepsilon}_{B_2}\right|}$$
$$=\log\frac{n^2\cdot\left|R^{\varepsilon}_{B_1}\cap R^{\varepsilon}_{B_2}\right|}{\left|R^{\varepsilon}_{B_1}\right|\cdot\left|R^{\varepsilon}_{B_2}\right|}=I^{\varepsilon}(B_1;B_2)\ .$$

(3) $H^{\varepsilon}(B_1)-H^{\varepsilon}(B_1|B_2)=\log\frac{\left|n^2\right|}{\left|R^{\varepsilon}_{B_1}\right|}-\log\frac{\left|R^{\varepsilon}_{B_2}\right|}{\left|R^{\varepsilon}_{B_1}\cap R^{\varepsilon}_{B_2}\right|}$

$$=\log\frac{n^2\cdot\left|R^{\varepsilon}_{B_1}\cap R^{\varepsilon}_{B_2}\right|}{\left|R^{\varepsilon}_{B_1}\right|\cdot\left|R^{\varepsilon}_{B_2}\right|}=I^{\varepsilon}(B_1;B_2).$$

Similarly, we have $H^{\varepsilon}(B_2)-H^{\varepsilon}(B_2|B_1)=I^{\varepsilon}(B_1;B_2)$.

The first item shows the mutual discrimination index of B_1 and B_2 is symmetric. The second says that the mutual discrimination index is the difference between the sum of the discrimination indexes of two feature subsets and their joint discrimination index. The last item shows that the mutual discrimination index is the difference between the discrimination index of one of two feature subsets and their conditional discrimination index. It reflects that mutual discrimination index is the common part of distinguishment information of two feature subsets. The relationship between neighborhood, conditional and mutual discrimination indexes can be explained in Figure 1.

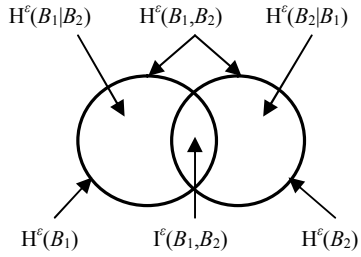

Fig. 1 Relationship diagram of discrimination indexes

Remark 2. Given a decision table $<U,A,D>$, $B\subseteq A$ and neighborhood radius ε, R^{ε}_B is a neighborhood relation induced by B and ε, R_D is an equivalence relation induced by D. Just like Shannon conditional entropy, $H(D|B)$ can be used to characterize the ability of B to distinguish samples with different decisions. The less the value of $H(D|B)$ is, the greater the distinguishment ability of B is. When all samples are rightly grouped into their respective categories, then $H(D|B)=0$. According to the relationship between conditional and mutual discrimination indexes, we can easily conclude that mutual discrimination index grows greater as the distinguishment ability of a feature subset increases. Moreover, we know that $H(D|B)$ and $I(D;B)$ are not monotonic with the size of feature subset B from Remark 1.

In many practical problems, we always assign a class label to a sample according to other samples' labels in its neighborhood. If all samples in the neighborhood have the same labels, then the sample is called consistent. Otherwise, the sample is inconsistent. Let ε be a neighborhood radius, if all samples are consistent, then $<U,A,D>$ is called consistent. Otherwise, it is called inconsistent. It is obviously seen that $<U,A,D>$ is consistent with respect to A if and only if $R^{\varepsilon}_A\subseteq R_D$.

Proposition 8. If a decision table is consistent with respect to B, i.e., $R^{\varepsilon}_B\subseteq R_D$, then

1. $H^{\varepsilon}(D\,|\,B)=0$, 2. $I^{\varepsilon}(D;B)=H^{\varepsilon}(D)$.

The first item shows that the conditional discrimination index equals to zeros if the classification is consistent. In this case, all samples can be rightly classified into their respective classes by feature subset B. The second item shows the mutual discrimination index between B and D equals to the distinguishment information quantity of D if the classification is consistent.

As we know, Shannon mutual information is widely used in feature selection algorithms for categorical data. An optimal feature subset for classification learning should be sufficient

and necessary. Because conditional entropy is not monotonic with the size of feature subset, sufficiency should guarantee that the selected features have the maximal capability in distinguishing samples with different decisions. Necessity requires no redundant features in the selected feature subset. Inspired by the idea, we present an axiomatic approach to feature selection as follows.

Axiom 1 (*Maximum of classification information*). Given a decision table $<U,A,D>$, the expected feature subset B is sufficient if $I^{\varepsilon}(D;B)\geq I^{\varepsilon}(D;A)$ under neighborhood radius ε.

Axiom 2 (*Minimum encoding length*). Given a decision table $<U,A,D>$, $\mathbb{N}$ is a set of sufficient feature subsets, and $B\in\mathbb{N}$. Then B is favored with respect to its predictive capability if $I^{\varepsilon}(D,B)=\max_{C\in\mathbb{N}} I^{\varepsilon}(D,C)$.

The proposed axiomatic system presents a multi-granular way to describe the classification ability of a set of numerical features if neighborhood radius ε is considered as a variable.

The axiomatic system also shows a goal for feature selection. It can be formally expressed as the following definition.

Definition 5. Given a decision table $<U,A,D>$, B is a subset of A and $a\in B$. a is called redundant in B relative to D if $I^{\varepsilon}(D;B)\leq I^{\varepsilon}(D;B-\{a\})$. Otherwise, we say a is indispensable in B relative to D; B is called dependent if any attribute in B is indispensable relative to D. B is called a reduct of A relative to decision D if B satisfies:

(1) $I^{\varepsilon}(D;B)\geq I^{\varepsilon}(D;A)$,

(2) $I^{\varepsilon}(D;B-\{a\})<I^{\varepsilon}(D;B)$, $\forall a\in B$.

Obviously, a reduct of A relative to D is the minimal feature subset to keep or improve the mutual discrimination index of A and D.

According to the relationships between neighborhood, conditional and mutual discrimination indexes, we can easily know that the above two conditions for feature selection is equivalent to the following conditions.

(1) $H^{\varepsilon}(D\mid B)\leq H^{\varepsilon}(D\mid A)$,

(2) $H^{\varepsilon}(D\mid B-\{a\})>H^{\varepsilon}(D\mid B)$ $\forall a\in B$.

Example 1. Given a set $X=\{x_1,x_2,x_3\}$, R_1, R_2, and R_3 are relations defined on X, where

$$R_1=\begin{bmatrix}1&0&1\\0&1&0\\1&0&1\end{bmatrix}\ R_2=\begin{bmatrix}1&0&0\\0&1&1\\0&1&1\end{bmatrix},\ R_3=\begin{bmatrix}1&1&1\\1&1&0\\1&0&1\end{bmatrix}.$$

We have

$$R_1\cap R_2=\begin{bmatrix}1&0&0\\0&1&0\\0&0&1\end{bmatrix},R_1\cap R_3=\begin{bmatrix}1&0&1\\0&1&0\\1&0&1\end{bmatrix},\ R_2\cap R_3=\begin{bmatrix}1&0&0\\0&1&0\\0&0&1\end{bmatrix}.$$

Suppose the decision equivalence relation

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

We compute

$H(R_1)=\log\frac{9}{5}=0.8480$, $H(R_2)=\log\frac{9}{3}=1.5850$,

$H(R_3)=\log\frac{9}{7}=0.3626$, $H(R_1R_2)=\log\frac{9}{3}=1.5850$,

$H(R_2R_3)=\log\frac{9}{3}=1.5850$, $H(R_1R_3)=\log\frac{9}{5}=0.8480$.

According to Proposition 6, we know,

$H(R_1\mid R_2)=H(R_1R_2)-H(R_2)=\log\frac{9}{3}-\log\frac{9}{3}=0$,

$H(R_2\mid R_1)=H(R_1R_2)-H(R_1)=\log\frac{9}{3}-\log\frac{9}{5}=\log\frac{5}{3}=0.7370$.

We can find $H(R_d\mid R_1R_2)=H(R_d\mid R_2R_3)=H(R_d\mid R_1R_2R_3)$. Hence, $\{a_1,a_2\}$ and $\{a_2,a_3\}$ are two reducts.

IV. Feature Selection Algorithm Based on Neighborhood Discrimination Index

As discussed above, the proposed discrimination indexes can be used to measure the distinguishment ability of a relation or a feature subset. The less a decision attribute has conditional discrimination index with respect to a feature subset, the more the feature subset has distinguishment ability and the more important the feature subset is. According to the definition of conditional discrimination index, adding a new feature into the selected feature subset, the conditional distinguishment index of a decision attribute may increase or decrease. A feature can lead to decrease the index only when it is irrelevant to the selected feature subset. The decrement of conditional discrimination index reflects the increment of distinguishment ability produced by a new feature subset. So the significance of a feature can be defined as follows.

Definition 6. Given decision table $<U,A,D>$, $B\subseteq A$, $a\in A-B$, the significance degree of feature a with respect to B and D is defined as:

$$SIG(a,B,D)=H^{\varepsilon}(D\mid B)-H^{\varepsilon}(D\mid B\cup\{a\}). \tag{16}$$

When $B=\varnothing$, we define $H^{\varepsilon}(D\mid B)=H^{\varepsilon}(D)$. The significance of attribute a depends on the increment of distinguishment information after adding a into B. A big value of $SIG(a,B,D)$ indicates that attribute a is more important for decision D.

Based on the above definition, a greedy algorithm for computing an optimal feature subset can be designed as follows.

Algorithm: Heuristic algorithm based on neighborhood discrimination index (HANDI)

Input: decision table $<U,A,D>$ and ε // ε is the neighborhood radius.

Output: one reduct *red*.

1: Initialize: $red=\varnothing$, $B=A-red$, start=1; // red is the pool to contain the selected attributes and B is for the left attributes.
2: while start
3: for each $a_i \in B$
4: Compute neighborhood relation $R^{\varepsilon}_{red\cup\{a_i\}}$.
5: Compute
$$SIG(a_i, red, D) = H^{\varepsilon}(D|red) - H^{\varepsilon}(D|red\cup\{a_i\});$$
6: end for
7: Find a_k with maximum value $SIG(a_k, red, D)$.
8: if $SIG(a_k, red, D) > \delta$
9: $red \leftarrow red \cup \{a_k\}$;
10: $B \leftarrow B - red$;
11: else
12: start=0;
13: end if
14: end while
15: return red.

The parameter δ is used to stop the main loop in this algorithm. It need to be set up in advance. For a given data set, generally speaking, the number of the selected features gets bigger if the value of the parameter δ increases. The algorithm employs $SIG(a,B,D)$ to evaluate which attribute is optimal and will be added into the current selected feature subset in each loop. This algorithm terminates when the addition of any remaining attribute does not decrease the evaluating function. For a dimensionality of N, the time complexity for computing neighborhood similarity relation is N, the worst search time for a reduct will result in $(N^2+N)/2$ evaluations of the evaluation function. The overall time complexity of the algorithm is $O\left((N^2+N)/2\right)$.

V. Experimental Analysis

In order to verify the feasibility and effectiveness of the proposed algorithm, we compare the proposed algorithm with neighborhood rough set based algorithm (NRS) [15], neighborhood entropy based algorithm (NEIEN) [13], fuzzy information entropy based algorithm (FINEN) [14], [55] and fuzzy rough dependency constructed by intersection operations of fuzzy similarity relations (FRINT) [18]. We employ Chebychev distance function to compute neighborhood similarity relations. We first compare (1) the numbers of selected features, (2) the running time of reduction, and (3) classification accuracies based on these algorithms. Then, we discuss the influence of neighborhood radius ε on our proposed algorithm. All the algorithms are performed in Matlab 2013b and run in a hardware environment with a Intel (R) Core (TM) i7-4790 CPU @ 3.60 GHz, with 16.0 GB RAM.

We employ ten-fold cross validation and two classical classifiers to evaluate these algorithms. The two classifiers are support vector machine (RBF-SVM) and k-nearest neighbor rule (K-NN, K=3). Since our main purpose is to compare the performances of different feature selection algorithms, the parameter selection for RBF-SVM is not our concern. Thus, in this experiment, we consistently set the control term C as 100 and the Gaussian kernel parameter g as 1. Such parameter specifications can perform well on real-world problem [61]. The experimental comparison is conducted based on a ten-fold cross-validation. That is to say, the original data set is randomly divided into ten subsets, of which one is used as the testing data and the remaining nine are used for training. Feature selection is performed on the training set; the reduced training and testing sets are then sent to a classifier to produce the classification accuracy. After ten rounds, the average value and variation of the classification accuracies are computed as the final performance. Thirteen data sets are used in the experimental analysis. They are selected from UCI Machine Learning Repository [3] and Keng Ridge Bio-medical (KRBM) Data Set Repository [62]. The information of these data sets is outlined in Tables 1. All the numerical attributes are first normalized into the interval [0,1].

Table 1 Description of data sets

No	Data sets	Sample	Attributes	Class
1	Wine	178	13	3
2	Wdbc	569	31	2
3	Wpbc	198	33	2
4	Sonar	208	60	2
5	Credit	690	15	2
6	Sick	2800	29	2
7	Gearbox	1603	72	4
8	Segmentation	2310	19	7
9	DLBCL	77	5469	2
10	Leukemia	72	11225	3
11	MLL	72	12582	3
11	Prostate	136	12600	2
13	Tumors	327	12558	7

There are two parameters in HANDI algorithm, ε and δ. The parameter ε is introduced to control sample similarity; it has a great impact on the performance of the algorithm. Generally speaking, different values of neighborhood radius can lead to different classification accuracies, therefore, we select an optimal feature subset for each data set by adjusting the value of the parameter to vary from 0 to 1 with a step of 0.05. The parameter δ is set as 0.001 for low dimensional data and 0.01 for high dimensional data. As different learning algorithms may require different feature subsets to produce the best classification accuracy, all the experimental results reported in the following tables are presented at highest classification accuracy.

Table 2 presents the comparison of the average sizes of the selected features with different algorithms. Because the highest classification accuracy of each data set is searched by adjusting the values of ε, the values of parameter ε are different for the highest accuracies of data sets. The last column in Table 2 marked with ε shows the values of the neighborhood radius in HANDI algorithm, where the best classification performances are produced on the corresponding datasets.

From the Table 2, we can find that these reduction methods

can effectively reduce attributes. The numbers of selected features with HANDI are fewer than other four algorithms in most of the cases. For Sonar data set, HANDI gets more features than FRSINT algorithm, but less than NRS, NEIEN, and FINEN algorithms. For Tumors, HANDI gets more features than NRS and FRSINT, but less than NEIEN and FINEN. This implies the proposed algorithm is more effective to reduce redundant attributes.

Table 2 Average sizes of feature subsets selected with 10-fold cross validation

Data sets	Raw data	NRS	NEIEN	FINEN	FRSINT	HANDI	ε
Wine	13	9.1	10.2	12.3	8.1	8.3	0.2
Wdbc	30	17.3	11.8	12.1	11.9	11.2	0.1
Wpbc	32	11.6	5.3	6.4	7.8	5.1	0.6
Sonar	60	24.8	28.9	25.8	18.7	21.6	0.6
Credit	15	10.2	4.4	8.1	9.8	4.4	0.35
Sick	29	8.3	13.1	12.7	8.4	7.6	0.05
Gearbox	72	17.4	10.9	11.4	10.1	9.3	0.4
Segmentation	19	10.7	9.2	9.5	8.4	8.7	0.15
DLBCL	5469	8.3	5.3	6.1	8.8	5.2	0.25
Leukemia	11225	14.7	8.2	8.5	9.8	6.3	0.4
MLL	12582	6.4	8.2	9.5	10.2	6.9	0.45
Prostate	12600	6.5	7.7	8.4	8.9	3.4	0.4
Tumors	12558	10.6	17.1	15.8	9.5	15.7	0.35
Average	4208	11.99	10.79	11.28	10.03	8.77	

Table 3. Comparison of classification accuracies of reduced data with SVM (%)

Data sets	Raw data	NRS	NEIEN	FINEN	FRSINT	HANDI
Wine	97.58 ± 4.68	96.09 ± 3.93	96.49 ± 4.39	97.93 ± 2.90	96.80 ± 3.85	97.91 ± 2.16
Wdbc	96.32 ± 2.51	97.06 ± 2.53	96.72 ± 2.23	95.99 ± 3.69	96.88 ± 3.26	97.42 ± 2.45
Wpbc	76.66 ± 8.79	77.79 ± 7.50	80.09 ± 8.21	79.64 ± 10.89	79.48 ± 9.46	81.48 ± 8.46
Sonar	86.26 ± 7.16	87.29 ± 9.94	86.71 ± 6.35	87.54 ± 5.86	87.77 ± 8.68	87.32 ± 5.12
Credit	82.40 ± 5.16	83.33 ± 2.77	83.61 ± 3.98	83.72 ± 4.50	84.09 ± 5.22	85.54 ± 4.31
Sick	95.57 ± 1.46	95.38 ± 0.72	96.49 ± 0.68	96.49 ± 0.81	95.02 ± 1.68	96.21 ± 0.51
Gearbox	98.90 ± 1.15	99.41 ± 0.30	98.75 ± 1.74	98.93 ± 1.63	98.24 ± 0.44	99.25 ± 1.34
Segmentation	95.20 ± 1.34	92.82 ± 6.76	94.67 ± 1.11	94.84 ± 1.67	95.75 ± 1.37	96.77 ± 1.58
DLBCL	75.50 ± 11.89	98.35 ± 6.04	97.10 ± 6.59	95.85 ± 8.27	98.35 ± 5.95	98.35 ± 9.51
Leukemia	46.79 ± 15.19	96.17 ± 4.52	94.20 ± 6.13	95.74 ± 4.52	97.55 ± 3.01	98.49 ± 6.45
MLL	41.11 ± 13.24	98.14 ± 6.90	99.67 ± 6.03	98.09 ± 7.21	98.07 ± 5.41	99.60 ± 3.02
Prostate	57.29 ± 15.23	90.31 ± 6.78	93.60 ± 6.69	95.74 ± 5.35	91.17 ± 5.90	93.89 ± 5.64
Tumors	27.88 ± 12.36	82.69 ± 7.53	83.52 ± 6.40	84.52± 7.10	80.39 ± 7.31	84.34 ± 6.78
Average	75.27 ± 7.70	91.91 ± 5.09	92.43 ± 4.66	92.69 ± 4.95	92.27 ± 4.73	93.58 ± 4.41

The classification accuracies of the raw data and the reduced data sets based on the five algorithms are shown in Tables 3 and 4, where the underline symbols highlight the highest classification accuracies among the reduced data sets. From the results of Tables 3-4, it is easily seen that the classification accuracies based on NRS method are lower than other four methods. Out of 26 cases of ten-fold cross validation, the HANDI and FINEN methods achieve highest classification accuracy in 13 and 7 cases, while the NRS, NEIEN, and FRSINT methods obtain it in 3, 5 and 2 cases, respectively. As for SVM, HANDI outperforms the raw data 12 times over the 13 classification tasks. In the same time, it outperforms the raw data 11 times with respect to 3NN. Moreover, the average accuracy of HANDI outperforms any other feature selection algorithm in terms of SVM and 3NN learning algorithms.

From table 5, we can find that the running time of reduction of NRS algorithm is the shortest in the five different algorithms. HANDI algorithm runs slower than NRS algorithm, but faster than other three algorithms. The running time of FRSINT algorithm is the longest. As NEIEN, FINEN, FRSINT and HANDI algorithms are based on the similarity relations, they have to spend a lot of time to compute similarity relations of attributes. NRS algorithm does not compute similarity relations, it just take some time to judge if samples in a neighborhood are similar or not. So NRS algorithm runs fastest. Because NEIEN and FINEN algorithms need additional time to compute the similarity class of each sample

based on similarity relations, they run slower than HANDI algorithm. For FRSINT algorithm, it not only depend on similarity relations, but also need time to compute the fuzzy-rough membership of each sample to different decision categories. Therefore, FRSINT algorithm run slowest. From Tables 3 and 4, we know most of the classification accuracies of the HANDI algorithm are higher than that of other four algorithms. The complexity of HANDI is lower than NEIEN, FINEN and FRSINT algorithms. Therefore, it can be concluded that the HANDI algorithm is both feasible and effective.

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

Data sets	Raw data	NRS	NEIEN	FINEN	FRSINT	HANDI
Wine	96.28 ± 4.36	96.36 ± 1.83	96.20 ± 2.53	95.77 ± 4.68	96.72 ± 3.91	97.86 ± 2.94
Wdbc	96.66 ± 2.34	97.01 ± 1.48	96.27 ± 2.69	95.38 ± 3.12	95.97 ± 2.31	96.39 ± 2.53
Wpbc	74.45 ± 9.69	77.89 ± 10.37	77.43 ± 8.11	76.82 ± 10.69	76.72 ± 13.60	78.38 ± 9.38
Sonar	83.66 ± 7.28	85.88 ± 6.82	83.26 ± 11.89	85.13 ± 8.56	85.04 ± 8.54	89.60 ± 8.16
Credit	84.42 ± 3.99	84.08 ± 4.51	86.32 ± 2.86	86.11 ± 3.19	82.32 ± 2.86	86.25 ± 1.94
Sick	95.01 ± 1.51	95.24 ± 1.24	96.04 ± 1.24	96.04 ± 1.01	95.68 ± 2.13	95.91 ± 0.88
Gearbox	99.69 ± 1.44	99.29 ± 0.33	99.35 ± 1.74	99.13 ± 1.63	99.16 ± 0.59	99.33 ± 1.02
Segmentation	95.89 ± 1.26	89.77 ± 9.56	96.08 ± 1.20	96.16 ± 1.12	96.24 ± 1.18	96.64 ± 1.53
DLBCL	86.99 ± 10.48	96.10 ± 5.27	95.85 ± 3.95	96.35 ± 5.36	96.35 ± 5.27	97.10 ± 6.04
Leukemia	84.61 ± 11.22	92.74 ± 6.03	95.63 ± 7.03	92.46 ± 6.60	96.33 ± 9.78	97.06 ± 8.71
MLL	84.29 ± 11.71	95.71 ± 4.35	95.22 ± 9.64	95.31 ± 6.90	94.85 ± 6.45	98.17 ± 4.52
Prostate	79.00 ± 12.21	83.29 ± 8.27	84.31 ± 10.06	86.89 ± 6.80	85.03 ± 7.48	86.46 ± 9.25
Tumors	76.76 ± 6.02	79.81 ± 7.31	80.05 ± 6.33	82.72 ± 4.91	79.71 ± 7.32	81.39 ± 9.42
Average	87.51 ± 6.42	90.24 ± 5.18	90.92 ± 5.33	91.10 ± 4.97	90.78 ± 5.49	92.35 ± 5.10

Table 5. Running time of reduction with different algorithms (s)

Data sets	NRS	NEIEN	FINEN	FRSINT	HANDI
Wine	0.04 ± 0.01	0.14 ± 0.03	0.11 ± 0.06	0.28 ± 0.04	0.03 ± 0.01
Wdbc	0.88 ± 0.03	4.29 ± 0.22	4.20 ± 0.37	4.37 ± 0.31	2.56 ± 0.30
Wpbc	0.11 ± 0.04	0.33 ± 0.08	0.33 ± 0.09	0.73 ± 0.11	0.20 ± 0.05
Sonar	0.54 ± 0.04	1.65± 0.10	1.73 ± 0.25	2.82 ± 0.22	1.01 ± 0.10
Credit	0.48 ± 0.03	2.18 ± 0.16	2.51 ± 0.10	3.45 ± 0.06	1.64 ± 0.23
Sick	6.21 ± 1.41	137.69 ± 10.41	135.81 ± 16.53	141.91 ± 13.21	71.06 ± 9.13
Gearbox	5.41 ± 0.36	115.05 ± 12.48	109.44 ± 10.23	176.81 ± 13.09	86.03 ± 10.81
Segmentation	1.18 ± 0.05	47.50 ± 6.14	45.74 ± 5.69	114.96 ± 11.22	35.35 ± 10.04
DLBCL	2.69 ± 0.08	13.18 ± 2.17	14.45 ± 3.55	46.05 ± 9.16	3.82 ± 0.51
Leukemia	8.85 ± 3.36	24.96 ± 4.68	29.33 ± 3.98	137.11 ± 12.53	8.05 ± 3.15
MLL	9.05 ± 2.96	34.62 ± 3.65	39.22 ± 2.11	127.59 ± 10.46	8.85 ± 1.53
Prostate	13.88 ± 3.32	79.98 ± 10.77	77.59 ± 9.99	234.05 ± 11.07	26.32 ± 1.55
Tumors	74.34 ± 8.98	648.35 ± 16.72	620.19 ± 18.65	1875.23 ± 23.16	315.01 ± 15.57
Average	9.51 ± 1.59	85.38 ± 5.20	83.13 ± 5.51	220.41 ± 8.05	43.07 ± 4.08

To show the selected feature subset of a data set, in the following we employ the NEIEN, FINEN and HANDI algorithms to reduce the entire data set based on parameters where the classification accuracies are obtained in the above experiments. The selected feature subsets are listed in Table 6. It is seen that the best feature subsets for SVM and 3NN are identical to each other in some cases, but they are different in general. This shows that no algorithm is consistently better than others for different learning tasks and classification algorithms. For the Sick and Segmentation data sets, the selected feature subsets are identical and the classification accuracies are almost the same for the NEIEN and FINEN algorithms. The slight differences for Segmentation may be due to the fact that the selected feature subsets are presented by reducing the entire data set, while the classification accuracies are based on ten-fold cross-validation.

Now, we present some figures to demonstrate the numbers of selected features and classification accuracies varying with ε, we only display the curves of some data sets with SVM. The data curves drawn by using 3NN are roughly consistent with SVM.

From Figures 2-13, it is easily observed that the parameter ε has great influence on the performance of HANDI algorithm. Most of data sets obtain high classification accuracies in a wide area. In particular, Wine, Wdbc, Credit, Gearbox, Segmentation and Leukemia exhibit stability in their

respective regions. These curves show the classification performance is stable and can provide a selection of an optimal subset of features. The optimal positions of classification accuracies are different among these datasets. Here, we recommend that ε should take values in the interval [0.1, 0.6].

Table 6. Optimal features selected by NEIEN, FINEN and HANDI algorithms

Data sets	NEIEN	FINEN	HANDI
Wine	12, 13, 1, 11, 5, 2, 10, 4, 3, 7	12, 13, 1, 10, 7, 2, 11, 4, 6, 8, 3, 9	7, 1, 11, 13, 5, 2, 10, 3
Wdbc	28, 21, 22, 8, 29, 13, 16, 10, 7, 27, 25, 1	28, 21, 22, 8, 29, 13, 16, 10, 7, 27, 25, 12	8, 21, 22, 12, 26, 28, 2, 25, 27, 9, 10
Wpbc	1, 24, 32, 5, 12	1, 13, 24, 16, 12, 32	24, 1, 32, 13, 5
Sonar	11, 45, 36, 17, 28, 54, 24, 41, 21, 32, 12, 26, 30, 15, 53, 42, 37, 20, 10, 23, 18, 48, 6, 39, 33, 50, 29, 40, 57	11, 45, 36, 9, 19, 1, 60, 46, 35, 22, 57, 12, 48, 37, 18, 26, 28, 27, 5, 32, 53, 29, 58, 59, 40, 10	12, 45, 20, 35, 22, 9, 21, 48, 37, 19, 18, 3, 60, 36, 8, 26, 29, 32, 6, 2, 17, 31
Credit	9, 10, 13, 15	9, 10, 13, 6, 12, 1, 5, 7	9, 10, 13, 4
Sick	20, 19, 26, 29, 24, 18, 2, 1, 3, 6, 10, 22, 17	20, 19, 26, 29, 24, 18, 2, 1, 3, 6, 10, 22, 17	29, 20, 26, 19, 6, 24, 2, 10
Gearbox	65, 56, 11, 48, 2, 53, 38, 8, 29, 44, 17	35, 20, 56, 47, 2, 11, 38, 62, 53, 26, 65	35, 20, 47, 2, 11, 56, 17, 38, 65
Segmentation	18, 11, 17, 2, 5, 12, 13, 7, 6	18, 11, 17, 2, 5, 12, 13, 7, 6	11, 2, 17, 13, 18, 1, 5, 7, 15
DLBCL	4767, 453, 2930, 5283, 3574	4767, 3257, 3127, 453, 1698, 1570	4767, 453, 4951, 1939, 1185
Leukemia	2833, 6720, 5555, 10127, 10038, 3479, 8964, 515	2833, 6720, 5555, 10127, 10038, 4839, 8952, 9053	2833, 6720, 5555, 788, 10127, 153
MLL	3634, 7754, 6565, 11395, 11297, 5265, 9121, 6410	3634, 7754, 6565, 11395, 11297, 5265, 4383, 8815, 8937, 145	3634, 7754, 6565, 5265, 1119, 6580, 1002
Prostate	8850, 4483, 6185, 6627, 8623, 9587, 12067, 4847	8850, 12067, 6185, 8623, 8129, 4483, 10753, 9850	4173, 6185, 4690
Tumors	5411, 6320, 7648, 3264, 3324, 6671, 4300, 6079, 6764, 10126, 8397, 8383, 9046, 7944, 10865, 8687, 2132	2543, 7648, 3264, 6320, 5411, 6671, 8548, 7781, 10126, 6764, 4178, 4448, 8337, 3043, 4831, 3880	2543, 6684, 6671, 2943, 3264, 7241, 7106, 5411, 10750, 11204, 12369, 4448, 4178, 7299, 3147, 3043

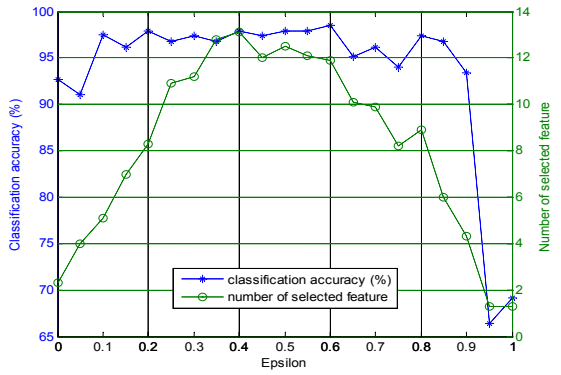

Fig.2 Numbers of selected features and accuracy varying with neighborhood radius ε (Wine)

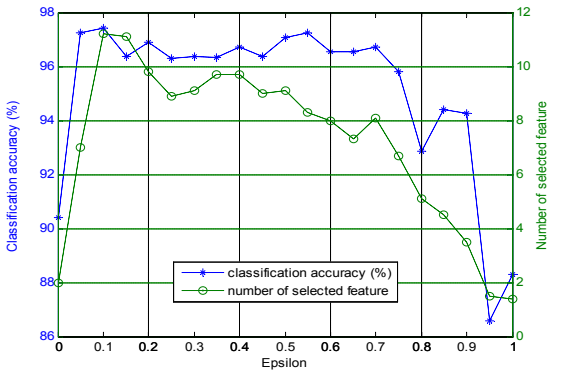

Fig.3 Numbers of selected features and accuracy varying with neighborhood radius ε (Wdbc)

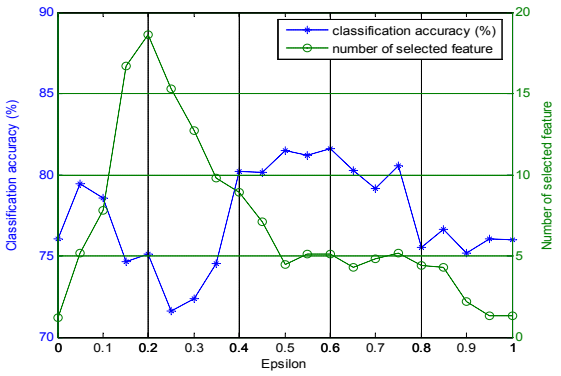

Fig.4 Numbers of selected features and accuracy varying with neighborhood radius ε (Wpbc)

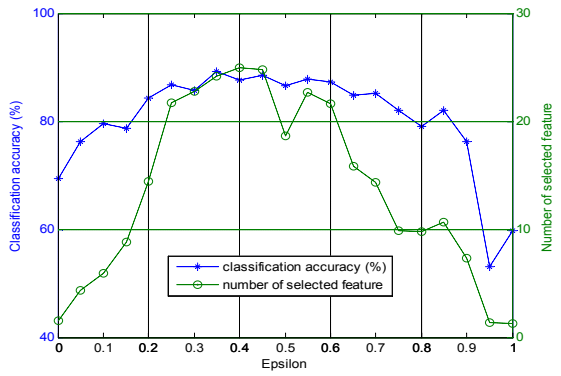

Fig.5 Numbers of selected features and accuracy varying with neighborhood radius ε (Sonar)

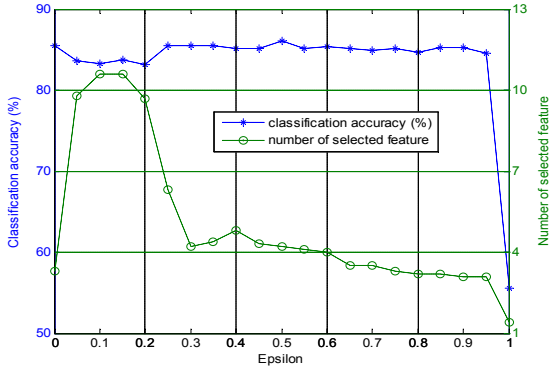

Fig.6 Numbers of selected features and accuracy varying with neighborhood radius ε (Credit)

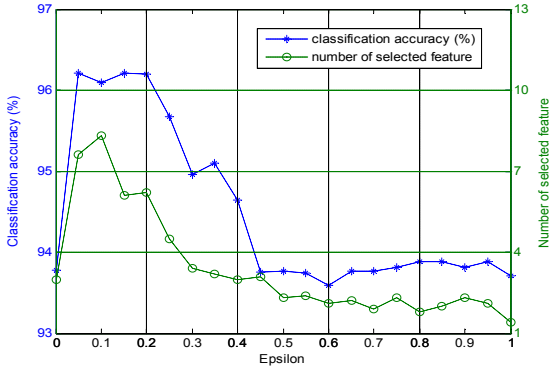

Fig.7 Numbers of selected features and accuracy varying with neighborhood radius ε (Sick)

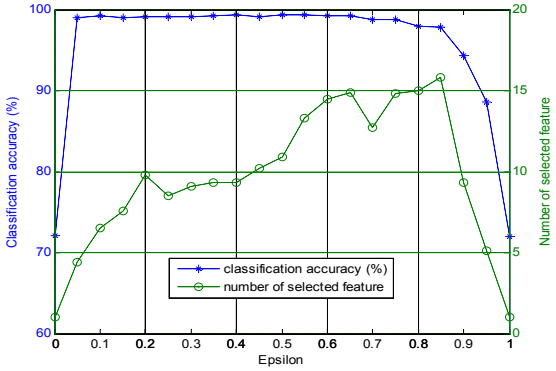

Fig.8 Numbers of selected features and accuracy varying with neighborhood radius ε (Gearbox)

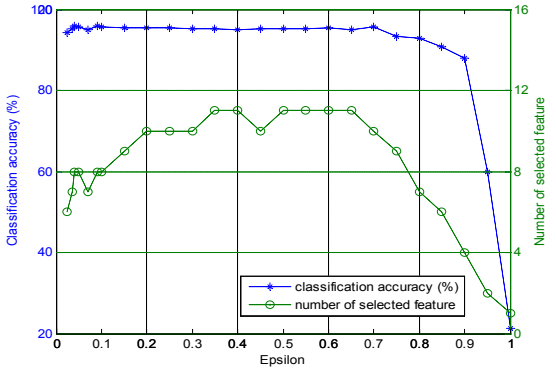

Fig.9 Numbers of selected features and accuracy varying with neighborhood radius ε (Segmentation)

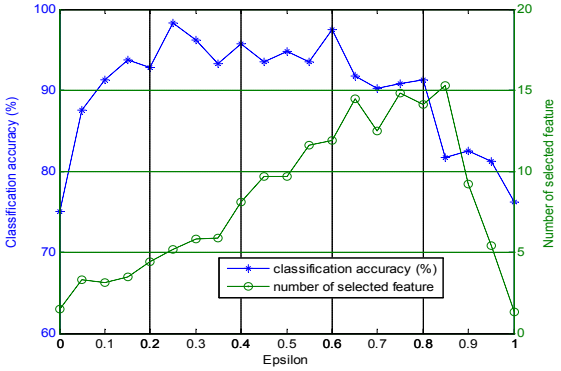

Fig.10 Numbers of selected features and accuracy varying with neighborhood radius ε (DLBCL)

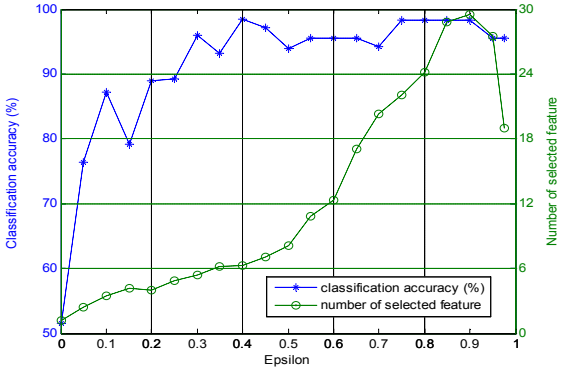

Fig.11 Numbers of selected features and accuracy varying with neighborhood radius ε (Leukemia)

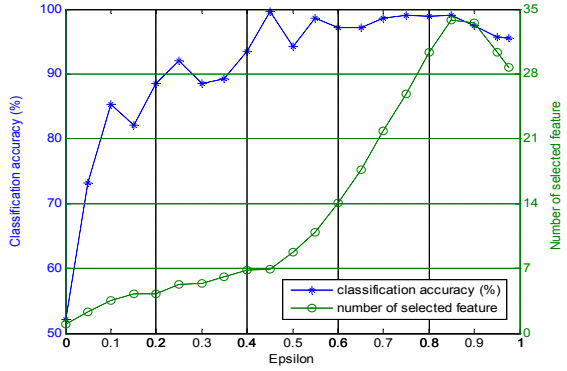

Fig.12 Numbers of selected features and accuracy varying with neighborhood radius ε (MLL)

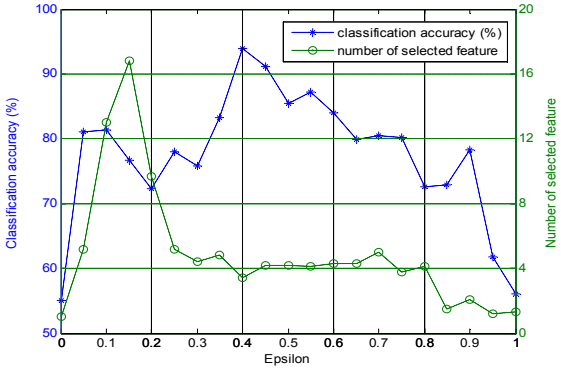

Fig.13 Numbers of selected features and accuracy varying with neighborhood radius ε (Prostate)

VI. Conclusion

Measures for computing the distinguishment capacity of a subset of features play an important role in classification learning and feature selection. A number of measures were developed for these tasks. Considering its effectiveness, information entropy is widely used and discussed for evaluating features. In this paper, we introduce some basic ideas in Shannon information theory into neighborhood relation context and propose some discrimination indexes to measure the distinguishment ability of a subset of features. The proposed discrimination indexes are directly defined on a neighborhood relation and computed by considering the cardinality of neighborhood relation rather than neighborhood similarity classes. The conditional discrimination index is used to measure the increment of discrimination information caused by adding a new feature, which is interpreted as the significance of an attribute. Based on the proposed discrimination measures, we put forward a new algorithm of feature selection. With thirteen public data sets, a series of experiments are conducted for evaluating the proposed method. The results show that the algorithm can select fewer features and keep higher classification accuracy and spend less time. What is more, most of classification accuracies are improved. We also find that different parameters have an impact on the performance of the feature selection algorithm. We should select the suitable value of threshold for each data set according to the curves of data sets.

References

[1] R. Battiti, "Using mutual information for selecting features in supervised neural net learning," IEEE Trans. Neural Netw., vol. 5, no. 4, pp. 537–550, 1994.

[2] T. Beaubouef, F.E. Petry, G. Arora, "Information-theoretic measures of uncertainty for rough sets and rough relational databases," Inf. Sci., vol.109, pp. 185–195, 1998.

[3] C.L. Blake and C.J. Merz, "UCI Repository of Machine Learning Databases," 1998. [Online]. Available: http://www. ics.uci.edu/~mlearn/ MLR- epository. html

[4] D. Chen, L. Zhang, S. Zhao, Q. Hu, P. Zhu, "A novel algorithm for finding reducts with fuzzy rough sets," IEEE Trans. on Fuzzy Syst., vol. 20, no.2, pp. 385-389, 2012.

[5] M. Dash and H. Liu, "Consistency-based search in feature selection," Artif. Intell., vol. 151, no. 1/2, pp. 155–176, 2003.

[6] J. Dai, W. Wang, Q. Xu, "An uncertainty measure for incomplete decision tables and its applications," IEEE Trans. Cybern., vol. 43, no.4, pp. 1277-1289, 2013.

[7] I. Duentsch, G. Gediga, "Uncertainty measures of rough set prediction," Artif. Intell., vol. 106, pp. 109–137, 1998.

[8] R. Gilad-Bachrachy, A. Navotz, and N. Tishbyy, "Margin-based feature selection: Theory and algorithms," in Proc. 21st Int. Conf. Mach. Learn. Banff, AB, Canada, pp. 43-50, 2004.

[9] S. Greco, B. Matarazzo, and R. Slowinski, "Rough approximation by dominance relations," Int. J. Intell. Syst., vol.17, pp.153-171, 2002.

[10] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," J. Mach. Learn. Res., vol. 3, pp. 1157–1182, 2003.

[11] M. A. Hall, "Correlation-based feature selection for discrete and numeric class machine learning," in Proc. 17th Int. Conf. Mach. Learn., pp. 359–366, 2000.

[12] T. K. Ho and M. Basu, "Complexity measures of supervised classification problems," IEEE Trans. Pattern Anal. Mach. Intell., vol. 24, no. 3, pp. 289–300, 2002.

[13] Q. Hu, L. Zhang, D. Zhang, W. Pan, S. An ,W. Pedrycz, "Measuring relevance between discrete and continuous features based On neighborhood mutual information," Expert Syst. Appl., vol. 38, pp. 10737-10750, 2011.

[14] Q. Hu, D. Yu, Z. Xie, and J. Liu, "Fuzzy probabilistic approximation spaces and their information measures," IEEE Trans. Fuzzy Syst., vol. 14, no. 2, pp. 191–201, 2006.

[15] Q. Hu, D. Yu, J. Liu, and C. Wu, "Neighborhood-rough-set based heterogeneous feature subset selection," Inf. Sci., vol. 178, no. 18, 3577–3594, 2008.

[16] E. Hernandez, J. Recasens, "A reformulation of entropy in the presence of indistinguishability operators," Fuzzy Sets Syst. Vol. 128, 185–196, 2002.

[17] M.Inuiguchi, Y.Yoshioka, Y.Kusunoki, "Variable-precision dominance based rough set approach and attribute reduction," Int. J. Approx. Reason., vol. 50, no. pp. 1199–1214, 2009.

[18] R. Jensen and Q. Shen, "Semantics-preserving dimensionality reduction: Rough-and fuzzy-rough-based approaches," IEEE Trans. Knowl. Data Eng., vol. 16, no. 12, pp. 1457–1471, 2004.

[19] R. Jensen, Q. Shen, "Fuzzy-rough sets assisted attribute selection," IEEE Trans. Fuzzy Syst., vol. 15, no. 1, pp. 73–89, 2007.

[20] D. Kim, "Data classification based on tolerant rough set," Pattern Recognit., vol. 34, no. 8, pp. 1613–1624, 2001.

[21] N. Kwak and C.-H. Choi, "Input feature selection by mutual information based on Parzen window," IEEE Trans. Pattern Anal. Mach. Intell.,vol. 24, no. 12, pp. 1667–1671, 2002.

[22] N. Kwak and C. H. Choi, "Input feature selection for classification problems," IEEE Trans. Neural Netw., vol. 13, no. 1, pp. 143–159, 2002.

[23] C. Lee and D. A. Landgrebe, "Feature extraction based on decision boundaries," IEEE Trans. Pattern Anal. and Mach. Intell., vol. 15, no. 4, pp. 388–400, 1993.

[24] J. Liang, K. Chin, C. Dang, R. Yam, "A new method for measuring uncertainty and fuzziness in rough set theory," Int. J. Gen. Syst., vol. 31, 331–342, 2002.

[25] J. Liang, F. Wang, C. Dang, "A Group Incremental Approach to Feature Selection Applying Rough Set Technique," IEEE Trans. Knowl. Data Eng., vol. 26, no. 2, pp. 294-304, 2014.

[26] H. Liu and L. Yu, "Toward integrating feature selection algorithms for classification and clustering," IEEE Trans. Knowl. Data Eng., vol. 17, no. 4, pp. 491–502, 2005.

[27] J. S. Mi, Y. Leung, H. Y. Zhao, T. Feng, "Generalized fuzzy rough sets determined by a triangular norm," Inf. Sci., vol. 178, no.16, pp. 3203-3213, 2008.

[28] J. Mi, Y. Leung, W. Wu, "An uncertainty measure in partition-based fuzzy rough sets," Int. J. Gen. Syst. vol. 34, no. 1, pp. 77–90, 2005.

[29] J. Mi, W. Wu, and W. Zhang, "Approaches to knowledge reduction based on variable precision rough sets model," Inf. Sci., vol. 159, no.3/4, pp. 255-272, 2004.

[30] P. Mitra, C.A. Murthy, and S.K. Pal, "Unsupervised feature selection using feature similarity," IEEE Trans. Pattern Anal. Mach. Intell., vol.24, no. 3, pp.301-312, Mar. 2002

[31] D. Muni, N. Pal, and J. Das, "Genetic programming for simultaneous feature selection and classifier design," IEEE Trans. Syst., Man, Cybern. B, Cybern., vol. 36, no. 1, pp. 106–117, 2006.

[32] F. Nie, H. Huang, X. Cai, Chris H. Q. Ding, Efficient and robust feature selection via joint $\ell_{2,1}$-norms minimization, in Proc. Advances in Neural Inf. Process. Syst., NIPS 2010, 2010, pp:1813-1821.

[33] I. Oh, J. Lee, and B. Moon, "Hybrid genetic algorithms for feature selection," IEEE Trans. Pattern Anal. Mach. Intell., vol. 26, no. 11, pp. 1424–1437, 2004.

[34] Z. Pawlak, "Rough Sets: Theoretical Aspects of Reasoning About Data," Dordrecht, The Netherlands: Kluwer, 1991.

[35] N. Parthalain, Q. Shen and R. Jensen, &ldquo, "A Distance measure approach to exploring the rough set boundary region for attribute reduction," IEEE Trans. Knowl. Data Eng., vol. 22, no. 3, pp. 305-317, 2010.

[36] H. Peng, F. Long, and C. Ding, "Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy," IEEE Trans. Pattern Anal. Mach. Intell., vol. 27, no. 8, pp. 1226–1238, 2005.

[37] Y. Qian, J. Liang, "Combination entropy and combination granulation in rough set theory," Fuzzin. Knowl. Syst., vol. 16 pp. 179–193, 2008.

[38] Y. Qian, J. Liang, W. Pedrycz and C. Dang, "Positive approximation: an accelerator for attribute reduction in rough set theory," Artif. Intell., vol. 174, pp. 597-618, 2010.

[39] R. Slowinski, D. vanderpooten, "A generalized definition of rough approximations based on similarity," IEEE Trans. Knowl. Data Eng., vol. 2, pp. 331-336, 2000.

[40] Y. She, X. He, “Uncertainty measures in rough algebra with applications to rough logic,” Int. J. Mach. Learn. Cybern., vol. 5, no.5, pp. 671-681, 2014.
[41] Y. Sun, S. Todorovic, and S. Goodison, “Local learning based feature selection for high-dimensional data analysis,” IEEE Trans. Pattern Anal. Mach. Intell.,vol. 32, no.9, pp.1610-1626, Sep. 2010
[42] H. Tao, C. Hou, F. Nie, Y. Jiao, D. Yi, "Effective discriminative feature selection with non-trivial solutions,"IEEE Trans. Neural Netw. Learning Syst. vol. 27, no. 4, pp. 796-808, 2016.
[43] K. Torkkola, “Feature extraction by nonparametric mutual information maximization,” J. Mach. Learn. Res., vol. 3, no. 7/8, pp. 1415–1438, 2003.
[44] X. Wang, L. Dong, J. Yan, “Maximum ambiguity based sample selection in fuzzy decision tree induction,” IEEE Trans. Knowl. Data Eng., vol. 24, no.8, pp. 1491–1505, 2012
[45] C. Wang, Q. He, D.g Chen, Q. Hu, “A novel method for attribute reduction of covering decision systems,” Inf. Sci., pp. 254:181–196, 2014.
[46] C. Wang, Y. Qi, M. Shao, Q. Hu, D. Chen, Y. Qian, Y. Lin, "A fitting model for feature selection with fuzzy rough sets," IEEE Trans. Fuzzy Syst., 10.1109/TFUZZ.2016.2574918
[47] H. Wang, “Nearest neighbors by neighborhood counting,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 28, no. 6, pp. 942–953, 2006.
[48] D. Wang, F. Nie, H. Huang, "Feature selection via global redundancy minimization," IEEE Trans. Knowl. Data Eng. vol. 27, no.10, pp. 2743-2755, 2015.
[49] W. Wu, W. Zhang, “Neighborhood operator systems and approximation,” Inf. Sci., vol. 144, no. 1–4, pp. 201–217, 2002.
[50] W. Wu, "Knowledge reduction in random incomplete decision tables via evidence theory," Fundam. Inform., vol. 115, no.2-3, pp. 203-218, 2012.
[51] W. Xu X. Zhang, W. Zhang, “Knowledge granulation, knowledge entropy and knowledge uncertainty measure in ordered information systems,” Appl. Soft Comput., vol. 9, no.4, pp. 1244-1251, 2009.
[52] S. Xiang, F. Nie, G. Meng, C. Pan, C. Zhang, "Discriminative least squares regression for multiclass classification and feature selection," IEEE Trans. Neural Netw. Learn. Syst. vol. 23, no. 11, pp. 1738-1754, 2012.
[53] Y. Yao, “Relational interpretations of neighborhood operators and rough set approximation operators,” Inf. Sci., vol. 111, no. 1-4, pp. 239– 259, 1998.
[54] Y. Yao, Y.H. She, "Rough set models in multigranulation spaces", Inf. Sc., vol. 327, pp. 40-56, 2016.
[55] R. R. Yager, “Entropy measures under similarity relations,” Int. J. Gen. Syst, vol. 20, pp. 341–358, 1992.
[56] S. Zhao, H. Chen, C. P. Li, M. Y. Zhai, X. Y. Du, "RFRR: Robust Fuzzy Rough Reduction," IEEE Trans. Fuzzy Syst., vol. 21, no. 5, pp. 825-841, 2013.
[57] S. Zhao, C.C. Tsang, D. Chen, "Building a rule-based classifier by using fuzzy rough set technique," IEEE Trans. Knowl. Data Eng., vol. 22, no. 5, pp. 624–638, 2010.
[58] S. Zhao, C.C. Tsang, "On Fuzzy approximation Operators in Attribute Reduction with Fuzzy Rough Sets," Inf. Sci., vol. 178, no. 16, pp.3162-3176, 2008.
[59] X. Zhang, B. Zhou, P. Li, “A general frame for intuitionistic fuzzy rough sets,” Inf. Sci., vol. 216, no. 2012, pp. 34-49.
[60] P. Zhu, Q. H. Hu, “Adaptive neighborhood granularity selection and combination based on margin distribution optimation,” Inf. Sci., vol. 249, pp.1-12, 2013.
[61] S. Ho and H. Wechsler, "Query by transduction," IEEE Trans. Pattern Anal. Mach. Intell., vol. 30, no. 9, pp. 1557–1571, 2008.
[62] Kent Ridge Bio-medical Dataset, http://datam.i2r.a-tar.edu.sg/datasets/krbd/index.html.

Changzhong Wang received the M.S. degree from Bohai University, Jinzhou, China, the Ph.D. degree from Harbin Institute of Technology, Harbin, China, in 2005, and 2008 respectively. He is currently an Professor with Bohai University.

His research interests are focused on fuzzy sets, rough sets, data mining, pattern recognition and statistical analysis. He has authored or coauthored more than 40 journal and conference papers in the areas of machine learning, data mining, and rough set theory.

Qinghua Hu received the B. Eng. and M. Eng. degrees in power engineering, and the Ph.D. degree in control science and engineering from Harbin Institute of Technology, Harbin, China, in 1999, 2002, and 2008, respectively. He is currently an Professor with Tianjin University.

His research interests include data mining and knowledge discovery with fuzzy and rough techniques. He is the author or a coauthor of more than 60 journal papers and conference proceedings in machine learning, data mining.

Xizhao Wang is Ph.D., Professor, IEEE Fellow and Editor-in-Chief of Springer Journal Machine Learning and Cybernetics. He received the Ph.D. degree in computer science from the Harbin Institute of Technology, Harbin, China, in 1998. From September 1998 to September 2001, he served as a Research Fellow in the Department of Computing, Hong Kong Polytechnic University, Hong Kong. He became a Full Professor and Dean of the College of Mathematics and Computer Science, Hebei University, in October 2001. Since March 2014 to now Prof. Wang has moved to college of computer science and software engineering in Shenzhen University as a professor and the vice director of Big Data Institute.

He has over 160 publications, including four books, seven book chapters, and over 90 journal papers in IEEE Transactions on PAMI/SMC/FS, Fuzzy Sets and Systems, Pattern Recognition, etc. His H-index is 18 (up to April 2013). His current research interests include learning from examples with fuzzy representation, fuzzy measures and integrals, neuro-fuzzy systems and genetic algorithms, feature extraction, multiclassifier fusion, and applications of machine learning.

Degang Chen received the M.S. degree from Northeast Normal University, Changchun, China, in 1994 and the Ph.D. degree from the Harbin Institute of Technology, Harbin, China, in 2000. He was a Postdoctoral Fellow with Xi’an Jiaotong University, Xi’an, China, from 2000 to 2002 and with Tsinghua University, Beijing, China, from 2002 to 2004. Since 2006, he has been a Professor with North China Electric Power University, Beijing.

His research interests include fuzzy groups, fuzzy analysis, rough sets, and support vector machines.

Yuhua Qian is a Professor of Key Laboratory of Computational Intelligence and Chinese Information Processing of Ministry of Education, China. He received the M.S. degree and the PhD degree in Computers with applications at Shanxi University in 2005 and 2011, respectively.

He is actively pursuing research in pattern recognition, feature selection, rough set theory, granular computing and artificial intelligence. He has published more than 50 articles on these topics in international journals.

Zhe Dong received his B.Sc. degrees in mathematics from Bohai University in 2012. Now, she is a graduate student for a Master's degree. Her main research interests include fuzzy sets, rough sets, pattern recognition and knowledge discovery.