Feature Selection Based on Neighborhood Discrimination Index

Changzhong Wang, Qinghua Hu, Xizhao Wang, Degang Chen, Yuhua Qian, and Zhe Dong

***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 distinguishing information of a neighborhood relation. It reflects the distinguishing ability of a feature subset. The proposed discrimination index is computed by considering the cardinality of a neighborhood relation rather than neighborhood similarity classes. Variants of the discrimination index, including joint discrimination index, conditional discrimination index, and mutual discrimination index, are introduced to compute the change of distinguishing information caused by the 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 address the 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. Data sets selected from public data sources are used to compare the proposed algorithm with existing algorithms. The experimental results confirm that the discrimination index-based algorithm yields superior performance compared to other classical algorithms.**

***Index Terms*— Discrimination index, distinguishing information, feature selection, neighborhood relation.**

I. Introduction

With the development of computer and database technology, the amount of data is growing exponentially. Ideally, the information provided is useful; however, data frequently contains redundant information. Therefore, before using a data set, it is necessary to preprocess the data to remove redundant features. Feature selection is an important tool to reduce redundant features. The majority of researchers are committed to processing high-dimensional data with feature selection. The aim is to determine a subset of optimal features with strong classification ability according to 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 the data. Recently, this has attracted considerable attention in pattern recognition, machine learning, and data mining [5]–[28], [31]–[36], [44]–[50], [55].

Relations, produced by a subset of features, represent the similarity or dissimilarity between samples. Similar samples form a similarity class, dissimilar samples 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 [37], [47], clustering [41], attribute reduction [4], [6], [19]–[21], [53], and uncertainty reasoning and decision [29], [43], [62]. Furthermore, equivalence relations [31], [37], [57], similarity relations [22], [42], [59]–[61], neighborhood relations [15], [38], [51]–[53], [56], and dominance relations [9], [19], [54] are the foundations of a sequence of rough set models.

Entropy, as an uncertainty measure, is a useful tool for characterizing the distinguishing information of a subset of features. The less likely the conditional entropy a feature subset has with respect to decision attribute, the greater the capability the feature subset has in distinguishing samples with different decisions. Entropy has played an important role in pattern recognition and feature selection. Since Shannon first proposed information entropy to evaluate the uncertainty of discrete sample spaces, entropy has been applied in diverse fields [2], [6], [7], [16]–[18]. The extension of entropy and its variants were adapted for feature selections in [1], [16], [23], and [39]. To calculate the distinguishing information of fuzzy or numerical features, Yager [58] introduced the concept of entropy into fuzzy similarity relations. In fact, Yager's entropy is a generalization of Shannon entropy; it is defined using equivalence classes or fuzzy similarity classes. In 2002, Hernandez and Recasens [18] extended Yager's work and presented the formulae of joint entropy and conditional entropy based on Yager's entropy then, they used these measures to learn fuzzy decision trees from a set of data samples. Hu *et al.* [17] redefined joint entropy and conditional entropy based on Yager's work and used them to measure

Manuscript received January 27, 2016; revised July 25, 2016 and March 4, 2017; accepted May 24, 2017. Date of publication June 23, 2017; date of current version June 21, 2018. This work was supported in part by the National Natural Science Foundation of China under Grant 61572082, Grant 61673396, Grant 61363056, and Grant 61473111, in part by the Foundation of Educational Committee of Liaoning Province under Grant LZ2016003, and in part by the Natural Science Foundation of Liaoning Province under Grant 2014020142. *(Corresponding author: Changzhong Wang.)*

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

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

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

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

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

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TNNLS.2017.2710422

the uncertainty of the distinguishing ability of a set of fuzzy similarity relations. In 2005, Mi *et al.* [30] introduced a distinguishable measure of fuzzy equivalence relation based on a fuzzy rough set model. In 2008, Qian and Liang [40] proposed a combinational measure for evaluating the uncertainty of the distinguishing ability of a subset of features. In 2011, Hu *et al.* [16] 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 All these studies focused on extensions of Shannon entropy or Yager's entropy and their applications.

Neighborhood is one of the most important concepts in classification learning [15], [51], [52], [63]. Neighborhood can be used to generate similarity classes from the samples described by numerical features and used to distinguish samples. The distinguishing information of a feature subset is related to the neighborhood relations induced by the feature subset. In this paper, we propose a new measure of distinguishing information, called the neighborhood discrimination index, based on neighborhood relations. Compared to Yager's entropy [58] and its varieties [16], [17], the neighborhood discrimination index has similar properties to Shannon entropy. However, it is directly defined on neighborhood relations and acquired by computing the cardinality of the neighborhood relations rather than neighborhood similarity classes. Thus, the computational complexity of the proposed discrimination index is less. We define joint discrimination index, conditional discrimination index, and mutual discrimination index and discuss their basic properties. These measures are used to calculate the change of distinguishing information caused by the combination of multiple feature subsets. As with Shannon conditional entropy, the conditional discrimination index can be used to characterize the ability of a subset of features to distinguish samples with different decisions; the smaller the conditional discrimination index, the greater the distinguishing ability of the feature subset. We also discuss the influence of the neighborhood radius on the neighborhood discrimination index. Then, we define attribute importance and propose a feature selection algorithm based on the proposed discrimination measures. Finally, we use public standard data sets to verify the validity and stability of the proposed method and compare the proposed algorithm with existing methods. The experimental results confirm that the proposed measures are efficient and effective for feature selections.

This paper is organized as follows. In Section II, we review the basic concepts of Shannon entropy in learning. In Section III, we present the definitions of the 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 for feature selection based on a 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 Learning

Suppose that U is a nonempty set of samples, A is a set of discrete attributes describing the 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:

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

Suppose that the partition produced by R_B is denoted by $U/B = \{X_1, X_2, \ldots X_m\}$, 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 $\{X_1, X_2, \ldots X_m\}$, 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, $i = 1, 2, \ldots, m$.

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 $U/C = \{Y_1, Y_2, \ldots, Y_n\}$, 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

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

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

$H(B|C)$ 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 $p(X_i|Y_j) = 1$ for any $Y_j \in U/C$, then $H(B|C) = 0$. This means that the distinguishing ability of the attribute subset B is completely contained in C.

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. Further, we know that mutual information has the following properties:

$$\begin{aligned} &I(B; C) = I(C; B) \\ &I(B; C) = H(B) + H(C) - H(B \cup C) \\ &I(B; C) = H(B) - H(B|C) = H(C) - H(C|B). \end{aligned} \quad (7)$$

We consider the decision attribute D as a random variable on U and suppose the value space for D is $\{\omega_1, \omega_2, \ldots, \omega_r\}$, 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 smaller the $H(D|B)$, the greater the distinguishing ability of B. When the attribute subset B completely divides all samples into their respective categories, then $H(D|B) = 0$. According to the relationship between the conditional entropy and mutual information, we can clearly know that the mutual information increases with an increase of the distinguishing 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 $\langle U, A, D\rangle$, where $U = \{x_1, x_2, \ldots, x_n\}$ is a nonempty set of samples, called a universe, $A = \{a_1, a_2, \ldots, a_m\}$ 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 $U/D = \{D_1, D_2, \ldots D_r\}$. 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 distinguishing ability of a feature subset. We begin by 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 satisfies

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 $r_{ij} \in \{0, 1\}$, $i, j = 1, 2, \ldots, n$. There are many methods to calculate r_{ij}; 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 $x_l = [x_{l1}, x_{l2}, \ldots, x_{ls}]^T$, $l = i, j$ are two samples, T represents the transpose operation of a vector, B is a subset of attributes with $|B| = s$ and

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

In this case, $\|\cdot\|$ represents the absolute value. Δ_p^B is called the 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 the 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 the 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, the greater its distinguishing ability. There are two factors that influence a neighborhood similarity relation. One is the neighborhood radius ε, the other is the feature subset B. For a given parameter ε, the neighborhood relation becomes finer as the number of features in B increases. This property can be formulated as follows.

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

A neighborhood similarity relation characterizes the distinguishing ability of a feature subset. Property 1 demonstrates that the greater the number of features, the finer the neighborhood relation and the greater the distinguishing ability of the feature subset.

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

Definition 1: Given a decision table $\langle U, A, D\rangle$, where $U = \{x_1, x_2, \ldots, x_n\}$, $B \subseteq A$, 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 clearly 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 the distinguishing 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 nonnegative real numbers. With this mapping, the distinguishing abilities of different feature subsets can be compared.

Compared with neighborhood entropy [16], the neighborhood discrimination index has two main differences.

1) The concept of neighborhood discrimination index is based on neighborhood relations. It can be directly obtained by computing the cardinality of the neighborhood relations, whereas neighborhood entropy is defined on the neighborhood similarity classes and accumulatively obtained by considering the cardinality of the similarity classes. Thus, the computational complexity of the neighborhood discrimination index is somewhat less than the 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. Hence, neighborhood entropy is a generalization of Shannon entropy, whereas the neighborhood discrimination index is simply a measure of the distinguishing ability of a feature subset. This is the essential difference between these measures.

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

Proposition 1: If $\varepsilon_1 \leq \varepsilon_2$, then $H^{\varepsilon_1}(B) \geq H^{\varepsilon_2}(B)$.

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

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

Proposition 2: If $B_1 \subseteq B_2$, then $H^{\varepsilon}(B_1) \leq H^{\varepsilon}(B_2)$.

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

Proposition 2 demonstrates that the neighborhood discrimination index is influenced by the number of features. It increases monotonously with the size of the 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

$$H^{\varepsilon}(B_1, B_2) = \log \frac{n^2}{\left|R_{B_1}^{\varepsilon} \cap R_{B_2}^{\varepsilon}\right|}. \tag{12}$$

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

Proposition 3: $H^{\varepsilon}(B_1, B_2) \geq H^{\varepsilon}(B_1)$, $H^{\varepsilon}(B_1, B_2) \geq H^{\varepsilon}(B_2)$.

It is clear that the joint discrimination index of B_1 and B_2 is greater than any individual discrimination index. This is interpreted as meaning the distinguishing ability of the joint features strengthens with the addition of new features. This is because we can obtain 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 demonstrates that the addition of new features does not increment the discrimination index if these features are contained in other existing features. In this case, the distinguishing 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

$$H^{\varepsilon}(B_1|B_2) = \log \frac{\left|R_{B_2}^{\varepsilon}\right|}{\left|R_{B_1}^{\varepsilon} \cap R_{B_2}^{\varepsilon}\right|}. \tag{13}$$

Because $|R_{B_1}^{\varepsilon} \cap R_{B_2}^{\varepsilon}| \leq |R_{B_2}^{\varepsilon}|$, it is clearly seen that $H^{\varepsilon}(B_1|B_2) \geq 0$. When $B_1 \subseteq B_2$, then $R_{B_1}^{\varepsilon} \supseteq R_{B_2}^{\varepsilon}$. This means $H^{\varepsilon}(B_1|B_2) = 0$. When $|R_{B_2}^{\varepsilon}| = n^2$ and $R_{B_1}^{\varepsilon}$ is an identity matrix, the conditional discrimination index attains the maximum value. That is, $H^{\varepsilon}(B_1|B_2) = \log n$.

According to the above discussion, we obtain the following property.

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

1) $H^{\varepsilon}(B_1 \cup B_2) \geq \max\{H^{\varepsilon}(B_1), H^{\varepsilon}(B_2)\}$.
2) $H^{\varepsilon}(B_1|B_2) = 0$ 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 indicates that feature subset B_1 will not introduce distinguishing 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

$$H^{\varepsilon}(B_1|B_2) = H^{\varepsilon}(B_1, B_2) - H^{\varepsilon}(B_2). \tag{14}$$

Proof:

$$\begin{aligned} H^{\varepsilon}(B_1, B_2) - H^{\varepsilon}(B_2) &= \log \frac{n^2}{\left|R_{B_1}^{\varepsilon} \cap R_{B_2}^{\varepsilon}\right|} - \log \frac{n^2}{\left|R_{B_2}^{\varepsilon}\right|} \\ &= \log \frac{n^2}{\left|R_{B_1}^{\varepsilon} \cap R_{B_2}^{\varepsilon}\right|} \cdot \frac{\left|R_{B_2}^{\varepsilon}\right|}{n^2} \\ &= \log \frac{\left|R_{B_2}^{\varepsilon}\right|}{\left|R_{B_1}^{\varepsilon} \cap R_{B_2}^{\varepsilon}\right|}. \end{aligned}$$

It is clearly observed that the conditional discrimination index is the increment of the distinguishing information by introducing a new feature subset after one feature subset is known. This reflects the increment of the distinguishing ability with 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 $R_{B_1}^{\varepsilon}$, $R_{B_2}^{\varepsilon}$ 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

$$I^{\varepsilon}(B_1; B_2) = \log \frac{n^2\left|R_{B_1}^{\varepsilon} \cap R_{B_2}^{\varepsilon}\right|}{\left|R_{B_1}^{\varepsilon}\right| \cdot \left|R_{B_2}^{\varepsilon}\right|}. \tag{15}$$

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

$$\begin{aligned} &I^{\varepsilon}(B_1; B_2) = I^{\varepsilon}(B_2; B_1) \\ &I^{\varepsilon}(B_1; B_2) = H^{\varepsilon}(B_1) + H^{\varepsilon}(B_2) - H^{\varepsilon}(B_1, B_2) \\ &I^{\varepsilon}(B_1; B_2) = H^{\varepsilon}(B_1) - H^{\varepsilon}(B_1|B_2) = H^{\varepsilon}(B_2) - H^{\varepsilon}(B_2|B_1). \end{aligned} \tag{16}$$

Proof:

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

$$\begin{aligned} &= \log \frac{|n^2|}{\left|R_{B_1}^{\varepsilon}\right|} + \log \frac{|n^2|}{\left|R_{B_2}^{\varepsilon}\right|} - \log \frac{|n^2|}{\left|R_{B_1}^{\varepsilon} \cap R_{B_2}^{\varepsilon}\right|} \\ &= \log \frac{n^2 \cdot \left|R_{B_1}^{\varepsilon} \cap R_{B_2}^{\varepsilon}\right|}{\left|R_{B_1}^{\varepsilon}\right| \cdot \left|R_{B_2}^{\varepsilon}\right|} = I^{\varepsilon}(B_1; B_2). \end{aligned}$$

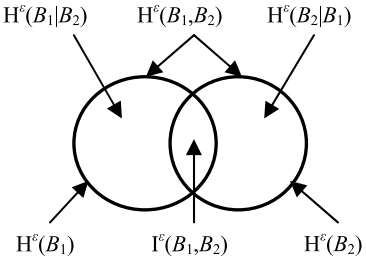


Fig. 1. Relationship diagram of discrimination indexes.

3) $H^{\varepsilon}(B_1) - H^{\varepsilon}(B_1|B_2)$

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

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

The first item indicates that the mutual discrimination index of B_1 and B_2 is symmetric. The second states 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 demonstrates that the mutual discrimination index is the difference between the discrimination index of one of the two feature subsets and their conditional discrimination index. It reflects that the mutual discrimination index is the common part of the distinguishing information of the two feature subsets. The relationship between neighborhood, conditional, and mutual discrimination indexes can be explained in Fig. 1.

Remark 2: Given a decision table $\langle U, A, D\rangle$, $B \subseteq A$ and neighborhood radius ε, R^{ε}_B is a neighborhood relation induced by B and ε, and R_D is an equivalence relation induced by D. Similar to Shannon conditional entropy, $H(D|B)$ can be used to characterize the ability of B to distinguish samples with different decisions. The smaller the value of $H(D|B)$, the greater the distinguishing ability of B. 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 conclude that the mutual discrimination index increases as the distinguishing 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 assign a class label to a sample according to other samples' labels in its neighborhood. If all samples in the neighborhood have the same label, then the sample is called consistent; otherwise, the sample is inconsistent. Let ε be a neighborhood radius, if all samples are consistent, then $\langle U, A, D\rangle$ is called consistent; otherwise, it is called inconsistent. It is clearly seen that $\langle U, A, D\rangle$ 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) $I^{\varepsilon}(D; B) = H^{\varepsilon}(D)$.

TABLE I
DESCRIPTION OF DATA SETS

No.	Dataset	Sample	Attribute	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

The first item reflects that the conditional discrimination index equals zero 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 states that the mutual discrimination index between B and D is equal to the distinguishing information quantity of D if the classification is consistent.

As we know, Shannon mutual information is widely used in the 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 the 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 this idea, we present an axiomatic approach to feature selection as follows.

Axiom 1 (Maximum of Classification Information): Given a decision table $\langle U, A, D\rangle$, 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 $\langle U, A, D\rangle$, $\mathbb{N}$ 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 multigranular method to describe the classification ability of a set of numerical features if neighborhood radius ε is considered as a variable.

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

Definition 5: Given a decision table $\langle U, A, D\rangle$, 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) $I^{\varepsilon}(D; B - \{a\}) < I^{\varepsilon}(D; B)$, $\forall a \in B$.

TABLE II
AVERAGE SIZE OF FEATURE SUBSETS SELECTED WITH TENFOLD CROSS VALIDATION

Dataset	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 III
COMPARISON OF CLASSIFICATION ACCURACY OF REDUCED DATA WITH SVM (%)

Dataset	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

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

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

1) $H^{\varepsilon}(D|B) \leq H^{\varepsilon}(D|A)$.
2) $H^{\varepsilon}(D|B-\{a\}) > H^{\varepsilon}(D|B) \forall a \in B$.

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

$$R_1 = \begin{bmatrix} 1 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 1 \end{bmatrix} \quad R_2 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 1 \\ 0 & 1 & 1 \end{bmatrix}$$

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

We have

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

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

Suppose the decision equivalence relation

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

TABLE IV
COMPARISON OF CLASSIFICATION ACCURACY OF REDUCED DATA WITH 3NN (%)

Dataset	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 V
RUNNING TIME OF REDUCTION WITH DIFFERENT ALGORITHMS (s)

Dataset	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

We compute

$$H(R_1) = \log\frac{9}{5} = 0.8480, \quad H(R_2) = \log\frac{9}{3} = 1.5850$$
$$H(R_3) = \log\frac{9}{7} = 0.3626, \quad H(R_1R_2) = \log\frac{9}{3} = 1.5850$$
$$H(R_2R_3) = \log\frac{9}{3} = 1.5850, \quad H(R_1R_3) = \log\frac{9}{5} = 0.8480.$$

According to Proposition 6, we know

$$H(R_1|R_2) = H(R_1R_2) - H(R_2) = \log\frac{9}{3} - \log\frac{9}{3} = 0$$
$$H(R_2|R_1) = H(R_1R_2) - H(R_1) = \log\frac{9}{3} - \log\frac{9}{5} = \log\frac{5}{3} = 0.7370.$$

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

Compared to neighborhood entropy [16], [17], the proposed discrimination indexes have wider application. These indexes can be not only used to compute the distinguishing ability of a reflexive relation but also address a more general binary relation. Next, we provide an example to illustrate this statement. We will discuss a more thorough analysis in the future.

Example 2: Given a set $X = \{x_1, x_2, x_3\}$, R_1, R_2, and R_3 are relations defined on X, where

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

According to neighborhood entropy [16], [17]

$$H(R_1) = -\frac{1}{3}\sum_{i=1}^{3}\log\frac{|[x_i]_{R_1}|}{3} = +\infty$$

TABLE VI
OPTIMAL FEATURES SELECTED BY NEIEN, FINEN, AND HANDI ALGORITHMS

Dataset	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

$$H(R_2) = -\frac{1}{3}\sum_{i=1}^{3}\log\frac{|[x_i]_{R_2}|}{3} = 0.5014$$

$$H(R_3|R_1) = -\frac{1}{3}\sum_{i=1}^{3}\log\frac{|[x_i]_{R_1}\cap[x_i]_{R_3}|}{|[x_i]_{R_1}|} = -\log\frac{1}{6} - \log\frac{0}{0}$$

where $[x_i]_{R_1}$ is the successor neighborhood of x_i with respect to R_1 (see [16], [17]). According to the discrimination index, we have

$$H(R_1) = \log\frac{9}{5} = 0.5878, \quad H(R_2) = \log\frac{9}{6} = 0.4055$$

$$H(R_3|R_1) = H(R_1R_3) - H(R_1) = \log\frac{5}{2} = 0.9163.$$

Although the differences between the distinguishing abilities of R_1 and R_2 are considerably small, the neighborhood entropy of R_1 is an infinite value. The value of the discrimination index of R_1, conversely, is more reasonable. Further, the conditional entropy of R_3 relative to R_1 is meaningless.

IV. FEATURE SELECTION ALGORITHM BASED ON NEIGHBORHOOD DISCRIMINATION INDEX

As discussed above, the proposed discrimination indexes can be used to measure the distinguishing ability of a relation or a feature subset. The smaller the conditional discrimination index of a feature subset, the greater the distinguishing ability of the feature subset and hence, the more important the feature subset. According to the definition of the conditional discrimination index, adding a new feature to the selected feature subset could increase or decrease the conditional discrimination index. A feature can lead to a decrease of the index only when it is irrelevant to the selected feature subset. The decrement of the conditional discrimination index reflects the increment of the distinguishing ability produced by a new

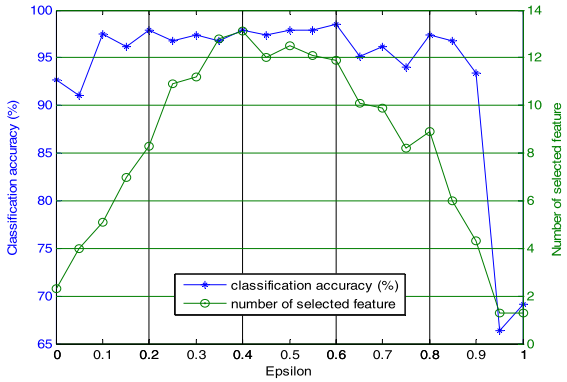


Fig. 2. Number of selected features and accuracy varying with neighborhood radius ε (Wine).

feature subset. Hence, the significance of a feature can be defined as follows.

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

$$\mathrm{SIG}(a, B, D) = H^{\varepsilon}(D|B) - H^{\varepsilon}(D|B\cup\{a\}). \tag{17}$$

When $B = \emptyset$, we define $H^{\varepsilon}(D|B) = H^{\varepsilon}(D)$. The significance of attribute a depends on the increment of the distinguishing information after adding a into B. A large value of $\mathrm{SIG}(a, B, D)$ 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 Algorithm 1.

The parameter δ is used to terminate the main loop in this algorithm. It must be set in advance. For a given data set, generally speaking, the number of the selected features increases if the value of the parameter δ decreases.

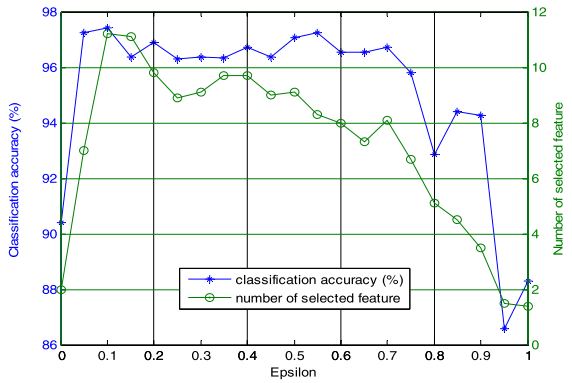


Fig. 3. Number of selected features and accuracy varying with neighborhood radius ε (Wdbc).

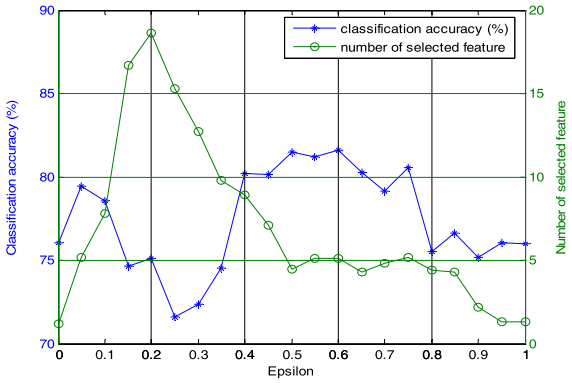


Fig. 4. Number of selected features and accuracy varying with neighborhood radius ε (Wpbc).

The algorithm employs SIG(a, B, D) to determine the optimal attribute is 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 the 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 $O((N^2+N)/2)$.

V. Experimental Analysis

To verify the feasibility and effectiveness of the proposed algorithm, we compared the proposed algorithm with the neighborhood rough set-based algorithm (NRS) [15], neighborhood entropy-based algorithm (NEIEN) [16], fuzzy information entropy-based algorithm (FINEN) [17], [58], and fuzzy rough dependence constructed by intersection operations of fuzzy similarity relations [20]. We employed the Chebychev distance function to compute neighborhood similarity relations. We first compared: 1) the numbers of selected features; 2) the running time of reduction; and 3) the classification accuracies based on these algorithms. Then, we discussed the influence of the neighborhood radius ε on the proposed algorithm. All the algorithms were executed in MATLAB 2013b and run in a hardware environment with a Intel (R) Core (TM) i7-4790 CPU at 3.60 GHz, with 16-GB RAM.

We employed ten-fold cross validation and two classical classifiers to evaluate these algorithms. The two classifiers were support vector machine (RBF-SVM) and k-nearest neighbor rule ($K=3$). Because our main purpose was to compare the performances of the different feature selection algorithms, the parameter selection for RBF-SVM was not a concern. Thus, in this experiment, we consistently set the control term C as 100 and the Gaussian kernel parameter g as one. Such parameter specifications can perform well on real-world problems [64]. The experimental comparison was conducted based on a ten-fold cross validation. That is, the original data set was randomly divided into ten subsets; one was used as the testing data and the remaining nine were used for training. Feature selection was performed on the training set; the reduced training and testing sets were then sent to a classifier to produce the classification accuracy. After ten rounds, the average value and variation of the classification accuracies were computed as the final performance. Thirteen data sets were used in the experimental analysis. They were selected from the UCI Machine Learning Repository [3] and Keng Ridge Biomedical Data set Repository [65]. The information regarding these data sets is outlined in Table I. All the numerical attributes were first normalized into the interval [0, 1].

There are two parameters in the HANDI algorithm, ε and δ. The parameter ε is introduced to control sample similarity; it has a significant influence on the performance of the algorithm. In general, different values of the neighborhood radius can lead to different classification accuracies; therefore, we selected an optimal feature subset for each data set by adjusting the value of the parameter to vary from zero to one with a step of 0.05. The parameter δ was 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

Algorithm 1 Heuristic Algorithm Based on Neighborhood Discrimination Index (HANDI)

Input: decision table $\langle U, A, D\rangle$ and ε // ε is the neighborhood radius.
Output: one reduct red.
1: Initialize: $red=\emptyset$, $B=A-red$, start $=1$; // red is the pool to contain the selected attributes and B is for the remaining 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.

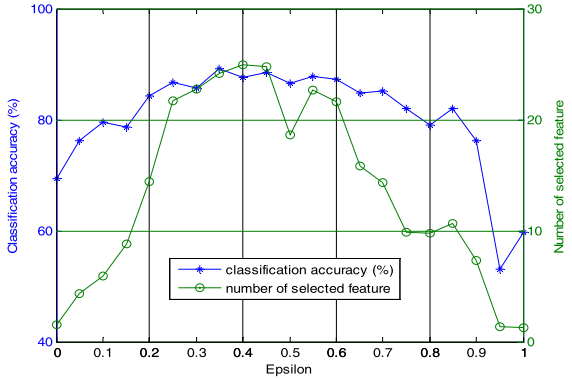


Fig. 5. Number of selected features and accuracy varying with neighborhood radius ε (Sonar).

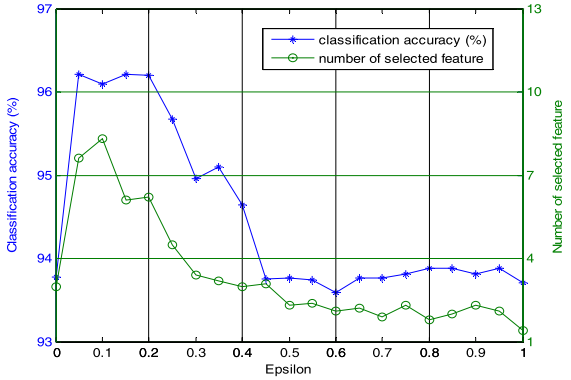


Fig. 7. Number of selected features and accuracy varying with neighborhood radius ε (Sick).

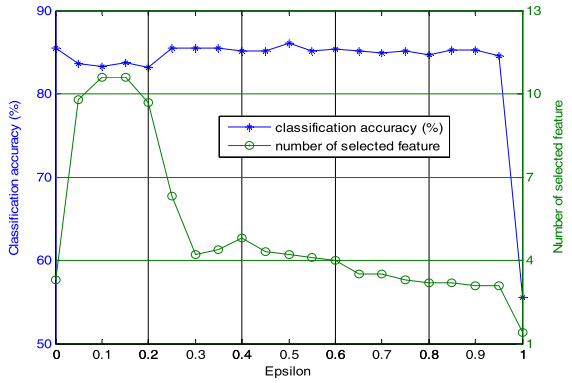


Fig. 6. Number of selected features and accuracy varying with neighborhood radius ε (Credit).

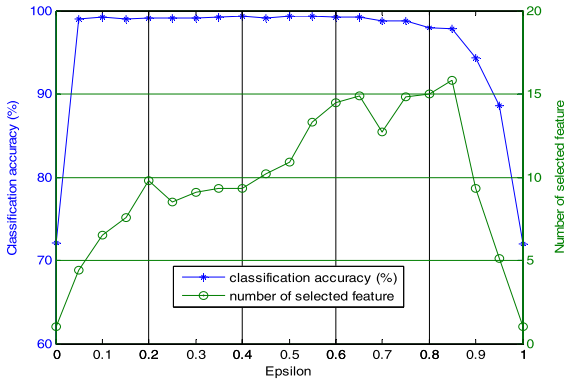


Fig. 8. Number of selected features and accuracy varying with neighborhood radius ε (Gearbox).

the best classification accuracy, all the experimental results reported in the following tables are presented at the highest classification accuracy.

Table II presents a comparison of the average size of the selected features with different algorithms. Because the highest classification accuracy of each data set was searched by adjusting the values of ε, the values of parameter ε were different for the highest accuracies of the data sets. The last column in Table II, labeled ε, indicates the value of the neighborhood radius used in the HANDI algorithm, where the best classification performances were produced on the corresponding data sets.

From the Table II, we can determine that these reduction methods can effectively reduce attributes. The number of selected features with HANDI was less than the other four algorithms in the majority of the cases. For the Sonar data set, HANDI identified more features than the FRSINT algorithm, yet less than the NRS, NEIEN, and FINEN algorithms. For Tumors, HANDI identified more features than NRS and FRSINT, yet less than NEIEN and FINEN. This implies that the proposed algorithm is more effective in reducing redundant attributes.

The classification accuracies of the raw data and the reduced data sets based on the five algorithms are presented in Tables III and IV, where the underlined symbols highlight the highest classification accuracy among the reduced data sets. From the results of Tables III and IV, it is clear that the classification accuracies based on the NRS method are lower than the other four methods. Out of 26 cases of ten-fold cross validation, the HANDI and FINEN methods achieved the highest classification accuracy in 13 and 7 cases, whereas the NRS, NEIEN, and FRSINT methods obtained the highest classification in 3, 5, and 2 cases, respectively. For SVM, HANDI outperformed the raw data 12 times over the 13 classification tasks; it outperformed the raw data 11 times with respect to 3NN. Moreover, the average accuracy of HANDI was superior to all other feature selection algorithms in terms of the SVM and 3NN learning algorithms.

From Table V, we can determine that the running time of the reduction of the NRS algorithm was the least of the five different algorithms. The HANDI algorithm executed more slowly than the NRS algorithm, yet faster than the other three algorithms. The running time of the FRSINT algorithm was the greatest. As the NEIEN, FINEN, FRSINT, and HANDI algorithms were based on similarity relations, they required significant time to compute the similarity relations of the attributes. The NRS algorithm does not compute similarity relations; it just required some time to determine if the samples in a neighborhood were similar. Hence, the NRS algorithm executed the fastest. Because the NEIEN and FINEN algorithms require additional time to compute the similarity class of each sample based on similarity relations, they executed more slowly than the HANDI algorithm. For the FRSINT algorithm, it not only depends on similarity relations but

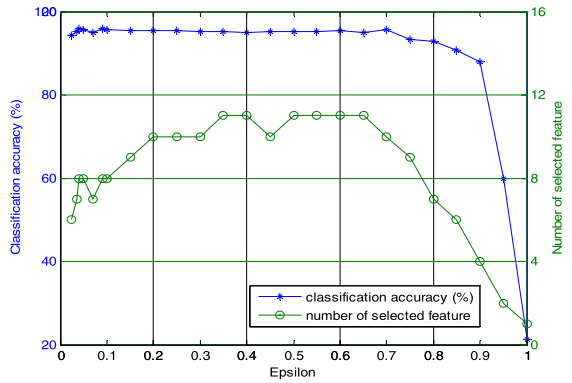


Fig. 9. Number of selected features and accuracy varying with neighborhood radius ε (Segmentation).

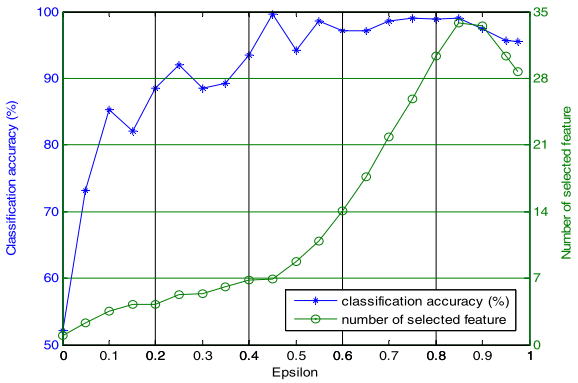


Fig. 12. Number of selected features and accuracy varying with neighborhood radius ε (MLL).

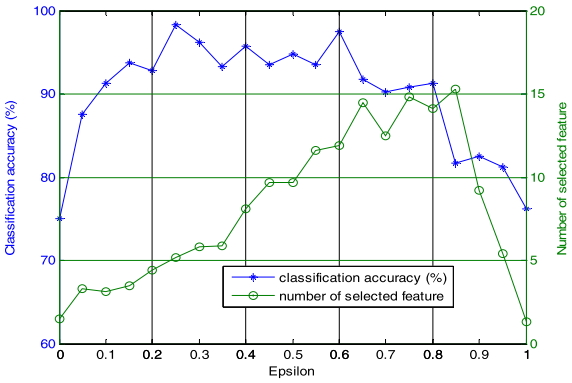


Fig. 10. Number of selected features and accuracy varying with neighborhood radius ε (DLBCL).

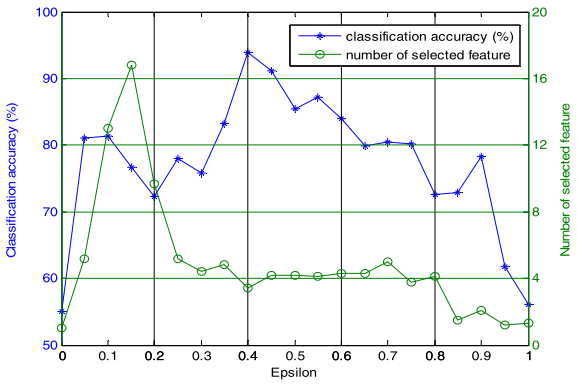


Fig. 13. Number of selected features and accuracy varying with neighborhood radius ε (Prostate).

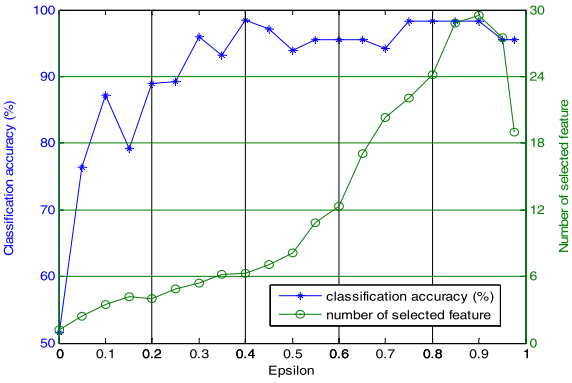


Fig. 11. Number of selected features and accuracy varying with neighborhood radius ε (Leukemia).

also requires time to compute the fuzzy-rough membership of each sample to the different decision categories. Therefore, the FRSINT algorithm executed the slowest. From Tables III and IV, we can observe that the majority of the classification accuracies of the HANDI algorithm are higher than those of the other four algorithms. The complexity of HANDI is less than the NEIEN, FINEN, and FRSINT algorithms. Therefore, it can be concluded that the HANDI algorithm is both feasible and effective.

To present 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 the parameters where the classification accuracies were obtained in the above experiments. The selected feature subsets are listed in Table VI. It can be observed that the optimal features selected by HANDI are virtually the subsets of the optimal features selected by NEIEN or FINEN in the majority of cases. For example, this is the case for the Wine data set if the effects of the fourth feature (Alcalinity) and fifth feature (Magnesium) are treated as equivalent. Other similar data sets include Wpbc, Credit, Sick, Gearbox, Leukemia, and Prostate. This result confirms that HANDI can reduce a greater number of redundant features than the NEIEN or FINEN methods. For other data sets, although the optimal features selected by these three algorithms were different, there were always common features in the selected feature subsets. The difference in the feature subsets indicates that there are multiple subsets of features that have acceptable classification power for a given classification task. It should be noted that HANDI always selected the optimal subsets having a fewer number of features. For the Sick and Segmentation data sets, the selected feature subsets were identical and the classification accuracies were virtually the same for the NEIEN and FINEN algorithms. The marginal differences for the Segmentation could be because the selected feature subsets were presented by reducing the entire data set, whereas the classification accuracies were based on ten-fold cross validation.

Finally, we present Figs. 2–13 to demonstrate the number of selected features and classification accuracies varying with ε; we only display the curves of some data sets with SVM. The data curves drawn using 3NN were reasonably consistent with SVM. From Figs. 2–13, it is clearly observed that the parameter ε has significant influence on the performance of the HANDI algorithm. The majority of the 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 illustrate that the classification performance is stable and can provide a selection of an optimal subset of features. The optimal positions of the classification accuracies are different among these data sets. We recommend that ε should be set to values in the interval [0.1, 0.6].

VI. Conclusion

Measures for computing the distinguishing ability of a subset of features have an important role in classification learning and feature selection. A number of measures have been developed for these tasks. Considering its effectiveness, information entropy is widely used and discussed for evaluating features. In this paper, we introduced basic ideas in Shannon information theory into a neighborhood relation context and proposed discrimination indexes to measure the distinguishing ability of a subset of features. The proposed discrimination indexes were directly defined on a neighborhood relation and computed by considering the cardinality of the neighborhood relations rather than neighborhood similarity classes. The conditional discrimination index was 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 proposed a new algorithm for feature selection. With thirteen public data sets, a series of experiments were conducted for evaluating the proposed method. The results confirm that the algorithm selected fewer features, retained higher classification accuracy, and required less time. Further, the majority of the classification accuracies were improved. We also determined that different parameters have an influence on the performance of the feature selection algorithm. It is important to select a suitable value for the threshold for each data set according to the curves of the data set.

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, Jul. 1994.

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

[3] C. L. Blake and C. J. Merz. (1998). *UCI Repository of Machine Learning Databases*. [Online]. Available: http://www.ics.uci.edu/~mlearn/MLRepository.html

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

[5] M. Dash and H. Liu, "Consistency-based search in feature selection," *Artif. Intell.*, vol. 151, nos. 1–2, pp. 155–176, 2003.

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

[7] I. Düentsch and G. Gediga, "Uncertainty measures of rough set prediction," *Artif. Intell.*, vol. 106, pp. 109–137, Nov. 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, Jul. 2004, pp. 43–50.

[9] S. Greco, B. Matarazzo, and R. Slowinski, "Rough approximation by dominance relations," *Int. J. Intell. Syst.*, vol. 17, no. 2, 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, Jan. 2003.

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

[12] S. Huang, C. Li, and Y. Liu, "Complex-valued filtering based on the minimization of complex-error entropy," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 24, no. 5, pp. 695–708, May 2013.

[13] 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, Mar. 2002.

[14] J. Hou, H. Gao, Q. Xia, and N. Qi, "Feature combination and the kNN framework in object classification," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 27, no. 6, pp. 1368–1378, Jun. 2016.

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

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

[17] 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, Apr. 2006.

[18] E. Hernández and J. Recasens, "A reformulation of entropy in the presence of indistinguishability operators," *Fuzzy Sets Syst.*, vol. 128, pp. 185–196, Jun. 2002.

[19] M. Inuiguchi, Y. Yoshioka, and Y. Kusunoki, "Variable-precision dominance-based rough set approach and attribute reduction," *Int. J. Approx. Reason.*, vol. 50, no. 8, pp. 1199–1214, 2009.

[20] 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, Dec. 2004.

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

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

[23] 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, Dec. 2002.

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

[25] J. Li, Y. Ren, C. Mei, Y. Qian, and X. Yang, "A comparative study of multigranulation rough sets and concept lattices via rule acquisition," *Knowl.-Based Syst.*, vol. 91, pp. 152–164, Jan. 2016.

[26] J. Liang, G. Yu, B. Chen, and M. Zhao, "Decentralized dimensionality reduction for distributed tensor data across sensor networks," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 27, no. 11, pp. 2174–2186, Nov. 2016.

[27] J. Liang, F. Wang, C. Dang, and Y. Qian, "A group incremental approach to feature selection applying rough set technique," *IEEE Trans. Knowl. Data Eng.*, vol. 26, no. 2, pp. 294–308, Feb. 2014.

[28] 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, Apr. 2005.

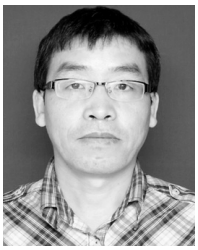
[29] J.-S. Mi, Y. Leung, H.-Y. Zhao, and T. Feng, "Generalized fuzzy rough sets determined by a triangular norm," *Inf. Sci.*, vol. 178, no. 16, pp. 3203–3213, 2008.

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

[31] J.-S. Mi, W.-Z. Wu, and W.-X. Zhang, "Approaches to knowledge reduction based on variable precision rough set model," *Inf. Sci.*, vol. 159, nos. 3–4, pp. 255–272, 2004.

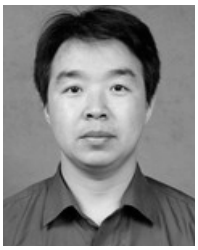
[32] P. Mitra, C. A. Murthy, and S. K. Pal, "Unsupervised feature selection using feature similarity," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 3, pp. 301–312, Mar. 2002.

[33] D. P. Muni, N. R. 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, Feb. 2006.
[34] S. Nan, L. Sun, B. Chen, Z. Lin, and K.-A. Toh, "Density-dependent quantized least squares support vector machine for large data sets," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 28, no. 1, pp. 94–106, Jan. 2017.
[35] F. Nie, H. Huang, X. Cai, and C. H. Ding, "Efficient and robust feature selection via joint ℓ2, 1-norms minimization," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, 2010, pp. 1813–1821.
[36] I.-S. Oh, J.-S. Lee, and B.-R. Moon, "Hybrid genetic algorithms for feature selection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 26, no. 11, pp. 1424–1437, Nov. 2004.
[37] Z. Pawlak, *Rough Sets: Theoretical Aspects of Reasoning About Data*. Dordrecht, The Netherlands: Kluwer, 1991.
[38] N. Parthaláin, Q. Shen, and R. Jensen, "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, Mar. 2010.
[39] 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, Aug. 2005.
[40] Y. Qian and J. Liang, "Combination entropy and combination granulation in rough set theory," *Int. J. Uncertainty, Fuzziness Knowl.-Based Syst.*, vol. 16, no. 2, pp. 179–193, 2008.
[41] Y. Qian, F. Li, J. Liang, B. Liu, and C. Dang, "Space structure and clustering of categorical data," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 27, no. 10, pp. 2047–2059, Oct. 2016.
[42] R. Slowinski and D. Vanderpooten, "A generalized definition of rough approximations based on similarity," *IEEE Trans. Knowl. Data Eng.*, vol. 2, no. 2, pp. 331–336, Mar./Apr. 2000.
[43] Y. She and 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.
[44] 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.
[45] H. Tao, C. Hou, F. Nie, Y. Jiao, and D. Yi, "Effective discriminative feature selection with nontrivial solution," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 27, no. 4, pp. 796–808, Apr. 2016.
[46] K. Torkkola, "Feature extraction by non-parametric mutual information maximization," *J. Mach. Learn. Res.*, vol. 3, nos. 7–8, pp. 1415–1438, 2003.
[47] X.-Z. Wang, L.-C. Dong, and J.-H. Yan, "Maximum ambiguity-based sample selection in fuzzy decision tree induction," *IEEE Trans. Knowl. Data Eng.*, vol. 24, no. 8, pp. 1491–1505, Aug. 2012.
[48] C. Wang, M. Shao, Q. He, Y. Qian, and Y. Qi, "Feature subset selection based on fuzzy neighborhood rough sets," *Knowl.-Based Syst.*, vol. 111, no. 1, pp. 173–179, 2016.
[49] C. Wang *et al.*, "A fitting model for feature selection with fuzzy rough sets," *IEEE Trans. Fuzzy Syst.*, to be published, doi: 10.1109/TFUZZ.2016.2574918.
[50] D. Wang, F. Nie, and H. Huang, "Feature selection via global redundancy minimization," *IEEE Trans. Knowl. Data Eng.*, vol. 27, no. 10, pp. 2743–2755, Oct. 2015.
[51] H. Wang, "Nearest neighbors by neighborhood counting," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 6, pp. 942–953, Jun. 2006.
[52] W.-Z. Wu and W.-X. Zhang, "Neighborhood operator systems and approximation," *Inf. Sci.*, vol. 144, nos. 1–4, pp. 201–217, 2002.
[53] W.-Z. Wu, "Knowledge reduction in random incomplete decision tables via evidence theory," *Fundam. Inf.*, vol. 115, nos. 2–3, pp. 203–218, 2012.
[54] W.-H. Xu, X.-Y. Zhang, and W.-X. Zhang, "Knowledge granulation, knowledge entropy and knowledge uncertainty measure in ordered information systems," *Appl. Soft Comput.*, vol. 9, no. 4, pp. 1244–1251, 2009.
[55] S. Xiang, F. Nie, G. Meng, C. Pan, and 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, Nov. 2012.
[56] Y. Y. Yao, "Relational interpretations of neighborhood operators and rough set approximation operators," *Inf. Sci.*, vol. 111, nos. 1–4, pp. 239–259, 1998.
[57] Y. Yao and Y. She, "Rough set models in multigranulation spaces," *Inf. Sci.*, vol. 327, pp. 40–56, Jan. 2016.
[58] R. R. Yager, "Entropy measures under similarity relations," *Int. J. General Syst.*, vol. 20, no. 4, pp. 341–358, 1992.
[59] S. Zhao, H. Chen, C. Li, M. Zhai, and X. Du, "RFRR: Robust fuzzy rough reduction," *IEEE Trans. Fuzzy Syst.*, vol. 21, no. 5, pp. 825–841, Oct. 2013.
[60] S. Zhao, E. C. C. Tsang, D. Chen, and X. Wang, "Building a rule-based classifier—A fuzzy-rough set approach," *IEEE Trans. Knowl. Data Eng.*, vol. 22, no. 5, pp. 624–638, May 2010.
[61] S. Zhao and E. C. C. Tsang, "On fuzzy approximation operators in attribute reduction with fuzzy rough sets," *Inf. Sci.*, vol. 178, no. 16, pp. 3162–3176, 2008.
[62] X. Zhang, B. Zhou, and P. Li, "A general frame for intuitionistic fuzzy rough sets," *Inf. Sci.*, vol. 216, no. 24, pp. 34–49, 2012.
[63] P. Zhu and Q. H. Hu, "Adaptive neighborhood granularity selection and combination based on margin distribution optimization," *Inf. Sci.*, vol. 249, pp. 1–12, Nov. 2013.
[64] S. S. Ho and H. Wechsler, "Query by transduction," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 9, pp. 1557–1571, Sep. 2008.
[65] *Kent Ridge Bio-Medical Dataset*, accessed on Apr. 2015. [Online]. Available: http://datam.i2r.a-tar.edu.sg/datasets/krbd/index.html



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

He is currently a Professor with Bohai University. He has authored or co-authored more than 40 journal and conference papers in the areas of machine learning, data mining, and rough set theory. His current research interests include fuzzy sets, rough sets, data mining, pattern recognition, and statistical analysis.



Qinghua Hu received the B.Eng. and M.Eng. degrees in power engineering and the Ph.D. degree in control science and engineering from the Harbin Institute of Technology, Harbin, China, in 1999, 2002, and 2008, respectively.

He is currently a Professor with Tianjin University, Tianjin, China. He has authored or co-authored more than 60 journal papers and conference proceedings in machine learning and data mining. His current research interests include data mining and knowledge discovery with fuzzy and rough techniques.

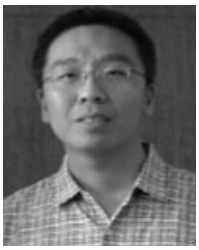
Xizhao Wang received the Ph.D. degree in computer science from the Harbin Institute of Technology, Harbin, China, in 1998.

He is currently a Professor with the Big Data Institute, Shenzhen University, Shenzhen, China. His current research interests include uncertainty modeling and machine learning for big data. He has edited more than ten special issues and published three monographs, two textbooks, and more than 200 peer-reviewed research papers. By the Google scholar, the total number of citations is over 5000. He is on the list of Elsevier 2015/2016 most cited Chinese authors.

Dr. Wang is the Chair of the IEEE SMC Technical Committee on Computational Intelligence, the Editor-in-Chief of *Machine Learning and Cybernetics Journal*, and Associate Editor for a couple of journals in the related areas. He was a recipient of the IEEE SMCS Outstanding Contribution Award in 2004 and a recipient of the IEEE SMCS Best Associate Editor Award in 2006.

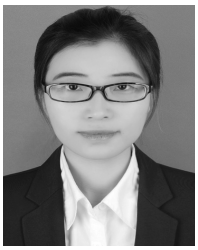
Degang Chen received the M.S. degree from Northeast Normal University, Changchun, China, in 1994, and the Ph.D. degree from the Harbin Institute of Technology, Harbin, China, in 2000.

He was a Post-Doctoral 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 current research interests include fuzzy groups, fuzzy analysis, rough sets, and support vector machines.



Yuhua Qian received the M.S. and Ph.D. degrees in computers with applications, from Shanxi University, Taiyuan, China, in 2005 and 2011, respectively.

He is currently a Professor with the Key Laboratory of Computational Intelligence and Chinese Information Processing, Ministry of Education, China. He has authored more than 50 articles in international journals. His current research interests include pattern recognition, feature selection, rough set theory, granular computing, and artificial intelligence.



Zhe Dong received the B.Sc. degree in mathematics from Bohai University, Jinzhou, China, in 2012. She is currently pursuing the master's degree.

Her current research interests include fuzzy sets, rough sets, pattern recognition, and knowledge discovery.