Hindawi Publishing Corporation
ISRN Applied Mathematics
Volume 2014, Article ID 382738, 11 pages
http://dx.doi.org/10.1155/2014/382738

Research Article

A Hybrid Feature Selection Method Based on Rough Conditional Mutual Information and Naive Bayesian Classifier

Zilin Zeng,[1,2] Hongjun Zhang,[1] Rui Zhang,[1] and Youliang Zhang[1]

[1] *PLA University of Science & Technology, Nanjing 210007, China*
[2] *Nanchang Military Academy, Nanchang 330103, China*

Correspondence should be addressed to Zilin Zeng; beauty1981@sohu.com and Hongjun Zhang; jsnjzhj@263.net

Received 17 December 2013; Accepted 12 February 2014; Published 30 March 2014

Academic Editors: A. Bellouquid, S. Biringen, H. C. So, and E. Yee

We introduced a novel hybrid feature selection method based on rough conditional mutual information and Naive Bayesian classifier. Conditional mutual information is an important metric in feature selection, but it is hard to compute. We introduce a new measure called rough conditional mutual information which is based on rough sets; it is shown that the new measure can substitute Shannon's conditional mutual information. Thus rough conditional mutual information can also be used to filter the irrelevant and redundant features. Subsequently, to reduce the feature and improve classification accuracy, a wrapper approach based on naive Bayesian classifier is used to search the optimal feature subset in the space of a candidate feature subset which is selected by filter model. Finally, the proposed algorithms are tested on several UCI datasets compared with other classical feature selection methods. The results show that our approach obtains not only high classification accuracy, but also the least number of selected features.

1. Introduction

With increase of data dimensionality in many domains such as bioinformatics, text categorization, and image recognition, feature selection has become one of the most important data mining preprocessing methods. The aim of feature selection is to find a minimal feature subset of the original datasets that is the most characterizing. Since feature selection can bring lots of advantages, such as avoiding overfitting, facilitating data visualization, reducing storage requirements, and reducing training times, it has attracted considerable attention in various areas [1].

In the past two decades, different techniques are proposed to address these challenging tasks. Dash and Liu [2] point out that there are four basic steps in a typical feature selection method, that is, subset generation, subset evaluation, stopping criterion, and validation. Most studies focus on the two major steps of feature selection: subset generation and subset evaluation. According to subset evaluation function, feature selection methods can be divided into two categories: filter method and wrapper method [3]. Filter methods are independent of predictor, whereas wrapper methods utilize their predictive power as the evaluation function. The merits of filter methods are high computation efficiency and its generality. However, the result of filter method is not always satisfactory. This is because the filter model separates feature selection from the classifier learning and selects the feature subsets that are independent from the learning algorithm. On the other hand, wrapper methods guarantee good results, but they are very slow when applied to large datasets.

In this paper, we propose a new algorithm which combined rough conditional entropy and naive Bayesian classifier to select features. First, in order to decrease the computational cost of wrapper search, a candidate feature set is selected by using rough conditional mutual information. Second, the candidate feature subset is then further refined by a wrapper procedure. We take advantages of both the filter and the wrapper. The main goal of our research is expected to obtain a few features while the classification accuracy is still very high. This approach provides the possibility of efficiently applying filter-wrapper model on some datasets from UCI [4], obtaining better results than other classical feature selection approaches.

In the remainder of the paper, related work is first discussed in the next section. Section 3 presents the preliminaries on Shannon's entropy and rough sets. Section 4 introduces the definitions of rough uncertainty measure and discusses their properties and interpretation. The proposed hybrid feature selection method is delineated in Section 5. The experimental results are presented in Section 6. Finally, a brief conclusion is given in Section 7.

2. Related Work

In filter based feature selection techniques, a number of relevance measures were applied to measure the performance of features for predicting decisions. These relevance measures can be divided into four categories: distance, dependency, consistency, and information. The most prominent distance-based method is relief [5]. This method uses Euclidean distance to select the relevance features. Since relief works only for binary classes, Kononenko generalized it to multiple classes called relief-F [6, 7]. However, relief and relief-F are unable to detect redundant features. Dependence measures or correlation measures quantify the ability to predict the value of one variable from the value of another variable. Hall's correlation-based feature selection (CFS) algorithm [8] is typical representative of this category. Consistency measures try to preserve the discriminative power of data in the original feature space. Rough set theory is a popular technique of this sort [9]. Among these measures, mutual information (MI) is the most widely used one in computing relevance. MI is a well-known concept from information theory and has been used to capture the relevance and redundancy among features. In this paper, we focus on the information-based measure in the filter model.

The main advantages of MI are its robustness to noise and transformation. In contrast to other measures, MI is not limited to linear dependencies but includes any nonlinear ones. Since Battiti proposed mutual information feature selector (MIFS) [10], more and more researchers began to study information-based feature selection. MIFS selects the feature that maximizes the information of the class, corrected by subtracting a quantity proportion to the average MI with the previously selected features. Battiti demonstrated that MI can be very useful in feature selection problems and the MIFS can be used in any classifying systems for its simplicity whatever the learning algorithm may be. Kwak and Choi [11] analyzed the limitations of MIFS and proposed method called MIFS-U, which, in general, makes a better estimation of the MI between input attributes and output classes than MIFS. They showed that MIFS does not work in nonlinear problems and proposed MIFS-U to improve MIFS for solving nonlinear problems. Another variant of MIFS is min-redundancy max-relevance (mRMR) criterion [12]. The method presented the theoretical analysis of the relationships of max-dependency, max-relevance, and min-redundancy. They proved that mRMR is equivalent to max-dependency for the first-order incremental search.

The limitations of MIFS, MIFS-U, and mRMR algorithms are as follows. Firstly, they are all incremental search schemes that select one feature at a time. At each pass, these methods select one feature with maximum criterion, without considering the interaction between groups of features. In many classification problems, groups of several features occurring simultaneously are relevant but not for the case of individual feature alone, for example, the XOR problem. Secondly, the coefficient β is a configurable parameter, which must be set experimentally. Thirdly, they are not accurate enough to quantify the dependency among features with respect to a given decision.

Assume $X = \{X_1, X_2, \ldots, X_n\}$ as an input feature set and Y as a target; our task is to select m ($m < n$) features from a pool such that their joint mutual information $I(\widetilde{X}_1, \widetilde{X}_2, \ldots, \widetilde{X}_m; Y)$ is maximized. However, the estimation of mutual information from the available data is a great challenge, especially multivariate mutual information. Martínez Sotoca and Pla [13] and Guo et al. [14] proposed different methods to approximate multivariate conditional mutual information, respectively. Nevertheless, their proofs are all based on the same inequality; that is, $I(X, Y \mid Z) \leq I(X, Y)$. The inequality does not hold under any conditions. Only if random variables X, Y, and Z satisfy Markovity, then the inequality holds. Many researchers try various methods to estimate mutual information. The most common methods are histogram [15], kernel density estimation (KDE) [16], and k-nearest neighbor estimation (K-NN) [17]. The standard histogram partitions the axes into distinct bins of width and then counts the number of observations; therefore, this estimation method is highly dependent on the choice of the width of the bins. Although the KDE is better than histogram, the bandwidth and kernel function are difficult to decide. The K-NN approach uses a fixed number of nearest neighbors to estimate the MI, but it seems more suitable for continuous random variables.

This paper will compute multivariate mutual information and multivariate conditional mutual information in a new perspective. Our method is based on rough entropy uncertainty measure. Several authors [18–21] have used Shannon's entropy and its variants to measure uncertainty in rough set theory. In this work, we will propose several rough entropy-based metrics. Some important properties and relationships of these uncertainty measures will be concluded. Then we will find a candidate feature subset by using rough conditional mutual information to filter the irrelevant and redundant features in the first stage. To overcome the limitations of the filter model, in the second stage, we will use the wrapper model with the sequential backward elimination scheme to search for an optimal feature subset from the candidate feature subset.

3. Preliminaries

In this section we briefly introduce some basic concepts and notations of the information theory and rough set theory.

3.1. Entropy, Mutual Information, and Conditional Mutual Information. Shannon's information theory, first introduced in 1948 [22], provides a way to measure the information of

random variables. The entropy is a measure of uncertainty of random variables [23]. Let $X = \{x_1, x_2, \ldots, x_n\}$ be a discrete random variable and let $p(x_i)$ be the probability of x_i; the entropy of X is defined by the following:

$$H(X) = -\sum_{i=1}^{n} p(x_i) \log p(x_i). \tag{1}$$

Here the base of log is 2 and the unit of entropy is the bit. If X and $Y = \{y_1, y_2, \ldots, y_m\}$ are two discrete random variables, the joint probability is $p(x_i, y_j)$, where $i = 1, 2, \ldots, n$ and $j = 1, 2, \ldots, m$. The joint entropy of X and Y is as follows:

$$H(X,Y) = -\sum_{i=1}^{n}\sum_{j=1}^{m} p(x_i, y_j) \log p(x_i, y_j). \tag{2}$$

When certain variables are known and others are not known, the remaining uncertainty is measured by the conditional entropy as follows:

$$\begin{aligned} H(X \mid Y) &= H(X,Y) - H(Y) \\ &= -\sum_{i=1}^{n}\sum_{j=1}^{m} p(x_i, y_j) \log p(x_i \mid y_j). \end{aligned} \tag{3}$$

The information found commonly in two random variables is of importance and this is defined as the mutual information between two variables as follows:

$$I(X;Y) = \sum_{i=1}^{n}\sum_{j=1}^{m} p(x_i, y_j) \log \frac{p(x_i \mid y_j)}{p(x_i)}. \tag{4}$$

If the mutual information between two random variables is large (small), it means two variables are closely (not closely) related. If the mutual information becomes zero, the two random variables are totally unrelated or the two variables are independent. The mutual information and the entropy have the following relation:

$$\begin{aligned} I(X;Y) &= H(X) - H(X \mid Y), \\ I(X;Y) &= H(Y) - H(X \mid Y), \\ I(X;Y) &= H(X) + H(Y) - H(X,Y), \\ I(X;X) &= H(X). \end{aligned} \tag{5}$$

For continuous random variables, the entropy and mutual information are defined as follows:

$$\begin{aligned} H(X) &= -\int p(x) \log p(x)\, dx \\ I(X;Y) &= \int p(x,y) \log \frac{p(x,y)}{p(x)\,p(y)} dx\, dy. \end{aligned} \tag{6}$$

Conditional mutual information is the reduction in the uncertainty of X due to knowledge of Y when Z is given. The conditional mutual information of random variables X and Y given Z is defined by the following:

$$I(X;Y \mid Z) = H(X \mid Z) - H(X \mid Y,Z). \tag{7}$$

Mutual information satisfies a chain rule; that is,

$$I(X_1, X_2, \ldots, X_n; Y) = \sum_{i=1}^{n} I(X_i; Y \mid X_{i-1}, X_{i-2}, \ldots, X_1). \tag{8}$$

3.2. Rough Sets. Rough sets theory, introduced by Pawlak [24], is a mathematical tool to handle imprecision, uncertainty, and vagueness. It has been applied in many fields [25] such as machine learning, data mining, and pattern recognition.

The notion of an information system provides a convenient basis for the representation of objects in terms of their attributes. An information system is a pair of (U, A), where U is a nonempty finite set of objects called the universe and A is a nonempty finite set of attributes; that is, $a : U \to V_a$ for $a \in A$, where V_a is called the domain of a. A decision table is a special case of information system $S = (U, A \cup \{d\})$, where attributes in A are called condition attributes and d is a designated attribute called the decision attribute.

For every set of attributes $B \subseteq A$, an indiscernibility relation $IND(B)$ is defined in the following way. Two objects, x_i and x_j, are indiscernible by the set of attribute B in A, if $b(x_i) = b(x_j)$ for every $b \in B$. The equivalence class of $IND(B)$ is called elementary set in B because it represents the smallest discernible groups of objects. For any element x_i of U, the equivalence class of x_i in relation $IND(B)$ is represented as $[x_i]_B$. For $B \subseteq A$, the indiscernibility relation $IND(B)$ constitutes a partition of U, which is denoted by $U/IND(B)$.

Given an information system $S = (U, A)$, for any subset $X \subseteq U$ and equivalence relation $IND(B)$, the B-lower and B-upper approximations of X are defined, respectively, as follows:

$$\begin{aligned} \underline{B}(X) &= \{x \in U : [x]_B \subseteq X\}, \\ \overline{B}(X) &= \{x \in U : [x]_B \cap X \neq \varnothing\}. \end{aligned} \tag{9}$$

4. Rough Entropy-Based Metrics

In this section, the concept of rough entropy is introduced to measure the uncertainty of knowledge in an information system and then some rough entropy-based uncertainty measures are presented. Some important properties of these uncertainty measures are deduced, respectively, and the relationships among them are discussed as well.

Definition 1. Given a set of samples $U = \{x_1, x_2, \ldots, x_n\}$ described by features F, $S \subseteq F$ is a subset of attributes. Then the rough entropy of the sample is defined by

$$\mathrm{RH}_{x_i}(S) = -\log \frac{|[x_i]_S|}{n}, \tag{10}$$

and the average entropy of the set of samples is computed as

$$\mathrm{RH}(S) = -\frac{1}{n}\sum_{i=1}^{n} \log \frac{|[x_i]_S|}{n}, \tag{11}$$

where $|[x_i]_S|$ is the cardinality of $[x_i]_S$.

Since for all x_i, $\{x_i\} \subseteq [x_i]_S \subseteq U$, $1/n \le |[x_i]_S|/n \le 1$, so we have $0 \le \mathrm{RH}(S) \le \log n$. $\mathrm{RH}(S) = \log n$ if and only if for all x_i, $|[x_i]|_S = 1$; that is, $U/IND(S) = \{\{x_1\}, \{x_2\}, \ldots, \{x_n\}\}$. $\mathrm{RH}(S) = 0$ if and only if for for all x_i, $|[x_i]|_S = n$; that is, $U/IND(S) = \{U\}$. Obviously, when knowledge S can distinguish any two objects, the rough entropy is the largest; when knowledge S can not distinguish any two objects, the rough entropy is zero.

Theorem 2. *Consider $RH(S) = H(S)$, where $H(S)$ is Shannon's entropy.*

Proof. Suppose $U = \{x_1, x_2, \ldots, x_n\}$ and $U/IND(S) = \{X_1, X_2, \ldots, X_m\}$, where $X_i = \{x_{i1}, x_{i2}, \ldots, x_{i|X_i|}\}$; then $H(S) = -\sum_{i=1}^{m}(|X_i|/n)\log(|X_i|/n)$. Because $X_i \cap X_j = \varnothing$ for $i \ne j$ and $X_i = [x]_S$ for any $x \in X_i$, we have

$$\begin{aligned}\mathrm{RH}(S) &= -\frac{1}{n}\sum_{i=1}^{n}\log\frac{|[x_i]_S|}{n}\\ &= \sum_{x\in X_1} -\frac{1}{n}\log\frac{|[x]_S|}{n} + \cdots + \sum_{x\in X_m} -\frac{1}{n}\log\frac{|[x]_S|}{n}\\ &= \left(-\frac{|X_1|}{n}\log\frac{|X_1|}{n}\right) + \cdots + \left(-\frac{|X_m|}{n}\log\frac{|X_m|}{n}\right)\\ &= -\sum_{i=1}^{m}\frac{|X_i|}{n}\log\frac{|X_i|}{n} = H(S).\end{aligned} \tag{12}$$

□

Theorem 2 shows that the rough entropy equals Shannon's entropy.

Definition 3. Suppose $R, S \subseteq F$ are two subsets of attributes; the joint rough entropy is defined as

$$\mathrm{RH}(R,S) = -\frac{1}{n}\sum_{i=1}^{n}\log\frac{|[x_i]_{R\cup S}|}{n}. \tag{13}$$

Due to $[x_i]_{R\cup S} = [x_i]_R \cap [x_i]_S$, therefore, $\mathrm{RH}(R,S) = -(1/n)\sum_{i=1}^{n}\log(|[x_i]_R \cap [x_i]_S|/n)$. According to Definition 3, we can observe that $\mathrm{RH}(R,S) = \mathrm{RH}(R \cup S)$.

Theorem 4. *Consider $RH(R,S) \ge RH(R)$ and $RH(R,S) \ge RH(S)$.*

Proof. Consider for all $x_i \in U$; we have $[x_i]_{S\cup R} \subseteq [x_i]_S$ and $[x_i]_{S\cup R} \subseteq [x_i]_R$, and then $|[x_i]_{S\cup R}| \le |[x_i]_S|$ and $|[x_i]_{S\cup R}| \le |[x_i]_R|$. Therefore, $\mathrm{RH}(R,S) \ge \mathrm{RH}(R)$ and $\mathrm{RH}(R,S) \ge \mathrm{RH}(S)$. □

Definition 5. Suppose $R, S \subseteq F$ are two subsets of attributes; the conditional rough entropy of R to S is defined as

$$\mathrm{RH}(R \mid S) = -\frac{1}{n}\sum_{i=1}^{n}\log\frac{|[x_i]_{R\cup S}|}{|[x_i]_S|}. \tag{14}$$

Theorem 6 (chain rule). *Consider $RH(R \mid S) = RH(R,S) - RH(S)$.*

Proof. Consider

$$\begin{aligned}&\mathrm{RH}(R,S) - \mathrm{RH}(S)\\ &\quad= -\frac{1}{n}\sum_{i=1}^{n}\log\frac{|[x_i]_{R\cup S}|}{n} + \frac{1}{n}\sum_{i=1}^{n}\log\frac{|[x_i]_S|}{n}\\ &\quad= -\frac{1}{n}\sum_{i=1}^{n}\log\frac{|[x_i]_{R\cup S}|}{|[x_i]_S|} = \mathrm{RH}(R \mid S).\end{aligned} \tag{15}$$

□

Definition 7. Suppose $R, S \subseteq F$ are two subsets of attributes; the rough mutual information of R and S is defined as

$$\mathrm{RI}(R;S) = \frac{1}{n}\sum_{i=1}^{n}\log\frac{n|[x_i]_{R\cup S}|}{|[x_i]_R|\cdot|[x_i]_S|}. \tag{16}$$

Theorem 8 (the relation between rough mutual information and rough entropy). *Consider*

(1) $RI(R;S) = RI(S;R)$.
(2) $RI(R;S) = RH(R) + RH(S) - RH(R,S)$.
(3) $RI(R;S) = RH(R) - RH(R \mid S) = RH(S) - RH(S \mid R)$.

Proof. The conclusions of (1) and (3) are straightforward; here we give the proof of property (2).

(2) Consider

$$\begin{aligned}&\mathrm{RH}(R) + \mathrm{RH}(S) - \mathrm{RH}(R,S)\\ &\quad= -\frac{1}{n}\sum_{i=1}^{n}\log\frac{|[x_i]_R|}{n} - \frac{1}{n}\sum_{i=1}^{n}\log\frac{|[x_i]_S|}{n}\\ &\qquad+ \frac{1}{n}\sum_{i=1}^{n}\log\frac{|[x_i]_{R\cup S}|}{n}\\ &\quad= \frac{1}{n}\sum_{i=1}^{n}\left(-\log\frac{|[x_i]_R|}{n} - \log\frac{|[x_i]_S|}{n} + \log\frac{|[x_i]_{R\cup S}|}{n}\right)\\ &\quad= \frac{1}{n}\sum_{i=1}^{n}\log\frac{n|[x_i]_{R\cup S}|}{|[x_i]_R|\cdot|[x_i]_S|} = \mathrm{RI}(R;S).\end{aligned} \tag{17}$$

□

Definition 9. The rough conditional mutual information of R and S given T is defined by

$$\mathrm{RI}(R;S \mid T) = \frac{1}{n}\sum_{i=1}^{n}\frac{|[x_i]_{R\cup S\cup T}|\cdot|[x_i]_S|}{|[x_i]_{R\cup S}|\cdot|[x_i]_{S\cup T}|}. \tag{18}$$

Theorem 10. *The following equations hold:*

(1) $RI(R;S \mid T) = RH(R \mid S) - RH(R \mid S,T)$;
(2) $RI(R;S \mid T) = RI(R,T;S) - RI(R;S)$.

Proof. (1) Consider

$$\begin{aligned}
&\mathrm{RH}(R \mid S) - \mathrm{RH}(R \mid S, T) \\
&\quad = -\frac{1}{n}\sum_{i=1}^{n} \log \frac{\left|[x_i]_{R\cup S}\right|}{\left|[x_i]_S\right|} + \frac{1}{n}\sum_{i=1}^{n} \log \frac{\left|[x_i]_{R\cup S\cup T}\right|}{\left|[x_i]_{S\cup T}\right|} \\
&\quad = \frac{1}{n}\sum_{i=1}^{n}\left(-\log \frac{\left|[x_i]_{R\cup S}\right|}{\left|[x_i]_S\right|} + \log \frac{\left|[x_i]_{R\cup S\cup T}\right|}{\left|[x_i]_{S\cup T}\right|}\right) \\
&\quad = \frac{1}{n}\sum_{i=1}^{n} \frac{\left|[x_i]_{R\cup S\cup T}\right| \cdot \left|[x_i]_S\right|}{\left|[x_i]_{R\cup S}\right| \cdot \left|[x_i]_{S\cup T}\right|} = \mathrm{RI}(R; S \mid T).
\end{aligned} \tag{19}$$

(2) Consider

$$\begin{aligned}
&\mathrm{RI}(R, T; S) - \mathrm{RI}(R; S) \\
&\quad = \mathrm{RI}(R \cup T; S) - \mathrm{RI}(R; S) \\
&\quad = \frac{1}{n}\sum_{i=1}^{n} \log \frac{n\left|[x_i]_{R\cup T\cup S}\right|}{\left|[x_i]_{R\cup T}\right| \cdot \left|[x_i]_S\right|} \\
&\qquad - \frac{1}{n}\sum_{i=1}^{n} \log \frac{n\left|[x_i]_{R\cup S}\right|}{\left|[x_i]_R\right| \cdot \left|[x_i]_S\right|} \\
&\quad = \frac{1}{n}\sum_{i=1}^{n}\left(\log \frac{n\left|[x_i]_{R\cup T\cup S}\right|}{\left|[x_i]_{R\cup T}\right| \cdot \left|[x_i]_S\right|}\right. \\
&\qquad\qquad \left. - \log \frac{n\left|[x_i]_{R\cup S}\right|}{\left|[x_i]_R\right| \cdot \left|[x_i]_S\right|}\right) \\
&\quad = \frac{1}{n}\sum_{i=1}^{n} \frac{\left|[x_i]_{R\cup S\cup T}\right| \cdot \left|[x_i]_S\right|}{\left|[x_i]_{R\cup S}\right| \cdot \left|[x_i]_{S\cup T}\right|} = \mathrm{RI}(R; S \mid T).
\end{aligned} \tag{20}$$

□

5. A Hybrid Feature Selection Method

In this section, we propose a novel hybrid feature selection method based on rough conditional mutual information and naive Bayesian classifier.

5.1. Feature Selection by Rough Conditional Mutual Information. Given a set of sample U described by the attribute set F, in terms of mutual information, the purpose of feature selection is to find a feature set S ($S \subseteq F$) with m features, which jointly have the largest dependency on the target class y. This criterion, called max-dependency, has the following form:

$$\max D(S, y), \quad D = I(\{f_1, f_2, \ldots, f_m\}; y). \tag{21}$$

According to the chain rule for information,

$$I(f_1, f_2, \ldots, f_m; y) = \sum_{i=1}^{m} I(f_i; yf_{i-1}, f_{i-2}, \ldots, f_1); \tag{22}$$

that is to say, we can select a feature which produces the maximum conditional mutual information, formally written as

$$\max_{f\in F-S} D(f, S, y), \quad D = I(f; y \mid S), \tag{23}$$

where S represents the selected feature set.

Figure 1 illustrates the validity of this criterion. Here, f_i represents a feature highly correlated with f_j, and f_k is much less correlated with f_i. The mutual information between vectors (f_i, f_j) and y is indicated by a shadowed area consisting of three different patterns of patches; that is, $I((f_i, f_j), y) = A + B + C$, where A, B, and C are defined by different cases of overlap. In detail,

(1) $(A + B)$ is the mutual information between f_i and y, that is, $I(f_i; y)$;

(2) $(B + C)$ is the mutual information between f_j and y, that is, $I(f_j; y)$;

(3) $(B + D)$ is the mutual information between f_i and f_j, that is, $I(f_i; f_j)$;

(4) C is the conditional mutual information between f_j and y given f_i, that is, $I(f_j; y \mid f_i)$;

(5) E is the mutual information between f_k and y, that is, $I(f_k; y)$.

This illustration clearly shows that the features maximizing the mutual information not only depend on their predictive information individually, for example, $(A+B+B+C)$, but also need to take account of redundancy between them. In this example, feature f_i should be selected first since the mutual information between f_i and y is the largest, and feature f_k should have priority for selection over f_j in spite of the latter having larger individual mutual information with y. This is because f_k provides more complementary information to feature f_i to predict y than does f_j (as $E > C$ in Figure 1); that is to say, for each round, we should select a feature which maximizes conditional mutual information. From Theorem 2, we know that rough entropy equals Shannon's entropy; therefore, we can select a feature which produces the maximum rough conditional mutual information.

We adopt the forward feature algorithm to select features. Each single input feature is added to selected features set based on maximizing rough conditional mutual information, that is, given selected feature set S, maximizing the rough mutual information of f_i and target class y, where f_i belongs to the remain feature set. In order to apply the rough conditional mutual information measure to the filter model well, a numerical threshold value β ($\beta > 0$) is set to $\mathrm{RI}(f_i; y \mid S)$. This can help the algorithm to be resistant to noise data and to overcome the overfitting problem to a certain extent [26]. The procedure can be performed until $\mathrm{RI}(f_i; y \mid S) > \beta$ is satisfied. The filter algorithm can be described by the following procedure.

(1) Initialization: set $F \leftarrow$ "initial set of all features," $S \leftarrow$ "empty set," and $y \leftarrow$ "class outputs."

(2) Computation of the rough mutual information of the features with the class outputs: for each feature ($f_i \in F$), compute $\mathrm{RI}(f_i; y)$.

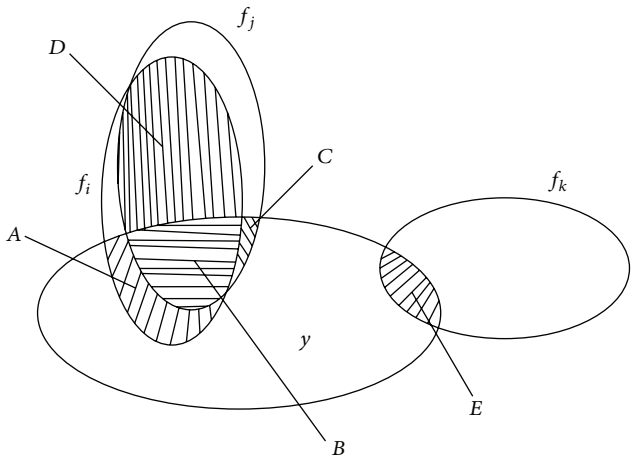

Figure 1: Illustration of mutual information and conditional mutual information for different scenarios.

(3) Selection of the first feature: find the feature f_i that maximizes $\mathrm{RI}(f_i; y)$; set $F \leftarrow F - \{f_i\}$ and $S \leftarrow \{f_i\}$.

(4) Greedy selection: repeat until the termination condition is satisfied:

   (a) computation of the rough mutual information $\mathrm{RI}(f_i; y \mid S)$ for each feature $f_i \in F$,

   (b) selection of the next feature: choose the feature f_i as the one that maximizes $\mathrm{RI}(f_i; y \mid S)$; set $F \leftarrow F - \{f_i\}$ and $S \leftarrow S \cup \{f_i\}$.

(5) Output the set containing the selected features: S.

5.2. Selecting the Best Feature Subset on Wrapper Approach. The wrapper model uses the classification accuracy of a predetermined learning algorithm to determine the goodness of the selected subset. It searches for features that are better suited to the learning algorithm, aiming at improving the performance of the learning algorithm; therefore, the wrapper approach generally outperforms the filter approach in the aspect of the final predictive accuracy of a learning machine. However, it is more computationally expensive than the filter models. Although many wrapper methods are not exhaustive search, most of them still incur time complexity $O(N^2)$ [27, 28] where N is the number of features of the dataset. Hence, it is worth reducing the search space before using wrapper feature selection approach. Through the filter model, it can reduce high computational cost and avoid encountering the local maximal problem. Therefore, the final subset of the features obtained contains a few features while the predictive accuracy is still high.

In this work, we propose the reducing of the search space of the original feature set to the best candidate which can reduce the computational cost of the wrapper search effectively. Our method uses the sequential backward elimination technique to search for every possible subset of features through the candidate space.

The features are ranked according to the average accuracy of the classifier, and then features will be removed one by one from the candidate feature subset only if such exclusion improves or does not change the classifier accuracy. Different kinds of learning models can be applied to wrappers. However, different kinds of learning machines have different discrimination abilities. Naive Bayesian classifier is widely used in machine learning because it is fast and easy to be implemented. Rennie et al. [29] show that its performance is competitive with the state-of-the-art models like SVM while the latter has too many parameters to decide. Therefore, we choose the naive Bayesian classifier as the core of fine tuning. The decrement selection procedure for selecting an optimal feature subset based on the wrapper approach can be seen as shown in Algorithm 1.

There are two phases in the wrapper algorithm, as shown in wrapper algorithm. In the first stage, we compute the classification accuracy Acc_C of the candidate feature set which is the results of filter model (step 1), where Classperf (D, C) represents the average classification accuracy of dataset D with candidate features C. The results are obtained by 10-fold cross-validation. For each $f_i \in C$, we compute the average accuracy *Score*. Then features are ranked according to *Score* value (steps 3–6). In the second stage, we deal with the list of the ordered features once; each feature in the list determines the first till the last ranked feature (steps 8–26). In this stage, each feature in the list considers the average accuracy of the naive Bayesian classifier only if the feature is excluded. If any feature is found to lead to the most improved average accuracy and the relative accuracy [30] is more than δ_1 (steps 11–14), the feature then will be removed. Otherwise, every possible feature is considered and the feature that leads to the largest average accuracy will be chosen and removed (step 15). The one that leads to the improvement or the unchanging of the average accuracy (steps 17–20) or the degrading of the relative accuracy not worse than δ_2 (steps 21–24) will be removed. In general, δ_1 should take value in [0, 0.1] and δ_2 should take value in [0, 0.02]. In the following, if not specified, $\delta_1 = 0.05$ and $\delta_2 = 0.01$.

This decrement selection procedure is repeated until the termination condition is satisfied. Usually, the sequential backward elimination is more computationally expensive than the incremental sequential forward search. However, it could yield a better result when considering the local maximal. In addition, the sequential forward search adding one feature at each pass does not take the interaction between the groups of the features into account [31]. In many classification problems, the class variable may be affected by grouping several features but not the individual feature alone. Therefore, the sequential forward search is unable to find the dependencies between the groups of the features while the performance can be degraded sometimes.

6. Experimental Results

This section illustrates the evaluation of our method in terms of the classification accuracy and the number of selected features in order to see how good the filter wrapper is in the

Input: data set D, candidate feature set C
Output: an optimal feature set B

(1) $\text{Acc}_C = \text{Classperf}(D, C)$
(2) set $B = \{\}$
(3) **for all** $f_i \in C$ do
(4) Score = Classperf (D, f_i)
(5) append f_i to B
(6) **end for**
(7) sort B in an ascending order according to Score value
(8) **while** $|B| > 1$ do
(9) **for all** $f_i \in B$ according to order do
(10) $\text{Acc}_{f_i} = \text{Classperf}(D, B - \{f_i\})$
(11) **if** $(\text{Acc}_{f_i} - \text{Acc}_C)/\text{Acc}_F > \delta_1$ **then**
(12) $B = B - \{f_i\}$, $\text{Acc}_C = \text{Acc}_{f_i}$
(13) go to *Step 8*
(14) **end if**
(15) Select f_i with the maximum Acc_{f_i}
(16) **end for**
(17) **if** $\text{Acc}_{f_i} \geq \text{Acc}_C$ **then**
(18) $B = B - \{f_i\}$, $\text{Acc}_C = \text{Acc}_{f_i}$
(19) go to *Step 8*
(20) **end if**
(21) **if** $(\text{Acc}_C - \text{Acc}_{f_i})/\text{Acc}_C \leq \delta_2$ **then**
(22) $B = B - \{f_i\}$, $\text{Acc}_C = \text{Acc}_{f_i}$
(23) go to *Step 8*
(24) **end if**
(25) go to *Step 27*
(26) **end while**
(27) Return an optimal feature subset B

Algorithm 1: Wrapper algorithm.

Table 1: Experimental data description.

Number	Datasets	Instances	Features	Classes
1	Arrhythmia	452	279	16
2	Hepatitis	155	19	2
3	Ionosphere	351	34	2
4	Segment	2310	19	7
5	Sonar	208	60	2
6	Soybean	683	35	19
7	Vote	435	16	2
8	WDBC	569	16	2
9	Wine	178	12	3
10	Zoo	101	16	7

situation of large and middle-sized features. In addition, the performance of the rough conditional mutual information algorithm is compared with three typical feature selection methods which are based on three different evaluation criterions, respectively. These methods include correlation based feature selection (CFS), consistency based algorithm, and min-redundancy max-relevance (mRMR). The results illustrate the efficiency and effectiveness of our method.

In order to compare our hybrid method with some classical techniques, 10 databases are downloaded from UCI repository of machine learning databases. All these datasets are widely used by the data mining community for evaluating learning algorithms. The details of the 10 UCI experimental datasets are listed in Table 1. The sizes of databases vary from 101 to 2310, the numbers of original features vary from 12 to 279, and the numbers of classes vary from 2 to 19.

6.1. Unselect versus CFS, Consistency Based Algorithm, mRMR, and RCMI. In Section 5, rough conditional mutual information is used to filter the redundant and irrelevant features. In order to compute the rough mutual information,

Table 2: Number and accuracy of selected features with different algorithms tested by naive Bayes.

Number	Unselect	CFS		Consistency		mRMR		RCMI	
	Accuracy	Accuracy	Feature number	Accuracy	Feature number	Accuracy	Feature number	Accuracy	Feature number
1	75.00%	78.54%	22	75.66%	24	75.00%	21	77.43%	16
2	84.52%	89.03%	8	85.81%	13	87.74%	7	85.13%	8
3	90.60%	92.59%	13	90.88%	7	90.60%	7	94.87%	6
4	91.52%	93.51%	8	93.20%	9	93.98%	5	93.07%	3
5	85.58%	77.88%	19	82.21%	14	84.13%	10	86.06%	15
6	92.09%	91.80%	24	84.63%	13	90.48%	21	91.22%	24
7	90.11%	94.71%	5	91.95%	12	95.63%	1	95.63%	2
8	95.78%	96.66%	11	96.84%	8	95.78%	9	95.43%	4
9	98.31%	98.88%	10	99.44%	5	98.88%	8	98.88%	5
10	95.05%	95.05%	10	93.07%	5	94.06%	4	95.05%	10
Average	89.86%	90.87%	13	89.37%	11	90.63%	9.3	**91.28%**	**9.3**

Table 3: Number and accuracy of selected feature with different algorithms tested by CART.

Number	Unselect	CFS		Consistency		mRMR		RCMI	
	Accuracy	Accuracy	Feature number	Accuracy	Feature number	Accuracy	Feature number	Accuracy	Feature number
1	72.35%	72.57%	22	71.90%	24	73.45%	21	75.00%	16
2	79.35%	81.94%	8	80.00%	13	83.23%	7	85.16%	8
3	89.74%	90.31%	13	89.46%	7	86.89%	7	91.74%	6
4	96.15%	96.10%	8	95.28%	9	95.76%	5	95.54%	3
5	74.52%	74.04%	19	77.88%	14	76.44%	10	77.40%	15
6	92.53%	91.65%	24	85.94%	13	92.24%	21	91.36%	24
7	95.63%	95.63%	5	95.63%	12	95.63%	1	96.09%	2
8	93.50%	94.55%	11	94.73%	8	95.43%	9	94.90%	4
9	94.94%	94.94%	10	96.07%	5	93.26%	8	94.94%	5
10	92.08%	93.07%	10	92.08%	5	92.08%	4	93.07%	10
Average	88.08%	88.48%	13	87.90%	11	88.44%	9.3	**89.52%**	**9.3**

we employ Fayyad and Irani's MDL discretization algorithm [32] to transform continuous features into discrete ones.

We use naive Bayesian and CART classifier to test the classification accuracy of selected features with different feature selection methods. The results in Tables 2 and 3 show the classification accuracies and the number of selected features obtained by the original feature (unselect), RCMI, and other feature selectors. According to Tables 2 and 3, we can find that the selected feature by RCMI has the highest average accuracy in terms of naive Bayes and CART. It can also be observed that RCMI can achieve the least average number of selected features which is the same as mRMR. This shows that RCMI is better than CFS and consistency based algorithm and is comparable to mRMR.

In addition, to illustrate the efficiency of RCMI, we experiment on Ionosphere, Sonar, and Wine datasets, respectively. A different number of the selected features obtained by RCMI and mRMR are tested on naive Bayesian classifier, as presented in Figures 2, 3, and 4. In Figures 2–4, the classification accuracies are the results of 10-fold cross-validation tested by naive Bayes. The number k in x-axis refers to the first k features with the selected order by different methods. The results in Figures 2–4 show that the average accuracy of classifier with RCMI is comparable to mRMR in the majority of cases. We can see that the maximum value of the plots for each dataset with RCMI method is higher than mRMR. For example, the highest accuracy of Ionosphere achieved by RCMI is 94.87% while the highest accuracy achieved by mRMR is 90.60%. At the same time, we can also notice that the RCMI method has the number of maximum values higher than mRMR. It shows that RCMI is an effective measure for feature selection.

However, the number of the features selected by the RCMI method is still more in some datasets. Therefore, to improve performance and reduce the number of the selected features, these problems were conducted by using the wrapper method. With removal of the redundant and irreverent features, the core of wrappers for fine tuning can perform much faster.

6.2. Filter Wrapper versus RCMI and Unselect. Similarly, we also use naive Bayesian and CART to test the classification accuracy of selected features with filter wrapper, RCMI, and unselect. The results in Tables 4 and 5 show the classification accuracies and the number of selected features.

Now we analyze the performance of these selected features. First, we can conclude that although most of features are removed from the raw data, the classification accuracies

Table 4: Number and accuracy of selected features based on naive Bayes.

Number	Unselect		RCMI		Filter wrapper	
	Accuracy	Feature number	Accuracy	Feature number	Accuracy	Feature number
1	75.00%	279	77.43%	16	78.76%	9
2	84.52%	19	85.13%	8	85.81%	3
3	90.60%	34	94.87%	6	94.87%	6
4	91.52%	19	93.07%	3	94.33%	3
5	85.58%	60	86.06%	15	86.54%	6
6	92.09%	35	91.22%	24	92.24%	10
7	90.11%	16	95.63%	2	95.63%	1
8	95.78%	16	95.43%	4	95.43%	4
9	98.31%	12	98.88%	5	96.07%	3
10	95.05%	16	95.05%	10	95.05%	8
Average	89.86%	50.6	91.28%	9.3	**91.47%**	**5.3**

Table 5: Number and accuracy of selected features based on CART.

Number	Unselect		RCMI		Filter wrapper	
	Accuracy	Feature number	Accuracy	Feature number	Accuracy	Feature number
1	72.35%	279	75.00%	16	75.89%	9
2	79.35%	19	85.16%	8	85.81%	3
3	89.74%	34	91.74%	6	91.74%	6
4	96.15%	19	95.54%	3	95.80%	3
5	74.52%	60	77.40%	15	80.77%	6
6	92.53%	35	91.36%	24	91.95%	10
7	95.63%	16	96.09%	2	95.63%	1
8	93.50%	16	94.90%	4	94.90%	4
9	94.94%	12	94.94%	5	93.82%	3
10	92.08%	16	93.07%	10	93.07%	8
Average	88.08%	50.6	89.52%	9.3	**89.94%**	**5.3**

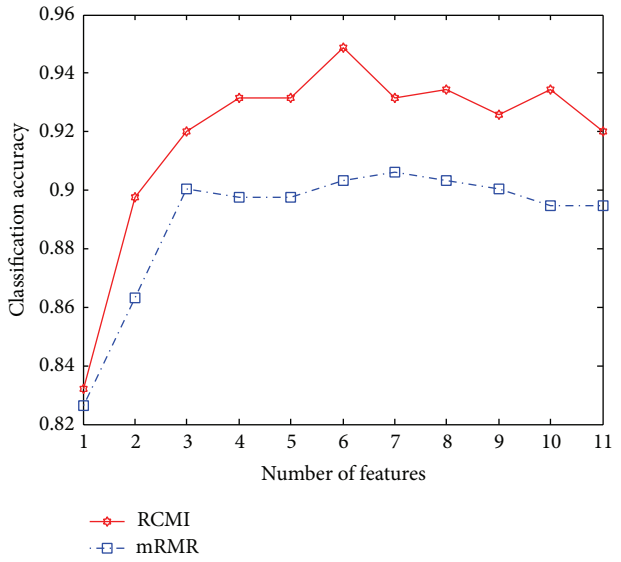

Figure 2: Classification accuracy of different number of selected features on Ionosphere dataset (naive Bayes).

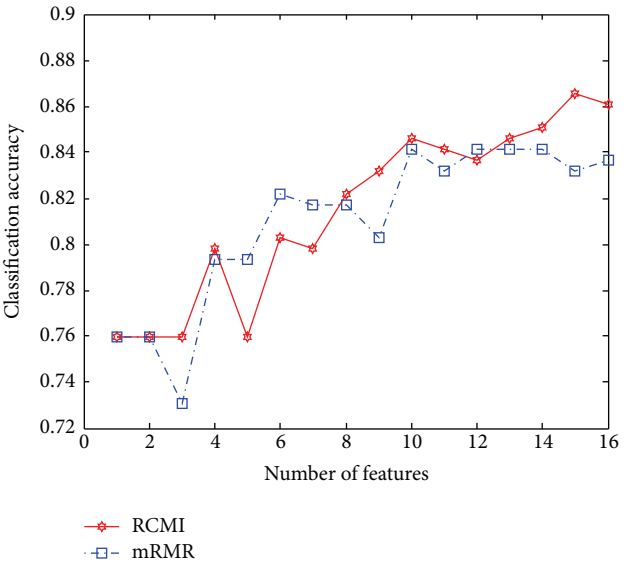

Figure 3: Classification accuracy of different number of selected features on Sonar dataset (naive Bayes).

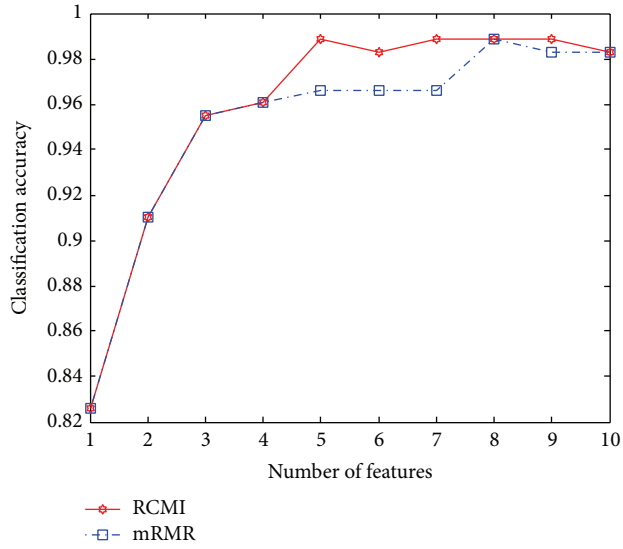

Figure 4: Classification accuracy of different number of selected features on Wine dataset (naive Bayes).

do not decrease; on the contrary, the classification accuracies increase in the majority of datasets. The average accuracies derived from RCMI and filter-wrapper method are all higher than the unselect datasets with respect to naive Bayes and CART. With respect to naive Bayesian learning algorithm, the average accuracy is 91.47% for filter wrapper, while 89.86% for unselect. The average classification accuracy increased 1.8%. With respect to CART learning algorithm, the average accuracy is 89.94% for filter wrapper, while 88.08% for unselect. The average classification accuracy increased 2.1%. The average number of selected features is 5.3 for filter wrapper, while 9.3 for RCMI and 50.6 for unselect. The average number of selected features reduced 43% and 89.5%, respectively. Therefore, the average value of classification accuracy and number of features obtained from the filter-wrapper method are better than those obtained from the RCMI and unselect. In other words, using the RCMI and wrapper methods as a hybrid improves the classification efficiency and accuracy compared with using the RCMI method individually.

7. Conclusion

The main goal of feature selection is to find a feature subset as small as possible, while the feature subset has highly prediction accuracy. A hybrid feature selection approach which takes advantages of filter model and wrapper model has been presented in this paper. In the filter model, measuring the relevance between features plays an important role. A number of measures were proposed. Mutual information is widely used for its robustness. However, it is difficult to compute mutual information, especially multivariate mutual information. We proposed a set of rough based metrics to measure the relevance between features and analyzed some important properties of these uncertainty measures. We have proved that the RCMI can substitute Shannon's conditional mutual information; thus, RCMI can be used as an effective measure to filter the irrelevant and redundant features. Based on the candidate feature subset by RCMI, naive Bayesian classifier is applied to the wrapper model. The accuracy of naive Bayesian and CART classifier was used to evaluate the goodness of feature subsets. The performance of the proposed method is evaluated based on ten UCI datasets. Experimental results on ten UCI datasets show that the filter-wrapper method outperformed CFS, consistency based algorithm, and mRMR at most cases. Our technique not only chooses a small subset of features from a candidate subset but also provides good performance in predictive accuracy.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

Acknowledgment

This work was supported by the National Natural Science Foundation of China (70971137).

References

[1] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," *Journal of Machine Learning Research*, vol. 3, pp. 1157–1182, 2003.

[2] M. Dash and H. Liu, "Feature selection for classification," *Intelligent Data Analysis*, vol. 1, no. 1–4, pp. 131–156, 1997.

[3] M. Dash and H. Liu, "Consistency-based search in feature selection," *Artificial Intelligence*, vol. 151, no. 1-2, pp. 155–176, 2003.

[4] C. J. Merz and P. M. Murphy, *UCI Repository of Machine Learning Databases*, Department of Information and Computer Science, University of California, Irvine, Calif, USA, 1996, http://mlearn.ics.uci.edu/MLRepository.html .

[5] K. Kira and L. A. Rendell, "Feature selection problem: traditional methods and a new algorithm," in *Proceedings of the 9th National Conference on Artificial Intelligence (AAAI '92)*, pp. 129–134, July 1992.

[6] M. Robnik-Šikonja and I. Kononenko, "Theoretical and empirical analysis of ReliefF and RReliefF," *Machine Learning*, vol. 53, no. 1-2, pp. 23–69, 2003.

[7] I. Kononenko, "Estimating attributes: analysis and extension of RELIEF," in *Proceedings of European Conference on Machine Learning (ECML '94)*, pp. 171–182, 1994.

[8] M. A. Hall, *Correlation-based feature subset selection for machine learning [Ph.D. thesis]*, Department of Computer Science, University of Waikato, Hamilton, New Zealand, 1999.

[9] J. G. Bazan, "A comparison of dynamic and non-dynamic rough set methods for extracting laws from decision table," in *Rough Sets in Knowledge Discovery*, L. Polkowski and A. Skowron, Eds., pp. 321–365, Physica, Heidelberg, Germany, 1998.

[10] R. Battiti, "Using mutual information for selecting features in supervised neural net learning," *IEEE Transactions on Neural Networks*, vol. 5, no. 4, pp. 537–550, 1994.

[11] N. Kwak and C. H. Choi, "Input feature selection for classification problems," *IEEE Transactions on Neural Networks*, vol. 13, no. 1, pp. 143–159, 2002.

[12] H. Peng, F. Long, and C. Ding, "Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 8, pp. 1226–1238, 2005.

[13] J. Martínez Sotoca and F. Pla, "Supervised feature selection by clustering using conditional mutual information-based distances," *Pattern Recognition*, vol. 43, no. 6, pp. 2068–2081, 2010.

[14] B. Guo, R. I. Damper, S. R. Gunn, and J. D. B. Nelson, "A fast separability-based feature-selection method for high-dimensional remotely sensed image classification," *Pattern Recognition*, vol. 41, no. 5, pp. 1670–1679, 2008.

[15] D. W. Scott, *Multivariate Density Estimation: Theory, Practice and Visualization*, John Wiley & Sons, New York, NY, USA, 1992.

[16] B. W. Silverman, *Density Estimation for Statistics and Data Analysis*, Chapman & Hall, London, UK, 1986.

[17] A. Kraskov, H. Stögbauer, and P. Grassberger, "Estimating mutual information," *Physical Review E*, vol. 69, no. 6, Article ID 066138, 16 pages, 2004.

[18] T. Beaubouef, F. E. Petry, and G. Arora, "Information-theoretic measures of uncertainty for rough sets and rough relational databases," *Information Sciences*, vol. 109, no. 1–4, pp. 185–195, 1998.

[19] I. Düntsch and G. Gediga, "Uncertainty measures of rough set prediction," *Artificial Intelligence*, vol. 106, no. 1, pp. 109–137, 1998.

[20] G. J. Klir and M. J. Wierman, *Uncertainty Based Information: Elements of Generalized Information Theory*, Physica, New York, NY, USA, 1999.

[21] J. Liang and Z. Shi, "The information entropy, rough entropy and knowledge granulation in rough set theory," *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, vol. 12, no. 1, pp. 37–46, 2004.

[22] C. E. Shannon, "A mathematical theory of communication," *The Bell System Technical Journal*, vol. 27, pp. 379–423, 1948.

[23] T. M. Cover and J. A. Thomas, *Elements of InformationTheory*, John Wiley & Sons, New York, NY, USA, 1991.

[24] Z. Pawlak, "Rough sets," *International Journal of Computer and Information Sciences*, vol. 11, no. 5, pp. 341–356, 1982.

[25] X. Hu and N. Cercone, "Learning in relational databases: a rough set approach," *Computational Intelligence*, vol. 11, no. 2, pp. 323–338, 1995.

[26] J. Huang, Y. Cai, and X. Xu, "A hybrid genetic algorithm for feature selection wrapper based on mutual information," *Pattern Recognition Letters*, vol. 28, no. 13, pp. 1825–1844, 2007.

[27] H. Liu and L. Yu, "Toward integrating feature selection algorithms for classification and clustering," *IEEE Transactions on Knowledge and Data Engineering*, vol. 17, no. 4, pp. 491–502, 2005.

[28] L. Yu and H. Liu, "Efficient feature selection via analysis of relevance and redundancy," *Journal of Machine Learning Research*, vol. 5, pp. 1205–1224, 2003/04.

[29] J. D. M. Rennie, L. Shih, J. Teevan, and D. Karger, "Tackling the poor assumptions of naive bayes text classifiers," in *Proceedings, Twentieth International Conference on Machine Learning*, pp. 616–623, Washington, DC, USA, August 2003.

[30] R. Setiono and H. Liu, "Neural-network feature selector," *IEEE Transactions on Neural Networks*, vol. 8, no. 3, pp. 654–662, 1997.

[31] S. Foithong, O. Pinngern, and B. Attachoo, "Feature subset selection wrapper based on mutual information and rough sets," *Expert Systems with Applications*, vol. 39, no. 1, pp. 574–584, 2012.

[32] U. Fayyad and K. Irani, "Multi-interval discretization of continuous-valued attributes for classification learning," in *Proceedings of the 13th International Joint Conference on Artificial Intelligence*, pp. 1022–1027, Morgan Kaufmann, San Mateo, Calif, USA, 1993.