Selecting Discrete and Continuous Features Based on Neighborhood Decision Error Minimization

Qinghua Hu, Witold Pedrycz, *Fellow, IEEE*, Daren Yu, and Jun Lang

***Abstract*—Feature selection plays an important role in pattern recognition and machine learning. Feature evaluation and classification complexity estimation arise as key issues in the construction of selection algorithms. To estimate classification complexity in different feature subspaces, a novel feature evaluation measure, called the neighborhood decision error rate (NDER), is proposed, which is applicable to both categorical and numerical features. We first introduce a neighborhood rough-set model to divide the sample set into decision positive regions and decision boundary regions. Then, the samples that fall within decision boundary regions are further grouped into recognizable and misclassified subsets based on class probabilities that occur in neighborhoods. The percentage of misclassified samples is viewed as the estimate of classification complexity of the corresponding feature subspaces. We present a forward greedy strategy for searching the feature subset, which minimizes the NDER and, correspondingly, minimizes the classification complexity of the selected feature subset. Both theoretical and experimental comparison with other feature selection algorithms shows that the proposed algorithm is effective for discrete and continuous features, as well as their mixture.**

***Index Terms*—Continuous feature, decision error minimization, discrete feature, feature selection, neighborhood, rough sets.**

I. Introduction

A CLASSIFIER that is learned through inductive learning assigns a given pattern to one of the classes. A typical representation of a pattern comes in the form of a vector of features. Patterns are points in the feature space. Classification performance substantially depends on the selection of the feature space (for example, see [47], [51], and [52]). Given a limited size of learning sets, excessive numbers of features may greatly deteriorate the quality of the classifiers, because irrelevant and redundant features are highly confusing in the learning process [9], [44]. Feature selection becomes much more essential for pattern recognition [55], [59], [60].

Manuscript received November 5, 2008; revised February 25, 2009 and May 8, 2009. First published July 17, 2009; current version published October 30, 2009. This work was supported in part by the National Natural Science Foundation of China under Grant 60703013 and by the Development Program for Outstanding Young Teachers, Harbin Institute of Technology, under Grant HITQNJS.2007.017. This paper was recommended by Associate Editor X. Wang.

Q. Hu and J. Lang are with Harbin Institute of Technology, Harbin 150001, China (e-mail: huqinghua@hit.edu.cn; billlangjun@gmail.com).

W. Pedrycz is with the Department of Electrical and Computer Engineering, University of Alberta, Edmonton, AB T6G 2V4, Canada, and also with the Systems Science Institute, Polish Academy of Sciences, 01-447 Warsaw, Poland (e-mail: pedrycz@ee.ualberta.ca).

D. Yu is with the School of Energy Science and Engineering, Harbin Institute of Technology, Harbin 150001, China (e-mail: yudaren@hit.edu.cn).

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/TSMCB.2009.2024166

There are four problems in feature selection [28]:

1) feature evaluation;
2) search strategies;
3) stopping criterion;
4) validation strategies.

Among them, feature evaluation and search strategies play essential roles in this process. A search strategy is a procedure for finding optimal subsets of features with regard to a certain evaluation function. Greedy selection [2], [6], [27], [24], [30], branch and bound (B&B) [30], [41], floating search [35], and genetic algorithms (GAs) [37], [44], [55] were studied in feature selection.

Feature evaluation functions are used to measure the quality of the candidate subsets [21], [22], [46]. Evaluation criteria play a very important role in feature selection. An optimal criterion should naturally relate the Bayes error rate of classification in the feature space [34], [43]. However, computing Bayes error rates requires detailed knowledge of the class probability distribution, whereas in practice, class probabilities are unknown. One has to estimate these probabilities by making use of a finite size of samples, which is very difficult, particularly when dealing with highly dimensional feature spaces [21], [23], [34], [36], [43]. Quite commonly, we focus on the design of performance measures to determine the relevance between features and decision. Distance [12], [39], correlation [9], [11], mutual information [3], [60], consistency [7], and dependency [18] are usually considered feasible alternatives. Mutual information is widely applied to characterize the relevance between categorical attributes and classification decisions [13], [24], [34]. Wang introduced an axiomatic framework for feature selection based on mutual information [46]. A dependency-based feature selection algorithm was proposed in [17], where dependency is defined as the ratio of the so-called positive region in the rough set (RS) theory over the whole set of samples. Samples with the same attribute values and different decisions are called classification boundary. However, the rate of positive region is not an effective estimate of classification accuracy. According to the Bayes rule, samples with the same feature values will be classified as belonging to the majority class. Therefore, only the samples in the minority classes are misclassified in this case. Based on this observation, Dash and Liu introduced the measure of consistency and employed it to evaluate the quality of features [7], where consistency is treated as the ratio of the samples that can be recognized with the Bayes rule. We may contemplate that consistency captures the natural objective of feature selection, i.e., selecting the feature subset that minimizes the Bayes decision error rate. Unfortunately, mutual information, dependency,

and consistency are all just applicable to evaluate discrete features.

As far as continuous features are concerned, discretization is introduced to segment their domains into several intervals [11], [26]. Subsequently, the discretized features are evaluated by making use of the aforementioned techniques. The quality of the resulting feature subset depends not only on feature selection algorithms but also on the discretization method that is used at the beginning of the entire process. In addition, there are also some techniques that were proposed for a direct selection of numerical features. Distance is a general measure for characterizing the class separability in a metric space [12], [19], [39]. Intuitively, it is desired to find a feature space where intraclass distance is minimal, whereas interclass distance is maximal [8]. The ReliefF algorithm [39] relies on this idea. It tries to find two sets of k-nearest samples ($k \geq 1$) from the same class and other classes, respectively, and then compute the distance between two sets. Obviously, this algorithm is computationally expensive for repeated calculation of k neighbors and the determination of their distances. Moreover, the theory about margin [10] shows that classification complexity depends on the boundary samples, i.e., the so-called support vectors, whereas ReliefF randomly selects samples to compute the weights of features, which is not consistent with the essence of the margin theory. The size of the decision boundary region is another kind of measure for evaluating numerical features. Lee and Landgrebe first captured this idea in their feature extraction algorithm [25]. Thawonmas and Abe introduced this idea for feature selection. They used hyperboxes or ellipsoids to approximate decision region and calculate the overlap of classes as a decision boundary [1], [42]. Obviously, this approximation is rather coarse if the class regions are not of regular shapes.

We propose a novel evaluation measure that is applicable to discrete and continuous features by introducing a neighborhood RS (NRS) model to compute the decision boundary in mixed feature spaces [49], [57], [58]. The definition of NRS were introduced and discussed in several literatures [57], [58]. Later, this model is extended to deal with classification learning with numerical features [49]. Usually, RSs evaluate the quality of features based on the size of classification boundary; a number of researches take the rate of boundary over the sample set as the measure of feature quality [17], [18], [32], [49], [61]. However, as we know, not all the boundary samples are misclassified [7], [61]. In this paper, the samples in boundary regions are further divided into two subsets based on the information of class distribution in samples' neighborhoods: 1) samples in the majority class and 2) samples in the minority classes. The samples in the minority classes are misclassified according to the Bayes rule. Then, the percentage of misclassified samples is taken as the estimate of the classification complexity encountered in the corresponding feature subspaces. We call it the neighborhood decision error rate (NDER). We show that the proposed measure is robust to outliers and complex nonlinear decision boundary. We present a strategy for feature subset selection based on the idea of neighborhood decision error minimization (NDEM). We compare the proposed technique with some current approaches by running experiments for some University of California, Irvine (UCI) data sets.

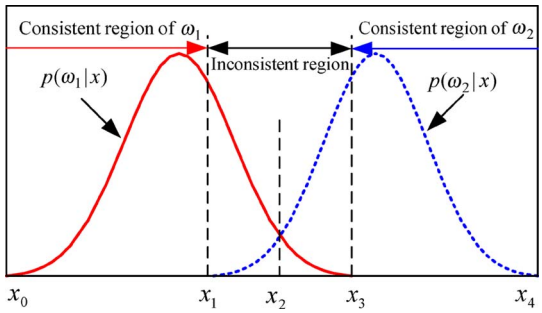


Fig. 1. Classification complexity in a 1-D feature space.

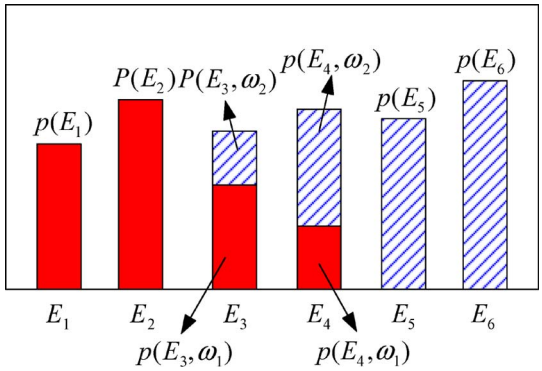


Fig. 2. Classification complexity in a discrete feature space.

II. Basic Idea and Related Work

In feature selection, one has to find features that can effectively distinguish between different classes. The optimal criterion for classification complexity of feature spaces would reflect the Bayes error rate that was observed in $X = \{x_1, x_2, \ldots, x_N\}$, i.e.,

$$e = \int \left[1 - \max_i p(\omega_i|X)\right] p(X)dX$$

where X is the value domain of features, ω_i stands for class i, and $p(\omega_i|X)$ is the conditional probability density function [8]. To compute the classification complexity that was expressed in feature space X, we have to know $p(X)$ and $p(\omega_i|X)$, which are usually not readily available in case of real-world classification tasks. Unfortunately, it is not feasible to estimate them in a high-dimensional space, given a finite and, sometimes, quite small number of samples.

For simplicity, we express the idea that refers to a binary classification problem in a 1-D space, as shown in Fig. 1. According to the class probability density function, the feature space becomes divided into three parts: 1) a consistent region of class 1; 2) a consistent region of class 2; and 3) an inconsistent region where the samples with the same feature values may belong to different classes, because the class probability densities of two classes overlap in this area. The size of the inconsistent region reflects the classification complexity of the corresponding feature spaces.

Fig. 2 shows a similar case in discrete spaces, where the samples are divided into a set of equivalence classes $\{E_1, E_2, \ldots, E_K\}$ based on their feature values. Samples with the same feature values are grouped into one equivalence class; the height of the rectangles in Fig. 2 denotes the probability

$p(E_i)$ of the equivalence class, and $p(\omega_i, E_j)$ is the joint probability of ω_i and E_j.

We can see that some of the equivalence classes are consistent, because their samples belong to one of the decision classes, e.g., E_1, E_2, E_5, and E_6. However, there are also some inconsistent equivalence classes like E_3 and E_4, where samples with the same feature values are assigned to different classes. According to the RS theory, this kind of samples forms the decision boundary region, and the union of consistent samples is called the decision positive region [32].

Based on the comparison of the discrete feature spaces and numerical spaces, it can be concluded that, in classification, a main source of classification complexity comes with inconsistent regions of decision, where samples with identical (in the discrete case) or similar (in the numerical case) feature values belong to different decision classes; hence, inconsistent samples lead to misclassification. The objective of feature selection is to find a subset of features that minimizes the inconsistent region, i.e., minimizes the Bayes decision error [20]. It is therefore desirable to have a measure to reflect the size of an inconsistent region for discrete and numerical spaces for feature selection.

Let us review some measures for estimating complexity in a discrete or numerical feature space. In the discrete space, a dependency function [32] in the theory of RSs is defined as

$$\gamma_B(D) = \frac{\|\text{POS}_B(D)\|}{\|U\|}$$

where U is the set of samples, $\|X\|$ is the cardinality of set X, $\text{POS}_B(D) = \cup_{i=1}^{N} \underline{B}X_i$, $\underline{B}X_i = \{x_j | [x_j]_B \subseteq X_i\}$, and $[x_j]_B$ is the equivalence class that was induced by x_j and attribute subset B. $[x_j]_B$ is the set of samples with the same attribute values as sample x_j in terms of attribute subset B.

Dependency reflects a ratio of consistent samples over the whole set of samples. Therefore, dependency does not take the boundary samples into account when computing the significance of specific attributes. Once there are inconsistent samples in an equivalence class, these equivalence classes are completely ignored. However, inconsistent samples can be divided into two groups: 1) a subset of samples from the majority class and 2) a subset from the minority classes. According to the Bayes rule, only samples under the minority classes are misclassified. For example, the samples in E_3 and E_4 are inconsistent (see Fig. 2). However, only the samples labeled as $P(\omega_2, E_3)$ and $P(\omega_1, E_4)$ are misclassified. The classification accuracy in this case is expressed as follows:

$$f = \sum_{i=1}^{6} P(E_i) - P(\omega_2, E_3) - P(\omega_1, E_4).$$

Consistency captures this idea [7], [16]. It is the percentage of the samples that are recognizable according to the Bayes rule. In the discrete case, $P(E_i)$ can be calculated from $\|E_i\|/\|U\|$, and $P(\omega_j|E_i)$ can be computed in a similar fashion.

For a numerical feature space, it is not easy to precisely compute the decision boundary region. In [12], some measures of the overlap of feature values were proposed to reflect complexity in feature spaces. Fisher's discriminant ratio is given by

$$\text{F1} = \frac{(u_1 - u_2)^2}{\sigma_1^2 + \sigma_2^2}$$

where u_1, u_2, σ_1^2, and σ_1^2 are the means and variances of two classes. Ho used the maximum F1 over all the features as complexity in [12]. However, F1 does not work if class probabilities do not satisfy the normal assumption, particularly in the case that the classification boundary is irregular. One similar measure, which is denoted by F2, quantifies an overlap of the tails of two class conditional distribution [12] defined as in the equation shown at the bottom of the page. Here, $\max(f_i, \omega_j)$ and $\min(f_i, \omega_j)$ are the maximum and minimum values of feature f_i in class ω_j. It is known that the maximum and minimum values are not robust to noise. F2 cannot reflect the real complexity of the feature space if there are several noisy samples. Furthermore, this measure completely overlooks the influence of class probability densities. Therefore, efficiency, which was denoted by F3, was introduced, where the efficiency of each feature is defined as the fraction of samples out of the overlap region [12]. F2 and F3 share two common disadvantages. First, these measures are sensitive to noisy information that was conveyed by samples, because they define the overlap region with the maximum and minimum values of classes. Second, they consider only separating hyperplanes that were perpendicular to the feature axes. Therefore, even for a linearly separable problem, F2 and F3 may be less than 1 if the optimal separating hyperplanes happens to be oblique. Subsequently, several other measures based on boundary region, e.g., N1, N2, N3, and T1 were introduced. Except for N3, these measures regard samples in the boundary region as the source of classification complexity. The sole difference lies in the way of defining the decision boundary region. In fact, as aforementioned, not all samples in the boundary region will be misclassified. It is not rational to take the measures of F2, F3, N1, N2, and T1, computing the probabilities of samples in the boundary region, as complexity [12]. According to the Bayes rule, only the boundary samples in the minority classes will be misclassified. Therefore, they are the real source that implies the classification complexity and the emergence of the decision error. Given this condition, it is desirable to form a theoretic framework and construct an algorithm for estimating the Bayes error rate in feature subset selection.

III. Theoretical Framework for NDEM

As underlined, classification complexity mainly results from the existence of inconsistent regions (e.g., overlap regions and decision boundary regions), where samples with identical or similar feature values would belong to different classes. Here, we introduce an RS methodology to form a theoretic

$$\text{F2} = \prod_i \frac{\min\left(\max(f_i, \omega_1), \max(f_i, \omega_2)\right) - \max\left(\min(f_i, \omega_1), \min(f_i, \omega_2)\right)}{\max\left(\max(f_i, \omega_1), \max(f_i, \omega_2)\right) - \min\left(\min(f_i, \omega_1), \min(f_i, \omega_2)\right)}$$

framework for discrete and numerical feature selection based on NDEM .

A. NRSs for Discrete and Numerical Features

Palawk's RS model [33] is defined in a discrete information space, where each attribute takes its values in a finite set. As a result, RS-based attribute reduction can only be used to deal with discrete features. Here, we first introduce an extended model, called NRSs, which can be used to deal with discrete and continuous features [49].

Formally, the samples for classification (learning) are expressed as $IS = \langle U, A\rangle$, where U is the nonempty set of samples $\{x_1, x_2, \ldots, x_n\}$, which is called a universe or a sample space, A is the nonempty set of attributes (which are referred to as features, inputs, or variables) $\{a_1, a_2, \ldots, a_m\}$ to characterize the samples, and $f(x, a)$ is the feature value of sample x. To be more specific, $\langle U, A\rangle$ is also called a decision table if $A = C \cup \{D\}$, where C is a set of condition attributes, and D is a decision variable.

Definition 1: Given arbitrary $x_i \in U$ and $B \subseteq C$, a neighborhood $\delta_B(x_i)$ of x_i in subspace B is defined as

$$\delta_B(x_i) = \{x_j | x_j \in U, \Delta_B(x_i, x_j) \leq \delta\}$$

where Δ is a metric. This relation means that, for all x_1, x_2, and x_3 in U, it satisfies the following three conditions: 1) $\Delta(x_1, x_2) \geq 0$, and $\Delta(x_1, x_2) = 0$ if and only if $x_1 = x_2$; 2) $\Delta(x_1, x_2) = \Delta(x_2, x_1)$; and 3) $\Delta(x_1, x_3) \leq \Delta(x_1, x_2) + \Delta(x_2, x_3)$.

There are a huge number of possible metrics that are considered in practice. Considering x_1 and x_2 to be two objects in an N-dimensional space, $A = \{a_1, a_2, \ldots, a_N\}$, with $f(x, a_i)$ denoting the value of sample x in the ith dimension a_i, a general alternative known as the Minkowski distance can be expressed as

$$\Delta_P(x_1, x_2) = \left(\sum_{i=1}^{N} |f(x_1, a_i) - f(x_2, a_i)|^P\right)^{1/P}.$$

It is well known that this distance translates into a Manhattan distance Δ_1 if $P = 1$, a Euclidean distance Δ_2 if $P = 2$, or a Tchebyshev distance if $P = \infty$.

There have been a number of proposed distance functions for mixed numerical and categorical data [45], [52], e.g., the heterogeneous Euclidean−overlap metric (HEOM) function, the value difference metric (VDM), heterogeneous VDM, and interpolated VDM. HEOM is defined as follows:

$$\text{HEOM}(x, y) = \sqrt{\sum_{i=1}^{m} w_{a_i} \times d_{a_i}^2(x_{a_i}, y_{a_i})}$$

where m is the number of attributes, w_{a_i} is the weight of attribute a_i, and $d_{a_i}(x, y)$ is the distance between samples x and y in terms of attribute a_i, which is defined as

$$d_{a_i}(x, y) = \begin{cases} 1, & \text{if the attribute value of } x \\ & \text{or } y \text{ is unknown} \\ \text{overlap}_a(x, y), & \text{if } a \text{ is a nominal attribute} \\ \text{rn\_diff}_a(x, y), & \text{if } a \text{ is a numerical attribute.} \end{cases}$$

Here

$$\text{overlap}(x, y) = \begin{cases} 0, & \text{if } x = y \\ 1, & \text{otherwise} \end{cases}$$

$$\text{rn\_diff}_a(x, y) = |x - y| / \max_a - \min_a .$$

With different metric functions, the proposed technique can be used to analyze categorical attributes, numerical attributes, interval-valued attributes, and their mixtures [45].

Definition 2: Given a set of samples U, N is a neighborhood relation on U, i.e., $\forall x, y \in U$, $R(x, y) = 1$ if $y \in \delta(x)$; otherwise, $R(x, y) = 0$. We call $\langle U, N\rangle$ a neighborhood approximation space.

Definition 3 [49]: Given $\langle U, N\rangle$ for arbitrary $X \subseteq U$, two subsets of objects, which are called the lower and upper approximations of X in terms of relation N, are defined as

$$\underline{N}X = \{x_i | \delta(x_i) \subseteq X, x_i \in U\}$$
$$\overline{N}X = \{x_i | \delta(x_i) \cap X \neq \varnothing, x_i \in U\}.$$

Definition 4 [49]: Given a neighborhood decision table (NDT; $\text{NDT} = \langle U, C, D\rangle$), $X_1, X_2, \ldots, X_N$ are the sample subsets with decisions $1-N$, and the lower and upper approximations of decision D with respect to attributes B are then defined as

$$\underline{N_B}D = \bigcup_{i=1}^{N} \underline{N_B}X_i$$
$$\overline{N_B}D = \bigcup_{i=1}^{N} \overline{N_B}X_i$$

where $\underline{N_B}X = \{x_j | \delta_B(x_j) \subseteq X, x_j \in U\}$, and $\overline{N_B}X = \{x_j | \delta_B(x_j) \cap X \neq \varnothing, x_j \in U\}$.

The decision boundary region of D with respect to attributes B is defined as

$$\text{BN}(D) = \overline{N_B}D - \underline{N_B}D.$$

A decision boundary is the subset of objects whose neighborhoods come from more than one decision class. The lower approximation of decision, which is called the positive region of decision and is denoted by $\text{POS}_B(D)$, is the subset of objects whose neighborhoods consistently belong to one of the decision classes.

Theorem 1: Given $\text{NDT} = \langle U, C, D\rangle$, $B \subseteq C$, we have the following three relations: 1) $\overline{N_B}D = U$; 2) $\text{POS}_B(D) \cap \text{BN}(D) = \varnothing$; and 3) $\text{POS}_B(D) \cup \text{BN}(D) = U$.

The NRS model divides the samples into two subsets: 1) the positive region and 2) the boundary region. Intuitively, the samples in the boundary region are easy to be misclassified. In data acquisition and preprocessing, one usually tries to find a feature space in which the classification task leads to the simplest classification boundary.

B. ND and NDER

The size of boundary samples reflects the classification complexity in a corresponding subspace. It also reflects the distinguishing capability or characterizing power of the condition attributes. One evaluating function of an attribute's significance, which is called the neighborhood dependency (ND), is introduced as follows.

Definition 5: Given NDT $= \langle U, C, D\rangle$, the ND of D to B in the neighborhood approximation space is defined as

$$\gamma_B(D) = \frac{\|\text{POS}_B(D)\|}{\|U\|}$$

where $\gamma_B(D)$ reflects the ability of B to approximate D. Obviously, $0 \le \gamma_B(D) \le 1$. We say that D totally depends on B or that the NDT is consistent if $\gamma_B(D) = 1$, which is denoted by $B \Rightarrow D$; otherwise, we say that $D\gamma -$ depends on B, which is denoted by $B \Rightarrow_r D$.

Dependency reflects the size of overlap between classes. If the samples are completely separable, i.e., consistent (linearly or nonlinearly separable), the dependency measure attains 1; otherwise, we have $\gamma_B(D) < 1$.

Theorem 2: Given NDT $= \langle U, C, D\rangle$, $B_1, B_2 \subseteq C$, and $B_1 \subseteq B_2$, with the same metric Δ and threshold δ for computing neighborhoods, we have the following three relations: 1) $N_{B_1} \supseteq N_{B_2}$; 2) $\forall X \subseteq U$, $\underline{N_{B_1}} X \subseteq \underline{N_{B_2}} X$; and 3) $\text{POS}_{B_1}(D) \subseteq \text{POS}_{B_2}(D)$, $\gamma_{B_1}(D) \le \gamma_{B_2}(D)$.

Theorem 2 shows that adding a new attribute to the current subset of attributes at least does not decrease the dependency. In general, we hope to determine a minimal feature subset with the same characterizing power as the whole features. The monotonicity of dependency is very important for constructing a greedy search algorithm [7], floating search [35], or B&B method [41].

Considering parameter δ as the level of granularity at which we analyze the classification problem, we can find that the complexity of classification depends not only on the given feature space but also on the assumed granularity level. Granularity, which was controlled by the values of parameter δ, can qualitatively be characterized as "fine," "coarse," and the like.

The dependency in NRSs reflects the rate of boundary samples in numerical features, discrete features, or their mixture spaces. In this sense, it naturally extends the definition of dependency in "standard" RSs to deal with numerical and discrete features without resorting to the discretization process. Discrete and numerical features usually coexist in real-world databases; thus, this extension greatly enhances the application scope of RSs.

However, as mentioned in Section II, not all samples in the decision boundary region are necessarily misclassified. Only the samples in minority classes cannot be recognized. Given this case, dependency cannot reflect the true classification complexity. In the discrete cases, we can observe this effect by comparing Figs. 2 and 3: Although the probabilities of inconsistent samples are identical, the probabilities of misclassification differ.

Fig. 4 illustrates a similar problem that arises in case of numerical feature spaces. In Fig. 4(a) and (b), the feature spaces are inconsistent, because two class probability densities are greater than zero. As a result, the neighborhood of any sample would not be "pure" (homogeneous), and the samples in it come from two classes. Therefore, the dependency is zero, whereas the probabilities of Bayes errors are less than 1 in these two cases. We can also note that, according to the Bayes rule, the error probability in A is far less than that in B. Dependency cannot capture these differences. Fig. 4(c) and (d) visualize a similar effect, i.e., the probabilities of inconsistent samples are of little difference, but the Bayes error rates become distinct.

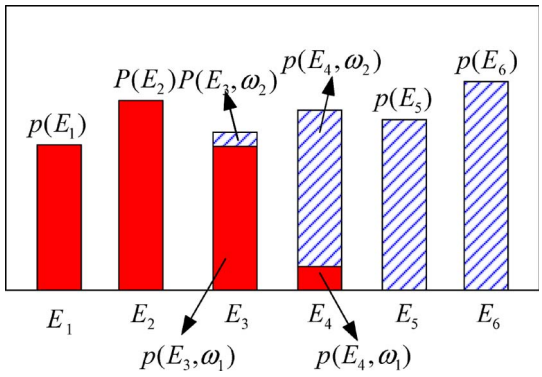


Fig. 3. Inconsistent discrete decision system.

Taking Fig. 4(c) as an example, we can observe that the region between x_0 and x_1 is the decision positive region of class ω_1, and the region between x_3 and x_4 is the decision positive region of class ω_2. The region between x_1 and x_3 is the decision boundary region. The inconsistency rate can be computed as

$$I = \sum_{i=1}^{2} \int_{x_1}^{x_3} p(\omega_i|x)dx.$$

However, the Bayes error rate is

$$e = \int_{x_1}^{x_2} p(\omega_2|x)dx + \int_{x_2}^{x_3} p(\omega_1|x)dx.$$

In general, $e \ll I$. Dependency is a good estimate of the inconsistency rate but not of the error rate. Now, we define the concept of neighborhood error rate. A neighborhood decision function $\text{ND}(x)$ is defined as follows.

Definition 6: Given NDT $= \langle U, C, D\rangle$, $x_i \in U$, $\delta(x_i)$ is the neighborhood of x_i, and $P(\omega_j|\delta(x_i))$, $j = 1, 2, \ldots, c$, is the class probability of class ω_j. The neighborhood decision of x_i is defined as $\text{ND}(x_i) = \omega_l$ if $P(\omega_l|\delta(x_i)) = \max_j P(\omega_j|\delta(x_i))$, where $P(\omega_j|\delta(x_i)) = n_j/K$, K is the number of samples in the neighborhood, and n_j is the number of samples with decision ω_j in $\delta(x_i)$.

$\text{ND}(x_i)$ is the class assigned to x_i according to the classification probability in the neighborhood of x_i. Obviously, $\text{ND}(x_i) = \omega(x_i)$ if x_i is the samples that belong to the lower approximation of decisions, where $\omega(x_i)$ is the real class of x_i; otherwise, we should compute the class probability in the neighborhood of x_i for giving a class label to x_i. $\text{ND}(x_i) \neq \omega(x_i)$ if majority of the samples in the neighborhood of x_i have different classes with x_i.

We introduce the following $0-1$ loss function for misclassified samples:

$$\lambda\left(\omega(x_i)|\text{ND}(x_i)\right) = \begin{cases} 0, & \omega(x_i) = \text{ND}(x_i) \\ 1, & \omega(x_i) \neq \text{ND}(x_i). \end{cases}$$

Definition 7: The NDER is defined as

$$\text{NDER} = \frac{1}{n} \sum_{i=1}^{n} \lambda\left(\omega(x_i)|ND(x_i)\right)$$

where n is the number of samples.

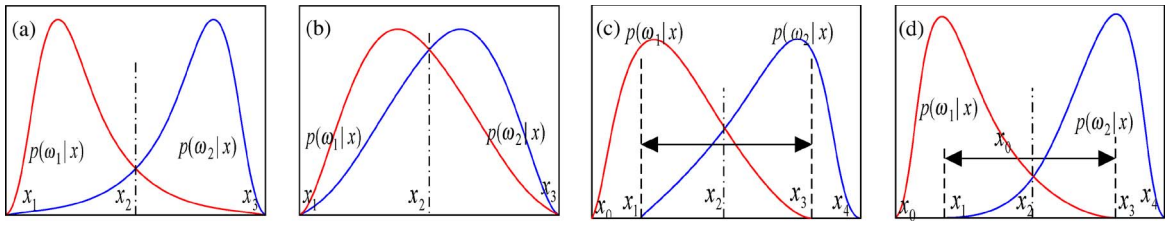


Fig. 4. Example of an inconsistent discrete system.

Theorem 3: Given NDT $= \langle U, C, D\rangle$, Δ is a metric on U, and $\delta \geq 0$ is a constant number. The following two conditions hold.

1) $\gamma_A(D) \leq 1 - \text{NDER}$.
2) $\gamma_A(D) = 1 - \text{NDER}$ if the NDT is consistent in the neighborhood approximation space. In this case, $\gamma_C(D) = 1$, and NDER $= 0$.

Proof: First, assuming that U is divided into subsets $D_1, D_2, \ldots, D_N$ by decision attribute D, $\forall x_i \in \text{POS}_C(D)$, $\exists D_j \in \{D_1, D_2, \ldots, D_N\}$ such that $\delta(x_i) \subseteq D_j$. $x_i \in D_j$; thus, $P(\delta(x_i)|D_j) = 1$, and $\text{ND}(x_i) = D_j$. Hence, $\text{ND}(x_i) = \omega(x_i)$. This result means that all the samples in the decision positive region have zero decision loss. Therefore, $\gamma_C(D)$ is not greater than $1 - \text{NDER}$. Second, if the NDT is consistent in the neighborhood approximation space, $\text{POS}_C(D) = U$, and $\gamma_C(D) = 1$. In this case, $\forall x_i \in U$, $\lambda(\omega(x_i)|\text{ND}(x_i)) = 0$; thus, NDER $= 0$. We get $\gamma_C(D) = 1 - \text{NDER}$.

For convenience, we call $1 - \text{NDER}$ a neighborhood recognition rate (NRR). The NDER is an estimate of the Bayes decision error. NDEM is an idea for feature selection by minimizing the NDER or maximizing the NRR in different feature subsets.

δ is a parameter for controlling the granularity of a neighborhood approximation space. Neighborhood decision errors vary with different resolution-granulated spaces. In essence, the NRR employs samples' neighborhoods to estimate the class probabilities of the local region. Therefore, a neighborhood can be considered as a Parzen window function that is used in problems of the probability density estimation. It is well known that the size of a window has great influence on the estimate; thus, it is important to select a proper value for parameter δ. We will discuss this problem in the experiment section. The NDER can be considered the Bayes error rate where the rate is estimated with a neighborhood window. We here estimate the local class probabilities with the distribution of samples in the neighborhood.

The following advantages are some properties of the NDEM strategy that was used in feature selection. First, the search algorithm based on NDEM can be used to deal with discrete and numerical data without discretizing numerical features, because the NDER is applicable to evaluating discrete and numerical features. Second, NDEM is robust to noisy samples. In real-world recognition tasks, learning samples are usually corrupted by various kinds of noise. NDEM , like the k-nearest neighbor (KNN) classifier, is robust to mislabeled samples [14], [45]. If there are some mislabeled samples, only these samples will be taken into account when computing the NDER. The samples around them will be recognized; thus, they will not be counted in computing neighborhood decision errors. Third, NDEM can approximate complex classification boundary regions. Similar to KNN and the neighborhood classifier (NEC) [14], NDEM can precisely localize the decision boundary region; thus, the classification complexity that was calculated with the NDER can reflect the real Bayes decision error rate. Unlike discretization, NRSs generate neighborhood granules of samples in a pointwise manner and then approximate the class local regions with these granules. We can thus approximate arbitrary complex nonlinear and multimodal class regions with the neighborhood granules if we specify a proper level of granularity. Moreover, compared with mutual information [3], [46] and the pattern recognition using information slicing method (PRISM) [38], it is easier to design a stopping criterion, because the NDER is linear with respect to the decision errors.

IV. Feature Selection Based on NDEM

Feature evaluation and optimal subset search are crucial to the overall process of feature selection. NDEM offers an idea for feature selection by minimizing the NDER. Here, the NDER is a measure, which is independent of a specific search algorithm, for evaluating the quality of feature subspaces.

There are a number of candidate search procedures. Greedy search strategies [17], [29], [43], [53] seem to be particularly computationally advantageous and robust against overfitting. They come in the following two variants: 1) sequentially forward selection (SFS) and 2) sequentially backward elimination (SBE). Both of them are suboptimal search procedures. One feature at a time is added to the current feature subset in the SFS. At each stage, the feature that will be included is selected from the remaining available features so that the new enlarged feature set yields a maximum value of the evaluation function that is used. The SBE starts with a set of all features and progressively eliminates the least promising ones. These algorithms may stop at some local minimum; thus, there is no guarantee that we can find an optimal solution. B&B [30] is a search technique in which all possible subsets are implicitly checked without running exhaustive search. B&B returns the optimal subset that is quantified in terms of the function if the feature evaluation function is monotonic. However, in the worst case, the B&B algorithm may exhibit exponential complexity. However, one may come up with some simple heuristics that, in practice, can result in substantial performance gains. Additional heuristics for facilitating further speedup of the B&B algorithm was proposed in [41]. Sequential forward floating search is also a near-optimal search procedure with computational complexity lower than the complexity of B&B. It performs sequential forward search with provision for backtracking. GA is a stochastic algorithm that mimics the genetic principles of natural evolution [37], [44]. The most distinct aspect of this algorithm is that it maintains a set of solutions (called individuals or chromosomes) in a population based on their fitness. Similar to the case of biological evolution, it has a

mechanism for selecting fitter chromosomes at each generation. To simulate the process of evolution, the selected chromosomes undergo genetic operations, e.g., crossover and mutation. The shortcomings in the design of GA-based feature selection are that it comes with a significant number of variations and parameters that need to properly be selected to achieve a reasonable or good performance of the optimization procedure.

In this paper, the objective is to design a sound evaluation function for feature selection and classification complexity estimation rather than to focus on search strategies. Hence, we introduce SFS to compare several feature evaluation functions with regard to their computational efficiency. The algorithm employs a bias of minimal features, i.e., it selects the minimal set of features to achieve the minimal classification complexity.

With the NRR, we define the significance of feature a relative to B as follows:

$$\text{SIG}(a, B, D) = \text{NRR}_{B\cup a}(D) - \text{NRR}_B(D)$$

where $\text{NDR}_B(D)$ stands for the NRR in feature subset spaces B.

Algorithm: SFS based on NDEM (**SFS−NDEM**)
Input: decision table $\langle U, C, d\rangle$; delta δ // Control the size of the neighborhood
Output: feature subset red.
1: $\varnothing \rightarrow$ red; // red is the pool for containing the selected features.
2: do while $C -$ red $\neq \varnothing$
3: for each $a_i \in C -$ red
4: Compute SIG(a_i, red, d)=NRR$_{\text{red}\cup a_i}(d)$–NRR$_{\text{red}}(d)$
5: end
6: select the attribute a_k that satisfies the condition:
7: SIG$(a_k, \text{red}, d) = \max_i(\text{SIG}(a_i, \text{red}, B))$
8: if SIG$(a_k, \text{red}, d) > 0$,
9: red $\cup\, a_k \rightarrow$ red
10: else
11: break
12: end
13: end
14: return red

There are several main steps in this reduction process. First, we rank the samples with each attribute and find the neighborhood of each sample with a sliding-window technique. We consider only the samples in the neighborhood window in searching the neighborhood. The time complexity of this step is $n \log n + kn$, where n is the number of samples, and k is a constant value that was used in searching the neighborhood of each sample. To compute the neighborhood of x in a multidimensional space, we use the intersection of the neighborhoods of x in each feature space. Then, we compute the class probability in neighborhoods whose time complexity is $O(n)$. Finally, we evaluate the remaining features and add the informative features into the reduct one by one. The overall complexity is $Nm(n \log n + kn + n)$ if there are N candidate features, and m features are selected. There exist some other strategies for speeding the algorithm up. For instance, ReliefF and RReliefF come with a time complexity of $Nn \log n$ [39].

The NRS model divides the samples into the following two parts: 1) the positive region and 2) the boundary region (boundary). Moreover, the boundary can also be classified into the following two subsets: 1) the set of samples that can correctly be recognized with the neighborhood information and 2) the remaining samples that cannot be recognized. With regard to the positive region, we have the following property.

Corollary 1: Given NDT $= \langle U, C, D\rangle$ and metric Δ, M, $N \subseteq C$, $M \subseteq N$, we have $x_i \in \text{POS}_N(D)$ if $x_i \in \text{POS}_M(D)$.

Corollary 1 shows that an object necessarily belongs to the positive region with respect to an attribute set if it belongs to the positive region with respect to its subset. In forward attribute selection, the attributes are added into the selected subset one by one according to their significance levels. Accordingly, we have $\forall M \supseteq B$, $x_i \in \text{POS}_M(D)$ if object $x_i \in \text{POS}_B(D)$. Therefore, we need not compute the objects in $\text{POS}_B(D)$ when computing $\text{POS}_M(D)$, because they are necessarily in $\text{POS}_M(D)$. In this case, we only need to discuss the objects in $U - \text{POS}_B(D)$. The objects in $U - \text{POS}_B(D)$ and the remaining features get much fewer as the attribute reduction goes on, and the computation size is reduced in selecting a new feature. Based on this observation, we can give a mark to the samples in the positive region with the current features and omit them in computing the significance of the rest features. For example, we just need to analyze 25% of the samples to compute the significance of $N-m$ features if 75% of the samples belong to positive regions with the selected m features. With this strategy, in some applications, the computation overhead is greatly reduced.

There are two stopping criteria for SFS−NDEM . The search stops if all candidate features have been selected or an inclusion of any new feature into the current subset does not reduce the NDER. In practice, this condition is very strong and restrictive, and it may result in overfitting. By applying SFS−NDEM to feature selection, we usually find that the NRR rapidly goes up when the first features have been added and then slowly rises until it completely stops. The last several features result in a very limited increase in the values of the recognition rates. NDEM employs the empirical risk minimization strategy; thus, the resulting feature subset may overfit the samples.

We can introduce a postpruning strategy to alleviate the effect of overfitting. In this case, feature selection consists of two steps. The SFS−NDEM searches a feature subset red such that $\forall a_k \in C -$ red : SIG$(a_k, \text{red}, D) = 0$ in the first step. In this step, it produces a series of nested feature subsets, i.e., $\text{red}_1 \subset \text{red}_2 \subset \cdots \subset \text{red}_m$. Then, we can introduce a learning algorithm to assess the quality of red_i one by one. The subset with the highest accuracy is then selected. This procedure can be understood as a feature selection algorithm that combines the filter and wrapper strategies [54].

V. Experimental Analysis

One of the important issues in feature selection is to estimate the classification complexity in different feature subspaces and rank the corresponding subsets of features, given this estimate. In this section, we will first test the influence of parameter δ on the estimate and get a good value domain for it. Then, we conduct NDEM-based feature selection and compare it with some existing techniques.

TABLE I
UCI DATA [FEATURES (CLASSES) SAMPLES] AND CLASSIFICATION COMPLEXITIES

	Dataset	F(C)-S	ASH	ASNN	ND(0.14)	NDEM(0.14)
1	balance	4(3)-625	0.96311	0.98793	0.5824	0.912
2	ecoli	7(8)-336	0.95398	0.98931	0.044643	0.87798
3	glass	9(7)-214	0.91079	0.97847	0.13551	0.68692
4	iono	34(2)-351	0.9875	0.99788	0.66667	0.96439
5	iris	4(3)-150	0.95935	0.99666	0.50667	0.95667
6	sonar	60(2)-208	0.98341	0.9972	0.79808	0.95433
7	wdbc	30(2)-569	0.98367	0.9984	0.18981	0.93322
8	wine	13(3)-178	0.98386	0.9992	0.86517	0.98876
9	yeast	8(10)-1484	0.84792	0.9463	0.0067385	0.42621
10	abalone	8(29)-4177	0.51613	0.81161	0.00023941	0.24611

TABLE II
AVERAGE ACCURACIES WITH TENFOLD CV USING FOUR CLASSIFIERS

Dataset	CART	SVM	KNN	NEC
balance	0.64814	0.88779	0.735	0.87206
ecoli	0.79615	0.85118	0.85082	0.86565
glass	0.53235	0.57908	0.65427	0.50511
iono	0.87546	0.93789	0.84116	0.7673
iris	0.96667	0.96667	0.95333	0.9600
sonar	0.72071	0.85095	0.81286	0.7931
wdbc	0.90501	0.98076	0.96839	0.95263
wine	0.86944	0.98333	0.96042	0.97153
yeast	0.5262	0.58202	0.54358	0.55084
abalone	0.19174	0.25727	0.22857	0.24709

TABLE III
CORRELATION COEFFICIENTS BETWEEN ACCURACIES AND CLASSIFICATION COMPLEXITIES

	CART	SVM	KNN	NEC
ASH	0.8726	0.9095	0.9154	0.8583
ASNN	0.8717	0.8964	0.9156	0.8484
ND	0.5426	0.6533	0.5667	0.5854
NDEM	**0.9074**	**0.9580**	**0.9433**	**0.9187**

A. Comparison of Different Complexity Measures

To compare the effectiveness of the estimates, ten data sets from the UCI Machine Learning Repository are considered (see Table I) [5]. We estimate their classification complexity by making use of the following four measures, as presented in Table I:

1) AS_H;
2) AS_{NN} [40];
3) ND;
4) NDEM .

We specify the size δ of the neighborhood to be 0.14; this particular value has been reported in the literature [49]. We also consider the following four well-known learning algorithms to estimate classification errors based on a tenfold cross validation [14]:

1) Classification and Regression Tree (CART);
2) Radial Base Function Support Vector Machine (RBF−SVM);
3) KNN;
4) NEC.

Intuitively, we could anticipate that the correlation coefficient between complexities and classification error rates should be high, which shows that the complexity is a sound estimate of the classification capabilities of the data. The classification complexity and classification accuracies are shown in Tables I and II, respectively. Then, we compute the values of the correlation coefficient between them, as given in Table III.

Table III shows the values of correlation between the classification accuracies that were obtained for four classifiers and the four complexity measures. NDEM consistently produces the highest values of correlation among these measures, whereas ND comes with the lowest correlation values. As pointed out in Section III, dependency reflects the ratio of samples in the boundary region over the whole sample set. However, not all the boundary samples are misclassified according to the Bayes rule. Therefore, dependency is not a good estimate of the classification complexity. Recall that NDEM is the percentage of the misclassified samples determined with the local information of samples. Based on Table III, we can conclude that NDEM is a better estimate of classification complexity than AS_H, AS_{NN}, and ND, and all the four coefficients assume values that were higher than 0.9. We also see that AS_H and AS_{NN} yield rather good estimates of complexity.

The values of ND and NDEM vary with respect to the size of the neighborhoods that were used in the computing boundary samples and misclassified samples. To show the influence of the values of parameter δ, we consider a series of numeric values, e.g., $\delta = 0.005, 0.006, 0.008, 0.01, 0.02, 0.04, 0.06, 0.08, 0.1, 0.12, 0.14, 0.16, 0.18, 0.20, 0.25, 0.30, 0.40,$ and compute the values of ND and NDEM for each size of neighborhood, respectively. The obtained values of the correlation coefficients between the classification accuracies and complexities are shown in Fig. 5. In Fig. 5, we can observe that the complexity of ND is much sensitive to the size of the neighborhood, whereas NDEM is rather robust. We can observe that a wide range of parameter δ yields good estimates of complexity with regard to NDEM . The size of the neighborhood can be regarded as the granularity of classification; thus, we can conclude that NDEM is robust to the granularity of the complexity analysis.

In Fig. 5, we can also observe that the four classifiers share a similar variation in correlation coefficients, which means that the estimates of complexity are independent of the classifiers that were used, and NDEM can be treated as a general measure of classification complexity. With the experimental results, we can also see that [0.1, 0.2] is an appropriate domain for parameter δ.

B. Comparison of Feature Selection Algorithms on Discrete Data

In what follows, we experimented with different algorithms of feature selection using 16 data sets, as outlined in Table IV, where four sets come with discrete features (i.e., lymphography, soybean, votes, and zoo), six data sets come with numerical features (i.e., Diab, Iono, sonar, WDBC, WPBC, and wine), and the rest of the data sets come with mixed numerical and categorical features. The last two columns show the classification rates that were obtained for the raw data sets without feature selection. Next, we use sequentially greedy forward search to form the best features when we compare the algorithms that evaluate features based on RSs, information entropy, NRSs, and consistency, respectively. That is, we only replace the evaluation function in Line 6 of the SFS−NDEM algorithm. With regard to correlation-based feature selection (CFS), ReliefF, and the support vector machine (SVM)-based algorithm, they have special search strategies, and we keep their search strategies in the experiments.

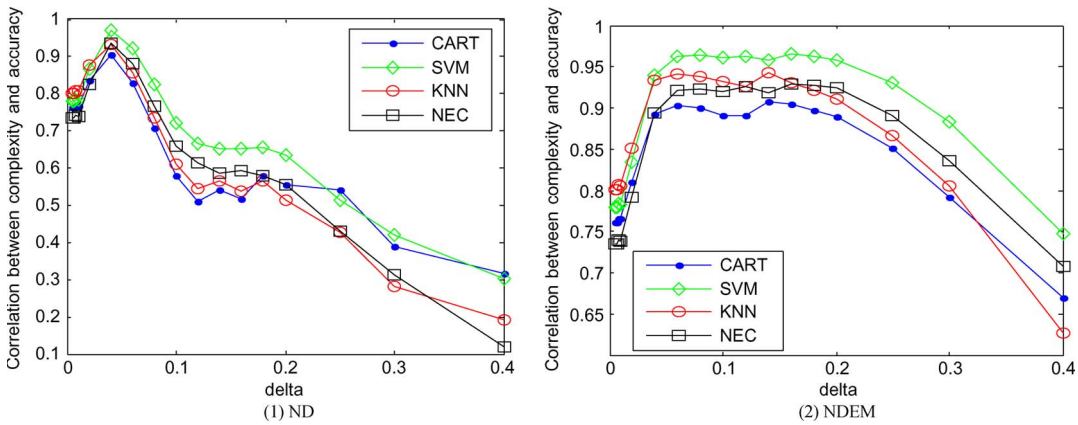


Fig. 5. Variation of correlation between accuracies and complexities with different sizes of neighborhoods. (a) ND. (b) NDEM.

TABLE IV
DATA DESCRIPTION

	Data	Samples	Numerical	Categorical	Class	CART	SVM
1	Anneal	798	6	32	5	99.89±0.35	99.89±0.35
2	Credit	690	6	9	2	82.73±14.86	81.44±7.18
3	Diab	768	8	0	2	72.27±5.12	77.47±4.30
4	Ecoli	336	5	2	7	81.97±4.44	85.12±5.91
5	Heart	270	7	6	2	74.07±6.30	81.11±7.50
6	Hepatitis	155	6	13	2	91.00±5.45	83.50±5.35
7	Horse	368	7	15	2	95.92±2.30	72.30±3.63
8	Iono	351	34	0	2	87.55±6.93	93.79±5.08
9	lymphography	148	0	18	4	69.94±21.95	56.23±5.83
10	Sonar	208	60	0	2	72.07±13.94	85.10±9.49
11	Soybean	683	0	35	19	91.84±5.21	52.85±6.35
12	Votes	435	0	16	2	96.50±3.04	93.07±5.00
13	WDBC	569	31	0	2	90.50±4.55	98.08±2.25
14	WPBC	198	33	0	2	70.63±7.54	80.37±5.33
15	Wine	178	13	0	3	89.86±6.35	98.89±2.34
16	Zoo	101	0	16	7	90.65±9.13	86.15±9.01

TABLE V
NUMBER OF FEATURES THAT WERE SELECTED WITH DIFFERENT MEASURES

Data	Raw data	Entropy	RS	VPRS	RDEM	Post-pruning
lymphography	18	6	6	7	7	2
soybean	35	13	16	16	13	7
votes	16	11	0	1	1	1
zoo	16	5	5	5	5	4
average	**21.25**	**8.75**	**--**	**7.25**	**6.50**	**3.50**

We first show the results of discrete feature selection. The number of selected features of the discrete data is given in Table V, where the second column is the number of features of the raw data sets. Entropy means mutual-information-based feature selection, whereas VPRS denotes feature selection based on variable precision RS. To distinguish numerical feature selection, we denote the NDEM algorithm by NDEM. NDEM is denoted as RDEM if there are only discrete features or numerical features that have been discretized before feature selection. In this case, RDEM is equivalent to the consistency-based feature selection [7]. There is a parameter that will be specified in the VPRS-based feature selection, i.e., α, whose intent is to control the approximation precision. We specify $\alpha = 0.8$ based on the suggestion in [4].

In Tables V–VII, we observe that most of the features in raw data have been deleted by all the feature selection algorithms. At the same time, there is no remarkably large deterioration in the classification performance. The results show that these algorithms are effective in retaining classification abilities. However, the RS-based algorithm yields an empty set when it is applied to the "votes" data, because no equivalence class is consistent at the first stage. In this case, the dependency of each single feature is zero. Therefore, the algorithm stops here. Noisy information has great influence on the results that were produced by RS-based algorithms. VPRS introduces a relaxing parameter to control the noise effect. RDEM further generalizes the idea to deal with noise with neighborhood decision. The noisy sample has little influence on the decisions of proximate samples. Furthermore, there is no parameter that will be specified with RDEM.

The last columns of Tables V–VII show the number of features and classification performance after postpruning. The number of the selected features has largely been reduced, and at the same time, the classification performance improved with data reduction.

C. Comparison on Numerical or Mixed Data and Overfitting Problems

Fig. 6 shows the variation of attribute significance with the number of selected features obtained for the soybean data. The significance of feature subset is computed with the RS-based dependency, VPRS-based dependency, and RDEM, respectively. The values of the RDEM rapidly increase before the set of features is formed by six features. Then, the increase slows down until it completely stops when the set of features has 13 elements. In fact, only a few samples are recognized when adding the remaining features, i.e., increasing the size of the feature space. In particular, the NRR achieves a value of 99.27% when we have included the tenth feature. Adding other three features contributes to further improvements to a very limited extent by distinguishing the remaining 0.073% of the samples. A similar behavior happens to VPRS and RS. We also observe the relationship RDEM $\geq$ VPRS $\geq$ RS.

Fig. 7 shows the relation between the number of selected features and the classification performance. In the beginning, the recognition rate steeply climbs to some maximum value and then decreases. This case shows that the features that were selected after the peak are superfluous for classification, although they augment the evaluation function. Their impact is negative, and they should be eliminated from the selected feature subsets.

Now, we test the algorithm on data sets with numerical or mixed features. We compare the proposed algorithm with some classical algorithms. The entropy and classical RS-based algorithms can be used to deal with discrete features; thus, we introduce the minimum descriptive length (MDL) discretization [24] to segment the numerical features into several intervals and

TABLE VI
CLASSIFICATION ACCURACIES OF FEATURES THAT WERE SELECTED WITH DIFFERENT MEASURES BASED ON CART (IN PERCENT)

Data	Raw data	Entropy	RS	VPRS	RDEM	Post-pruning
lymphography	69.94±21.95	68.25±18.22	68.25±18.22	72.79±23.46	68.25±20.00	74.22±22.58
soybean	91.84±5.21	88.01±6.23	85.60±6.64	87.75±6.07	88.01±6.23	89.66±6.12
votes	96.50±3.04	96.28±3.38	--	96.27±3.63	96.27±3.63	96.27±3.63
zoo	90.65±9.13	92.76±9.87	92.76±9.87	92.76±9.87	92.76±9.87	92.76±9.87
average	**87.23**	**86.33**	**--**	**87.39**	**86.32**	**88.23**

TABLE VII
CLASSIFICATION ACCURACIES OF FEATURES THAT WERE SELECTED WITH DIFFERENT MEASURES BASED ON SVM (IN PERCENT)

Data	Raw data	Entropy	RS	VPRS	RDEM	Post-pruning
lymphography	56.23±5.83	84.48±9.40	84.48±9.40	77.86±11.39	74.03±8.33	78.31±11.74
soybean	52.85±6.34	61.43±6.86	59.51±7.21	61.14±7.49	61.43±6.86	90.97±5.51
votes	93.07±5.00	96.26±3.79	--	96.26±3.63	96.27±3.63	96.27±3.63
zoo	86.15±9.01	92.39±9.24	92.39±9.24	92.39±9.24	92.39±9.24	92.39±9.24
average	**72.08**	**83.64**	**--**	**81.92**	**81.03**	**89.49**

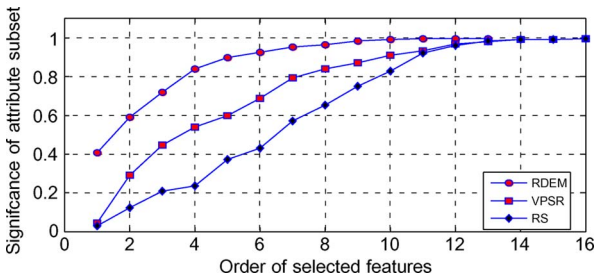


Fig. 6. Attribute significance versus the number of selected features.

form the discretized data sets. Then, feature selection based on mutual information, RS, and RDEM is employed to select discretized features. Meanwhile, we also apply ND and NDEM to directly select numerical features, where numerical features are normalized into the unit interval. We set $\delta = 0.14$ (refer to the discussion on the selection of the numeric value of this parameter).

The selected features with different evaluation functions are presented on the order of selecting in Table VIII, where MDL is the discretization algorithm; entropy, RS, RDEM, NRS, and NDEM are algorithms that were used for feature selection. The MDL+RS feature selection algorithm yields two empty sets for the heart and sonar data sets. This result states that, for these data sets, no sample is consistent in terms of a single feature if the numerical features are discretized with the MDL algorithm. However, all the other feature selection algorithms determine a subset of features. We can also find that the selected features are distinct when applying different algorithms.

The "classical" RS-based dependency, RDEM, NRS-based dependency, and NDEM reflect ratios of consistent samples or recognized samples, respectively. They take values in interval [0, 1].

Fig. 8 visualizes a change of significance with respect to the number of selected features. All four significance functions climb fast at the beginning of the selection process, and this phenomenon occurs for the heart, Iono, sonar, WDBC, and wine data sets. The feature significance of credit data slowly increases, and this result constitutes a different pattern of behavior compared with the four other data sets. Feature selection algorithms may stop very early if we specify a threshold to stop the search in this case.

Here, we show a technique that combines the idea of filter with wrapper to overcome this problem. That is, we first select the relevant features that were evaluated with significance in the forward selection step. We then employ a learning algorithm to evaluate the selected features by tenfold cross validation, where the selected features are added to the learning algorithm one by one on the order of selection. The results are presented in Tables IX and X. The results in Table IX are evaluated with CART, and the results in Table X are evaluated with SVM, where A is the optimal classification performance, and N is the corresponding feature number. We can find that there is little difference in the results if the entropy, RDEM, and NDEM algorithms are in the optimal states, respectively. However, entropy-based algorithms require discretizing numerical features in advance. The results also show that wrapper-based postpruning is necessary for feature selection. A number of features that were selected with filter techniques are excluded in the final results. The numbers of features in optimal subsets are greatly reduced in most of the cases. In addition, the optimal number of features varies from one learning algorithm to another. No general conclusion is applicable to various learning algorithms. It is efficient to use a filter in selecting a candidate subset for a subsequent wrapper. Therefore, this method integrates the advantage of a filter in efficiency with that of a wrapper in effectiveness.

Tables XI–XIII show the number of selected features and the corresponding classification performance based on fuzzy entropy [13], [15], NRS-based dependency, NDEM , the combined filter and wrapper (selection and postpruning) algorithm, CFS [11], ReliefF [37], and the SVM-based technique [10], respectively, where feature subsets are directly selected from numerical data. P-CART and P-SVM denote the number of features that were selected using NDEM + CART or NDEM + SVM (see Table XI). Among the 12 data and eight algorithms of feature selection, P-CART comes with the minimal number of features (which occurs for seven data sets), whereas P-SVM returns the minimal number of features as far as six data sets are concerned. On the average, P-CART and P-SVM select 5.17 and 5.25 features for classification, which are the least two values among the size of features that the eight algorithms offered.

With regard to the performance of CART-based classification, as shown in Table XII, P-CART, i.e., NDEM + CART, comes with the highest accuracy in six cases. At the same

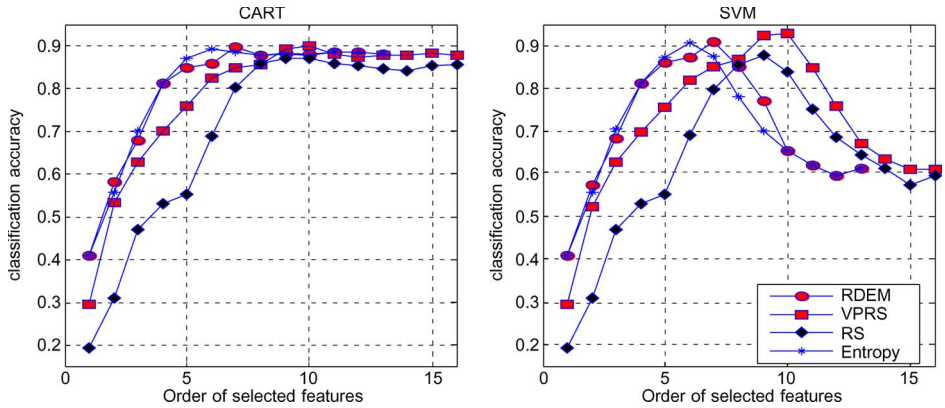


Fig. 7. Variation of average classification accuracies with the number of selected features.

TABLE VIII
FEATURES THAT WERE SEQUENTIALLY SELECTED WHEN USING DIFFERENT SIGNIFICANCE CRITERIA

Data	MDL+ entropy	MDL+RS	MDL+RDEM	NRS	NDEM
credit	9, 11, 15, 6, 14, 4, 8, 3, 1, 2, 7	4, 7, 9, 15, 1, 3, 11, 6, 14, 8, 2	9, 4, 10, 15, 14, 2, 6, 8, 3, 1, 11	1, 8, 6, 9, 7, 10, 12, 4, 2, 11, 13, 3	9, 13, 6, 10, 7, 1, 2, 12, 3, 4, 8
heart	13, 12, 3, 11, 1, 7, 2, 9, 8, 6, 10		13, 12, 3, 11, 7, 1, 8, 10, 2, 9, 6	10, 12, 13, 3, 11, 7, 1, 2, 8, 4, 6, 5	13, 12, 3, 11, 7, 4, 1, 8, 2, 5, 6
iono	5, 6, 3, 7, 33, 27, 12, 13	5, 3, 6, 34, 17, 14, 22, 4	34, 5, 17, 4, 13, 8, 14, 7	1, 5, 27, 28, 8, 24, 7, 31, 34, 25, 30, 26, 11, 3, 16, 32, 18, 22	5, 4, 3, 20, 34, 8, 30, 10
sonar	11, 4, 45, 36, 52, 47, 21, 13, 35, 51, 28, 5, 9, 49		11, 4, 36, 45, 46, 5, 20, 48, 9, 47, 13, 10, 21, 52	58, 1, 45, 35, 21, 27, 12, 29, 54, 16, 55, 22, 10, 33, 48, 34	11, 16, 25, 22, 47, 32, 38, 53, 30, 1, 10
wdbc	23, 28, 22, 14, 25, 9, 17	24, 8, 22, 26, 13, 5, 14	21, 27, 28, 22, 29, 13, 7	23, 28, 22, 12, 19, 1, 9, 25, 10, 29, 8, 16, 21, 2, 27, 7, 5, 11, 15, 4, 6, 18, 26	28, 21, 22, 2, 8, 11
wine	7, 1, 10, 2, 4	10, 13, 7, 2	7, 1, 10, 2, 4	13, 10, 7, 1, 5, 2, 11, 3, 8	7, 1, 11, 13, 3, 10

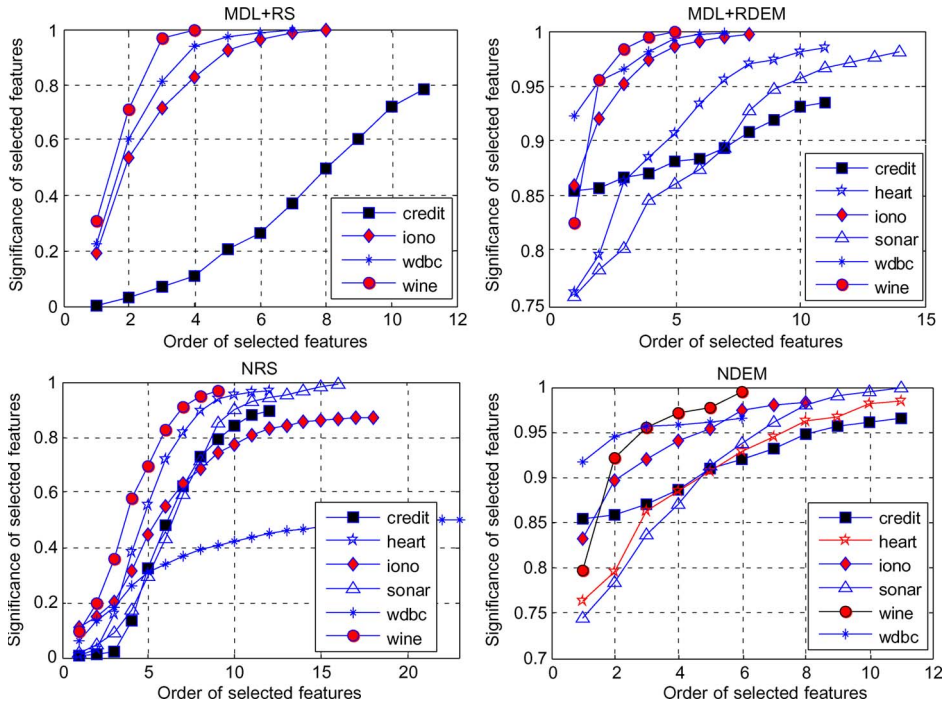


Fig. 8. Evaluation functions versus the number of selected features.

time, P-SVM comes with the highest accuracy in seven cases. By scanning the results in Tables XI–XIII, we conclude that NDEM + CART and NDEM + SVM give rise to the highest feature reduction while retaining classification performance that is comparable to the one reported for some well-known algorithms.

VI. CONCLUSION AND FUTURE WORK

A novel feature evaluation measure that is applicable to discrete and continuous features has been proposed in this paper. We first introduced the NRS model to define and compute decision positive regions and decision boundary in metric

TABLE IX
CLASSIFICATION PERFORMANCE OF OPTIMAL FEATURES FILTER + WRAPPER (CART)

Data	MDL+ entropy		MDL+RS		MDL+RDEM		NRS		NDEM	
	N	A	N	A	N	A	N	A	N	A
credit	1	0.8548	3	0.8534	1	0.8548	9	0.8217	1	0.8548
heart	3	0.8519	0	0	4	0.8259	4	0.8000	3	0.8519
iono	3	0.9181	8	0.9318	6	0.8979	6	0.8983	3	0.8925
sonar	6	0.7933	0	0	5	0.7693	8	0.7648	8	0.7643
wdbc	4	0.9439	4	0.9508	6	0.9420	4	0.9420	2	0.9315
wine	4	0.9437	3	0.9208	4	0.9437	4	0.9208	4	0.9437

TABLE X
CLASSIFICATION PERFORMANCE OF OPTIMAL FEATURES THAT WERE SELECTED WITH FILTER + WRAPPER (SVM)

Data	MDL+ entropy		MDL+RS		MDL+RDEM		NRS		NDEM	
	N	A	N	A	N	A	N	A	N	A
credit	1	0.8548	3	0.8476	1	0.8548	4	0.8505	6	0.8563
heart	3	0.8556	0	0	3	0.8556	4	0.8593	3	0.8556
iono	7	0.9353	8	0.9154	7	0.9262	24	0.9546	7	0.9345
sonar	12	0.8269	0	0	10	0.8276	8	0.8121	11	0.8269
wdbc	5	0.9650	3	0.9614	6	0.9667	18	0.9790	5	0.9650
wine	6	0.9556	4	0.9722	5	0.9556	6	0.9833	6	0.9944

TABLE XI
NUMBER OF SELECTED MIXED FEATURES

data	Raw data	F-entropy	NRS	NDEM	P-CART	P-SVM	CFS	ReliefF	SVMEval
anneal	38	3	12	3	3	3	5	13	13
credit	15	8	12	7	1	1	8	5	5
diab	8	8	8	7	6	6	4	3	3
ecoli	7	7	7	7	6	5	5	2	2
heart	13	11	12	11	3	3	10	4	4
hepatitis	19	11	11	10	10	4	6	6	6
horse	22	7	8	7	2	2	7	7	7
iono	34	13	18	8	3	3	4	11	11
sonar	60	12	16	11	8	11	9	20	20
wdbc	31	17	23	6	2	5	6	10	10
wpbc	33	19	20	15	12	14	3	11	11
wine	13	9	9	6	6	6	5	4	4
Average	**24.42**	**10.42**	**13.00**	**8.17**	**5.17**	**5.25**	**6**	**8**	**8**

TABLE XII
CLASSIFICATION ACCURACY (IN PERCENT) OF MIXED DATA THAT WERE OBTAINED FOR CART CLASSIFIERS

data	Raw data	F-entropy	NRS	NDEM	P-CART	CFS	ReliefF	SVMEval
anneal	99.89±0.35	**100.00±0.0**	**100.00±0.0**	**100.00±0.0**	**100.00±0.00**	96.33±1.81	**100.00±0.00**	**100.00±0.00**
credit	82.17±4.59	72.13±4.04	83.02±14.4	82.31±12.6	**85.48±18.51**	81.43±12.50	85.20±17.8	80.99±14.50
diab	72.27±5.12	72.40±4.85	72.40±4.85	72.67±4.93	**72.80±6.55**	70.45±4.26	72.52±3.09	66.79±5.60
ecoli	81.97±4.44	81.68±4.29	81.68±4.29	81.68±4.29	81.68±4.29	**81.78±5.44**	56.13±11.3	73.09±4.12
heart	74.07±6.30	74.44±6.16	74.44±6.16	74.81±6.72	**85.19±6.30**	74.81±7.37	82.59±5.53	80.00±7.65
hepatitis	91.00±5.45	90.33±4.57	91.00±5.45	89.00±7.71	89.00±7.71	91.00±5.45	86.17±5.45	**92.33±6.68**
horse	95.92±2.30	89.38±4.76	88.87±5.57	89.38±4.76	91.82±3.93	93.48±4.06	**96.47±1.30**	89.70±6.07
iono	87.55±6.93	**90.67±5.64**	87.26±5.15	88.94±3.92	89.25±5.48	90.01±5.41	90.64±5.23	88.65±4.33
sonar	72.07±13.9	71.60±8.67	73.55±7.22	74.52±5.52	**76.43±7.43**	70.17±12.36	71.60±9.02	75.45±11.88
wdbc	90.50±4.55	91.93±3.18	92.27±2.62	92.29±4.28	93.15±2.67	**94.56±3.45**	93.15±3.15	92.26±3.03
wpbc	69.63±8.26	71.33±6.34	69.03±10.4	69.10±9.97	**72.16±9.15**	71.58±12.87	70.13±10.7	71.68±5.60
wine	89.86±6.35	90.97±6.05	**91.53±6.09**	90.97±6.05	90.97±6.05	91.46±6.87	84.31±6.21	**91.53±4.83**
Ave.	**83.91**	**83.07**	**83.75**	**83.81**	**85.66**	**83.92**	**82.41**	**83.54**

TABLE XIII
CLASSIFICATION ACCURACY (IN PERCENT) OF MIXED DATA THAT WERE OBTAINED FOR SVM CLASSIFIERS

data	Raw data	F-entropy	NRS	NDEM	P-SVM	CFS	ReliefF	SVMEval
anneal	99.89±0.35	99.89±0.35	99.89±0.35	99.89±0.35	**100.00±0.00**	93.88±0.78	**100.00±0.00**	**100.00±0.00**
credit	81.44±7.18	77.47±4.30	81.44±7.18	81.44±7.18	85.48±18.51	85.48±18.5	85.48±18.51	**85.81±8.51**
diab	77.47±4.30	77.47±4.30	77.47±4.30	77.35±3.13	**78.26±4.13**	76.56±3.68	76.31±3.29	76.18±3.20
ecoli	85.12±5.91	85.12±5.91	85.12±5.91	84.24±6.96	**85.54±3.17**	84.35±3.34	63.75±6.18	76.69±5.50
heart	81.11±7.50	81.11±7.70	81.48±7.61	82.22±5.47	**85.56±6.16**	84.44±6.00	83.33±6.36	82.22±7.77
hepatitis	83.50±5.35	87.50±7.51	83.33±8.46	84.50±9.53	86.50±6.31	**91.50±6.40**	82.33±7.04	85.67±7.04
horse	72.30±3.62	87.24±4.74	88.86±2.99	87.24±4.74	91.82±3.93	90.76±4.82	92.94±4.96	**93.78±4.57**
iono	93.79±5.07	94.62±3.65	95.46±3.10	92.64±5.56	**92.67±4.55**	87.84±5.39	81.88±7.32	90.63±4.82
sonar	85.10±9.48	82.71±9.01	79.31±8.04	76.93±3.08	**82.69±10.96**	76.52±7.10	72.71±8.80	82.64±7.74
wdbc	98.08±2.25	97.02±2.48	**97.90±2.15**	95.80±2.49	96.50±2.60	96.32±2.26	97.02±1.66	97.37±2.22
wpbc	77.79±4.20	**80.90±5.98**	80.87±6.01	76.32±3.04	79.87±5.08	76.32±3.04	76.32±3.04	77.37±5.14
wine	98.89±2.34	98.33±2.68	98.33±2.68	**99.44±1.76**	**99.44±1.76**	95.49±3.54	84.31±5.63	95.49±3.54
Average	**86.21**	**87.45**	**87.46**	**86.50**	**88.69**	**86.62**	**83.03**	**86.99**

spaces. The size of classification boundary was used to reflect on the classification complexity in NRSs. We showed that not all samples but only the samples in minority classes in boundary regions become misclassified. This measure of dependency in RSs is not an unbiased estimation of classification complexity. Decision boundaries lead to the following two subsets based on the class distribution of neighborhoods: 1) recognized parts and 2) misclassified parts. Finally, the ratio of the misclassified samples is taken as the estimate of complexity. There are some potential advantages of using the strategy of NDEM for feature selection. First, the NDER is applicable to discrete and continuous features; thus, the search algorithm based on NDEM can be used to deal with mixed data without discretizing numerical features. Second, NDEM is robust to outliers. NDEM , like the KNN classifier, is robust to mislabeled samples. If there are some mislabeled samples, only these samples are taken into account in the computation of the NDER, whereas the samples around them may be recognized. However, dependency considers all these samples as being uncertain. Third, NDEM can approximate complex classification boundary regions. NDEM considers the local samples of feature spaces, which leads the model to compute the nonlinear classification boundary.

We have presented a forward greedy strategy for searching feature subsets to minimize the neighborhood decision error and, correspondingly, minimize the classification complexity in the selected feature subspaces. We compared the proposed algorithm with some classical algorithms, e.g., mutual information, CFS, ReliefF, and SVM-based feature selection. The results show that the proposed algorithm is effective when dealing with discrete data, numerical data, and their mixtures. We showed the phenomenon of overfitting, which may occur in forward feature selection. We integrated the filter-with-wrapper strategy and showed how one can delete the superfluous features by the postpruning technique. Experiments showed that the classification performance consistently increases with the pruned features.

Margin, which was first used to estimate the structure risk of classification in the statistical learning theory, was introduced in the research works of Gilad−Bachrachy *et al.* and Sun to evaluate the quality of a set of candidate features. The experiments showed that these techniques are promising. According to the statistical learning theory, we know that classification risks depend on empirical risks and the complexity of the learning machine. The latter factor can roughly be estimated by the margin of classification. Therefore, margin should be integrated with the classification error rate in evaluating the quality of features. This topic may offer a new direction for future studies.

REFERENCES

[1] S. Abe, R. Thawonmas, and Y. Kobayashi, "Feature selection by analyzing class regions approximated by ellipsoids," *IEEE Trans. Syst., Man, Cybern. C: Appl. Rev.*, vol. 28, no. 2, pp. 282–287, May 1998.

[2] H. Almuallim and T. G. Dietterich, "Learning Boolean concepts in the presence of many irrelevant features," *Artif. Intell.*, vol. 69, no. 1/2, pp. 279–305, Sep. 1994.

[3] 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.

[4] M. Beynon, "Reducts within the variable precision rough sets model: A further investigation," *Eur. J. Oper. Res.*, vol. 134, no. 3, pp. 592–605, Nov. 2001.

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

[6] R. Caruana and D. Freitag, "Greedy attribute selection," in *Proc. Int. Conf. Mach. Learn.*, 1994, pp. 28–36.

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

[8] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*, 2nd ed. Hoboken, NJ: Wiley, 2001.

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

[10] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik, "Gene selection for cancer classification using support vector machines," *Mach. Learn.*, vol. 46, no. 1–3, pp. 389–422, Jan. 2002.

[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] 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.

[13] Q. H. Hu, D. R. Yu, and Z. X. Xie, "Information-preserving hybrid data reduction based on fuzzy-rough techniques," *Pattern Recognit. Lett.*, vol. 27, no. 5, pp. 414–423, Apr. 2006.

[14] Q. H. Hu, D. R. Yu, and Z. X. Xie, "Neighborhood classifier," *Expert Syst. Appl.*, vol. 34, no. 2, pp. 866–876, Feb. 2008.

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

[16] Q. H. Hu, H. Zhao, Z. X. Xie, and D. R. Yu, "Consistency-based attribute reduction," in *Proc. PAKDD*, Z.-H. Zhou, H. Li, and Q. Yang, Eds., Berlin, Germany: Springer-Verlag, 2007, vol. 4226, LNAI, pp. 96–107.

[17] X. Hu and N. Cercone, "Learning in relational databases: A rough set approach," *Comput. Intell.*, vol. 11, no. 3, pp. 323–338, 1995.

[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, Dec. 2004.

[19] K. Kira and L. A. Rendell, "The feature selection problem: Traditional methods and a new algorithm," in *Proc. AAAI*, San Jose, CA, 1992, pp. 129–134.

[20] R. Kohavi, "Feature subset selection as search with probabilistic estimates," in *Proc. AAAI Fall Symp. Relevance*, 1994, pp. 122–126.

[21] R. Kohavi and B. Frasca, "Useful feature subsets and rough set reducts," in *Proc. Int. Workshop RSSC*, 1994, pp. 310–317.

[22] R. Kohavi and G. H. John, "Wrappers for feature subset selection," *Artif. Intell.*, vol. 97, no. 1/2, pp. 273–324, Dec. 1997.

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

[26] H. Liu, F. Hussain, and M. Dash, "Discretization: An enabling technique," *Data Mining Knowl. Discovery*, vol. 6, no. 4, pp. 393–423, Oct. 2002.

[27] H. Liu and H. Motoda, *Feature Selection for Knowledge Discovery and Data Mining*. Boston, MA: Kluwer, 1998.

[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] 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.

[30] P. M. Narendra and K. Fukunaga, "A branch-and-bound algorithm for feature subset selection," *IEEE Trans. Comput.*, vol. C-26, no. 9, pp. 917–922, Sep. 1977.

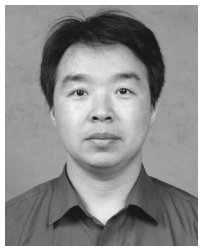
[31] J. Neumann, C. Schnorr, and G. Steidl, "Combined SVM-based feature selection and classification," *Mach. Learn.*, vol. 61, no. 1–3, pp. 129–150, Nov. 2005.

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

[33] Z. Pawlak, "Rough set," *Commun. ACM*, vol. 38, no. 11, pp. 88–95, Nov. 1995.

[34] 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.

[35] P. Pudil, J. Novovičová, and J. Kittler, "Floating search methods in feature selection," *Pattern Recognit. Lett.*, vol. 15, no. 11, pp. 1119–1125, Nov. 1994.
[36] S. J. Raudys and A. K. Jain, "Small-sample-size effects in statistical pattern recognition: Recommendations for practitioners," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 13, no. 3, pp. 252–264, Mar. 1991.
[37] 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.
[38] S. Singh, "PRISM: A novel framework for pattern recognition," *Pattern Anal. Appl.*, vol. 6, no. 2, pp. 134–149, Jun. 2003.
[39] M. R. Sikonja and I. Kononenko, "Theoretical and empirical analysis of ReliefF and RReliefF," *Mach. Learn.*, vol. 53, no. 1/2, pp. 23–69, Oct./Nov. 2003.
[40] S. Singh, "Multiresolution estimates of classification complexity," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 25, no. 12, pp. 1534–1539, Dec. 2003.
[41] P. Somol, P. Pudil, and J. Kittler, "Fast branch-and-bound algorithms for optimal feature selection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 26, no. 7, pp. 900–912, Jul. 2004.
[42] R. Thawonmas and S. Abe, "A novel approach to feature selection based on analysis of class regions," *IEEE Trans. Syst., Man, Cybern. B: Cybern.*, vol. 27, no. 2, pp. 196–207, Apr. 1997.
[43] K. Torkkola, "Feature extraction by nonparametric mutual information maximization," *J. Mach. Learn. Res.*, vol. 3, no. 7/8, pp. 1415–1438, Mar. 2003.
[44] J. Yang and H. Vasant, "Feature subset selection using a genetic algorithm," *IEEE Intell. Syst.*, vol. 13, no. 2, pp. 44–49, Mar./Apr. 1998.
[45] H. Wang, "Nearest neighbors by neighborhood counting," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 6, pp. 942–953, Jun. 2006.
[46] H. Wang, D. Bell, and F. Murtagh, "Axiomatic approach to feature subset selection based on relevance," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 21, no. 3, pp. 271–277, Mar. 1999.
[47] L. Yu and H. Liu, "Efficient feature selection via analysis of relevance and redundancy," *J. Mach. Learn. Res.*, vol. 5, pp. 1205–1224, Dec. 2004.
[48] L. A. Zadeh, "Toward a theory of fuzzy information granulation and its centrality in human reasoning and fuzzy logic," *Fuzzy Sets Syst.*, vol. 90, no. 2, pp. 111–127, Sep. 1997.
[49] Q. H. Hu, D. Yu, J. F. Liu, and C. Wu, "Neighborhood-rough-set-based heterogeneous feature subset selection," *Inf. Sci.*, vol. 178, no. 18, pp. 3577–3594, Sep. 2008.
[50] H. J. Shin and S. Z. Cho, "Invariance of neighborhood relation under input space to feature space mapping," *Pattern Recognit. Lett.*, vol. 26, no. 6, pp. 707–718, May 2005.
[51] E. Schmitt, V. Bombardier, and L. Wendling, "Improving fuzzy rule classifier by extracting suitable features from capacities with respect to the Choquet integral," *IEEE Trans. Syst., Man, Cybern. B, Cybern.*, vol. 38, no. 5, pp. 1195–1206, Oct. 2008.
[52] Y. Pan and S. A. Billings, "Neighborhood detection for the identification of spatiotemporal systems," *IEEE Trans. Syst., Man, Cybern. B, Cybern.*, vol. 38, no. 3, pp. 846–854, Jun. 2008.
[53] T. W. S. Chow, P. Y. Wang, and E. W. M. Ma, "A new feature selection scheme using a data distribution factor for unsupervised nominal data," *IEEE Trans. Syst., Man, Cybern. B, Cybern.*, vol. 38, no. 2, pp. 499–509, Apr. 2008.
[54] Z. X. Zhu, Y. S. Ong, and M. Dash, "Wrapper-filter feature selection algorithm using a memetic framework," *IEEE Trans. Syst., Man, Cybern. B: Cybern.*, vol. 37, no. 1, pp. 70–76, Feb. 2007.
[55] 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.
[56] Z. Pawlak, S. K. M. Wong, and W. Ziarko, "Rough sets: Probabilistic versus deterministic approach," *Int. J. Man-Mach. Stud.*, vol. 29, no. 1, pp. 81–95, Jul. 1988.
[57] Y. Y. Yao, "Relational interpretations of neighborhood operators and rough set approximation operators," *Inf. Sci.*, vol. 111, no. 1–4, pp. 239–259, Nov. 1998.
[58] W.-Z. Wu and W.-X. Zhang, "Neighborhood operator systems and approximations," *Inf. Sci.*, vol. 144, no. 1–4, pp. 201–217, Jul. 2002.
[59] D. Slezak, "Degrees of conditional (in)dependence: A framework for approximate Bayesian networks and examples related to the rough-set-based feature selection," *Inf. Sci.*, vol. 179, no. 3, pp. 197–209, Jan. 2009.
[60] D. Slezak, "Approximate entropy reducts," *Fundam. Inform.*, vol. 53, no. 3/4, pp. 365–390, May 2002.
[61] Y. Y. Yao and Y. Zhao, "Attribute reduction in decision-theoretic rough set models," *Inf. Sci.*, vol. 178, no. 17, pp. 3356–3373, Sep. 2008.
[62] 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, 2004, pp. 43–50.
[63] Y. J. Sun, "Iterative RELIEF for feature weighting: Algorithms, theories, and applications," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 6, pp. 1035–1051, Jun. 2007.



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 Associate Professor with Harbin Institute of Technology. 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, and rough-set theory.

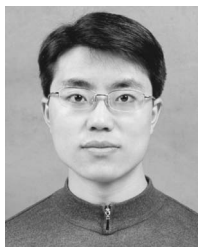
Witold Pedrycz (M'88–SM'94–F'99) received the M.Sc., Ph.D., and D.Sci. degrees from the Silesian University of Technology, Gliwice, Poland.

He is currently a Professor and Canada Research Chair (CRC) with the Department of Electrical and Computer Engineering, University of Alberta, Edmonton, AB, Canada. He is also with the Systems Research Institute, Polish Academy of Sciences, Warsaw, Poland. He has published numerous papers related to his research interests. He is also the author of nine research monographs on computational intelligence and software engineering. He is the Editor-in-Chief of *Information Sciences*. His research interests include computational intelligence, fuzzy modeling, knowledge discovery and data mining, fuzzy control (in particular fuzzy controllers), pattern recognition, knowledge-based neural networks, relational computation, bioinformatics, and software engineering.

Dr. Pedrycz was a member of numerous program committees of conferences in fuzzy sets and neurocomputing. He is the President of the International Fuzzy Systems Association and the North American Fuzzy Information Processing Society. He is currently the Associate Editor for the IEEE TRANSACTIONS ON SYSTEMS, MAN, AND CYBERNETICS, the IEEE TRANSACTIONS ON NEURAL NETWORKS, and the IEEE TRANSACTIONS ON FUZZY SYSTEMS.

Daren Yu received the M.Sc. and D.Sc. degrees from Harbin Institute of Technology, Harbin, China, in 1988 and 1996, respectively.

Since 1988, he has been with the School of Energy Science and Engineering, Harbin Institute of Technology. He has published more than 200 conference proceedings and journal papers on power control and fault diagnosis. His research interests include modeling, simulation, and control of power systems.



Jun Lang received the B.S. degree in computer science and technology in 2004 from Harbin Institute of Technology, Harbin, China, where he is currently working toward the Ph.D. degree with the Information Retrieval Laboratory, School of Computer Science and Technology.

In 2005, he was an Intern Student with the Natural Language Computing Group of Microsoft Research Asia, Beijing. In 2008–2009, he was invited to be a Visiting Student with the School of Computing, National University of Singapore, Singapore, and the Data Mining Group, Institute for Infocomm Research, Singapore. His research interests include text mining, natural language processing, and machine learning.