WALAILAK JOURNAL
http://wjst.wu.ac.th Information Technology

Min-Uncertainty & Max-Certainty Criteria of Neighborhood Rough-Mutual Feature Selection

Sombut FOITHONG[1,*], Phaitoon SRINIL[1], Ouen PINNGERN[2] and Boonwat ATTACHOO[3]

[1]Faculty of Science and Arts, Burapha University, Chanthaburi Campus, Chanthaburi 22170, Thailand
[2]Department of Computer Science, Faculty of Science, Ramkhamhaeng University, Bangkok 10240, Thailand
[3]Department of Computer Engineering, Faculty of Engineering, King Mongkut's Institute of Technology Ladkrabang, Bangkok 10520, Thailand

([*]Corresponding author's e-mail: sombut@buu.ac.th)

Received: 27 October 2015, Revised: 8 February 2016, Accepted: 8 March 2016

Abstract

Feature Selection (FS) is viewed as an important preprocessing step for pattern recognition, machine learning, and data mining. Most existing FS methods based on rough set theory use the dependency function for evaluating the goodness of a feature subset. However, these FS methods may unsuccessfully be applied on dataset with noise, which determine only information from a positive region but neglect a boundary region. This paper proposes a criterion of the maximal lower approximation information (Max-Certainty) and minimal boundary region information (Min-Uncertainty), based on neighborhood rough set and mutual information for evaluating the goodness of a feature subset. We combine this proposed criterion with neighborhood rough set, which is directly applicable to numerical and heterogeneous features, without involving a discretization of numerical features. Comparing it with the rough set based approaches, our proposed method improves accuracy over various experimental data sets. Experimental results illustrate that much valuable information can be extracted by using this idea. This proposed technique is demonstrated on discrete, continuous, and heterogeneous data, and is compared with other FS methods in terms of subset size and classification accuracy.

Keywords: Feature selection, mutual information, neighborhood rough sets, classification, boundary region

Introduction

Feature Selection (FS) is an essential technique used in data preprocessing in many fields of artificial intelligence, such as machine learning, pattern recognition, text categorization, and data mining. FS is a process which selects a subset of the original features of a data set while preserving the most essential information of the data set. FS has also been developed for decades, as in the examples of the statistical pattern recognition [1,2], machine learning [3-5], and data mining [6,7]. At the same time, it has been widely applied in a number of fields, such as text classification [8,9], intrusion detection [10,11], and gene expression analysis [12,13].

Over the past 10 years, a large number of feature selection methods have been proposed. The most widely used methods for filter-feature selection are rough set [14,15] and mutual information [16]. Most existing FS approaches [17-25] based on the rough set method take the subset evaluation method, which searches for a minimum subset of features that satisfies some goodness measures relying on the information gathered from the lower approximation alone. The mutual information (MI) approach is widely used for feature ranking [26-29], which assesses features individually and assigns them weights

according to their degrees of relevance. A subset of features is often selected from the top of the ranking list, which approximates the set of relevant features. However, the disadvantages of feature ranking are the difficulty in removing redundant features, because features are likely to have similar rankings. Besides this, this feature selection technique requires predefining of the number of features to be selected, and the optimal subset is taken from the best result of the classification accuracy.

The rough set (RS) theory, proposed by Pawlak [14,15], provides a new mathematic model for dealing with imprecise, uncertain, and incomplete information. The rough set approach analyzes data relying on 2 important concepts, namely, the lower and upper approximation of a set. In RS theory, we desire to achieve reducts of an information system, in order to extract rule-like knowledge. A reduct is a minimal attribute subset of the original attributes, which has the same classification of objects of the universe as the whole set of attributes. Most existing RS-based FS approaches have been presented [17,20,22,24,30]. These rely on the key concept of the lower approximation, or region of certainty, for evaluating the goodness of a feature subset in the process of determining an optimal reduct, such as dependency function [20,22] and the significance of attributes[17,24]. Although this concept has been successfully applied to numerous FS problems, the approaches neglect the information that is contained in the boundary region or the region of uncertainty. Therefore, using the information from the lower approximation alone is insufficient for efficient feature selection when applied to data in which no equivalence class is consistent. In addition, ignoring the information contained in the inconsistent region during the feature selection process may lead to a loss of much valuable information. While there are some researches based on RS which determine the boundary region information [31,32], these approaches determine by using only the knowledge of the upper approximation as a whole, rather than considering the lower approximation and the boundary region, which are supposed to be conceptually separated. Therefore, some papers have successfully applied the method to solve several problems [2,33] which consider the lower approximation and the boundary region separately.

We can divide feature selection methods into 2 categories: filter methods and wrapper methods [34,35]. Filter methods select a subset of features as a preprocessing step which is independent from the learning algorithm. Meanwhile, wrapper methods utilize the performance of the learning algorithm to evaluate the worth of feature subsets. Furthermore, we can roughly divide feature selection algorithms into 2 categories: discrete methods [22,36-39] and numerical methods [30,40]. However, those methods require the numerical features to be discretized before applying the FS techniques in order to segment the numerical features into several intervals and form discretized data sets. Similarly, discretization of numeric data is required in order to apply them to feature selection based on rough set theory [17-25].

Formally, discretization of numerical attributes does not determine the degrees of membership of numerical values to discretized values. Therefore, essential information or attributes may be lost. There are at least 2 categories of structure lost: neighborhood structure and order structure in real spaces [30]. Obviously, the distances between samples are different in real spaces, but similar in discretized spaces. Therefore, it is unreasonable to measure the similarity of discretized attributes in numerical methods with Euclidean distance. In paper [30], the authors introduce a neighborhood rough set model for heterogeneous feature subset selection and attribute reduction. In this method, neighborhood relations are also used to generate a family of the objects by using distances to measure the similarity. Therefore, the samples in the same neighborhood granule (family) are closer to each other, compared with those in the different neighborhood granule. However, evaluating the goodness of a feature subset still uses the concept of lower approximation based on rough set theory. Therefore, considering without boundary region information is not sufficient for feature selection in the case of dealing with high-dimensional or highly-noisy data.

This paper presents a feature selection method which is based on neighborhood rough sets and mutual information. The neighborhood granular of each sample is computed by measuring the Euclidean distance of the samples to each other. The samples in the same neighborhood granule are equivalent to equivalence classes of the classical rough set. This proposed method can be applied to both numerical and mixture features, and the discretization process of numeric data is not required. This method determines the different amounts of information in the lower approximation and the boundary region in order to select the feature subsets. Noisy data has little influence on the results that can be produced by our

proposed method. It can also result in outperformance of the classification accuracies, compared to those obtained by RS dependency-based approaches.

The remainder of this paper is structured as follows. Section 2 summarizes the theoretical background of neighborhood rough sets and mutual information. In Section 3, we propose an approach for feature selection based on neighborhood rough sets and mutual information. The pseudo-code of our algorithm is also presented in this section. Section 4 compares the proposed method with some current approaches, by running experiments for some data sets from the University of California, Irvine (UCI). Section 5 concludes the method proposed in this paper and points out some future research tasks.

Background

In this section, the basic concepts of the theories of neighborhood rough sets and mutual information based on rough sets are described.

A: Neighborhood rough sets

Let $IS = (U, A)$ be an information system, where U is a finite nonempty set of n objects $\{x_1, x_2, ..., x_n\}$, A is a finite nonempty set of attributes $\{a_1, a_2, ..., a_m\}$ used to describe the samples, and $f(x, a)$ is the feature value of sample x. Formally, $<U, A>$ 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.

For $x_i \in U$ and $B \subseteq C$, a neighborhood $\delta_B(x_i)$ of x_i in subspace B is defined as;

$$\delta_B(x_i) = \{x_j \mid x_j \in U, \Delta_B(x_i, x_j) \le \delta\} \quad (1)$$

where Δ is a metric and δ is neighborhood size. This relation means that, for all x_1, x_2, and x_3 in U, it satisfies the following 3 conditions:

1) $\Delta(x_1, x_2) \ge 0$, 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) $\Delta(x_1, x_3) \le \Delta(x_1, x_2) + \Delta(x_2, x_3)$.

Let x_1 and x_2 be 2 samples in m dimensional space $A = \{a_1, a_2, \ldots, a_m\}$, $f(x,a_i)$ denotes the value of sample x in the i^{th} attribute a_i, then the Minkowsky distance is defined as;

$$\Delta_P(x_1, x_2) = \left(\sum_{i=1}^{m} \left| f(x_1, a_i) - f(x_2, a_i) \right|^P \right)^{1/P} \quad (2)$$

From the above well-known distance measure, $P = 1$ represents the Manhattan distance (L_1 norm), $P = 2$ represents the Euclidean distance (L_2 norm), and $P = \infty$ is the distance for the Tchebyshev average (L_∞ norm).

The Heterogeneous Euclidean-Overlap Metric function (*HEOM*) has been proposed for distance measuring between samples which contain both numerical and categorical attributes. The *HEOM* distance between samples x and y, $HEOM(x,y)$, can be calculated as;

$$HEOM(x, y) = \sqrt{\sum_{i=1}^{m} d_{a_i}(x_{a_i}, y_{a_i})^2} \quad (3)$$

where $d_{a_i}(x, y)$ is the distance between samples x and y on different types of attributes, defined as;

$$d_{a_i}(x, y) = \begin{cases} 1 & \text{, if } x \text{ or } y \text{ is unknown} \\ overlap(x, y) & \text{, if } a \text{ is a nominal attribute} \\ rn\_dif_a f(x, y) & \text{, if } a \text{ is a numerical attribute} \end{cases}$$

here

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

and

$$rn\_diff_a(x,y)=\frac{|x-y|}{\max_a-\min_a}.$$

Given $<U, N>$ and $X \subseteq U$, the lower and upper approximations of X in terms of a neighborhood relation N are defined as;

$$\underline{NX}=\left\{x_j \mid \delta_B(x_j)\subseteq X, x_j\in U\right\}$$
$$\overline{NX}=\left\{x_j \mid \delta_B(x_j)\cap X\neq\phi, x_j\in U\right\}. \quad (4)$$

For a neighborhood decision table (NDT; NDT = $<U,C,D>$), $X_1, X_2, \ldots, X_l$ are the sample subsets with decisions D, and the lower and upper approximations of decision D with respect to attributes B are then defined as

$$\underline{N_B}D=\bigcup_{i=1}^{l}\underline{N_B}X_i$$
$$\overline{N_B}D=\bigcup_{i=1}^{l}\overline{N_B}X_i \quad (5)$$

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

$$POS_B(D)=\underline{N_B}D \quad (6)$$

$$BN(D)=\overline{N_B}D-\underline{N_B}D. \quad (7)$$

Note that, according to the above definitions of approximation sets, the lower approximation of set X can be interpreted as the collection of objects whose neighborhood sets can be classified into X. The upper approximation of X includes all the neighborhood sets that cannot be classified into $-X$. Finally, the boundary region is the subset of objects whose neighborhood comes from more than one decision class.

B: Mutual information based on rough sets

The information theory proposed by Shannon [41] provides useful tools to measure the information of a data set with entropy and mutual information. The mutual information is a measure of generalized correlation between 2 random variables, and can also be interpreted as the amount of information shared by 2 random variables. In information system, entropy can be an information measure for feature selection on probabilistic knowledge about a given feature.

In RS theory, an equivalence relation induces a partition of the universe. The partition can be regarded as a type of knowledge. The meaning of knowledge in information theoretical framework of rough sets is interpreted as follows.

For any subset $B \subseteq A$ of features, let $U/IND(B) = \{X_1, X_2, ..., X_n\}$ denote the partition induced by the equivalence relation $IND(B)$. The information entropy of knowledge B, $H(B)$, is defined as;

$$H(B) = -\sum_{i=1}^{n} p(X_i)\log(p(X_i)), \tag{8}$$

where $p(X_i) = \frac{|X_i|}{|U|}, 1 \le i \le n$.

Let B and D be the subset of A. Let $U/IND(B) = \{X_1, X_2, ..., X_n\}$, $U/IND(D) = \{Y_1, Y_2, ..., Y_m\}$ denote the partitions induced by the equivalence relations $IND(B)$ and $IND(D)$, respectively. The conditional entropy of knowledge D given by the knowledge B, $H(D|B)$, is defined as;

$$H(D \mid B) = -\sum_{i=1}^{n} p(X_i)\sum_{j=1}^{m} p(Y_j \mid X_i)\log(p(Y_j \mid X_i)), \tag{9}$$

where $p(X_i) = \frac{|X_i|}{|U|}, p(Y_j \mid X_i) = \frac{|Y_j \cap X_i|}{|X_i|}, 1 \le i \le n, 1 \le j \le m$.

The mutual information is a measure of the amount of information that knowledge B contains about knowledge D, which is defined as;

$$I(D;B) = \sum_{j=1}^{m}\sum_{i=1}^{n} p(Y_j, X_i)\log\frac{P(Y_j, X_i)}{P(Y_j)P(X_i)} \tag{10}$$

where $p(X_i) = \frac{|X_i|}{|U|}, p(Y_j, X_i) = \frac{|Y_j \cap X_i|}{|U|}, 1 \le i \le n, 1 \le j \le m$.

If the mutual information between B and D are large (small), it means B and D are closely (not closely) related. The relation between the mutual information and the entropy can be defined as;

$$I(B;D) = H(B) - H(D \mid B). \tag{11}$$

When applying mutual information in feature selection, mutual information plays a key role in measuring the relevance and redundancy among features. The main advantages of mutual information are its robustness to noise and geometrical transformations such as rotation, translation and scaling. In this paper, mutual information is used as an information measure of correlation between the lower approximation $\underline{N}X$ and class X. Furthermore, mutual information of the boundary region $BN(D)$ with respect to decision class is measured. More details on information measuring of the lower approximation and the boundary region can be seen in the next section.

Feature selection based on neighborhood rough sets and mutual information

In this section, we describe the problem of FS by using RS dependency-based approaches in which the equivalence classes are inconsistent. In addition, the concept of dividing the sample set into decision positive regions and decision boundary regions, which is a concept used for finding the set of certainty and uncertainty, are described in this section. From these concepts, we present a strategy for feature subset selection based on the uncertainty information minimization and certainty information maximization. This idea yields a nonempty set of reducts when it is applied to the data sets in which all equivalence classes are inconsistent in terms of a single feature.

A: Problems of rough set-based feature selection methods

As discussed previously, most existing RS-based FS approaches rely on the information of the lower approximation for evaluating the goodness of a feature subset in determining an optimal subset. Many approaches based on the theory of RS have employed the dependency function, which is based on the lower approximation as an evaluation step in the FS process. Unfortunately, these RS-based approaches yield an empty set of reducts when they are applied to data in which no equivalence class is consistent in terms of a single feature because the dependency of each single feature is zero.

Figure 1 illustrates the idea for a binary classification problem in a one-dimensional space. The class probability density function of the feature space is divided into 3 parts: 1) a consistent region of class 1 (ω_1); 2) a consistent region of class 2 (ω_2); and 3) an inconsistent region of between class 1 and class 2. The inconsistent region contains samples with the same feature values, but which belong to different classes.

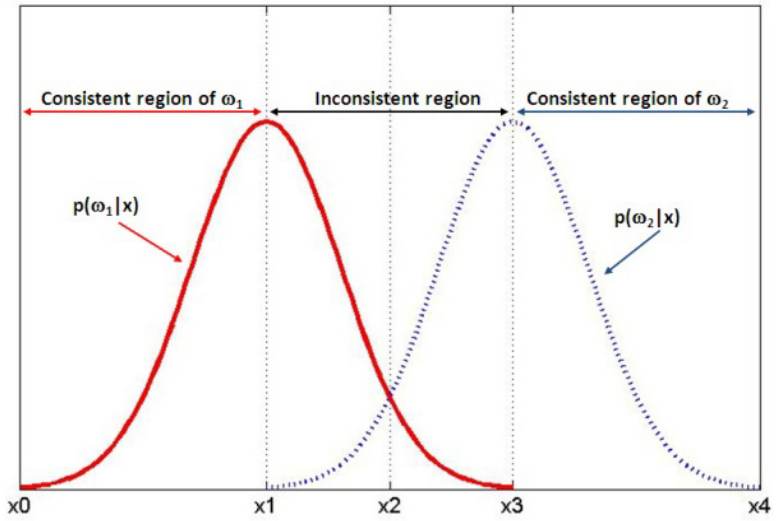

Figure 1 Binary classification in a 1-D numerical feature space.

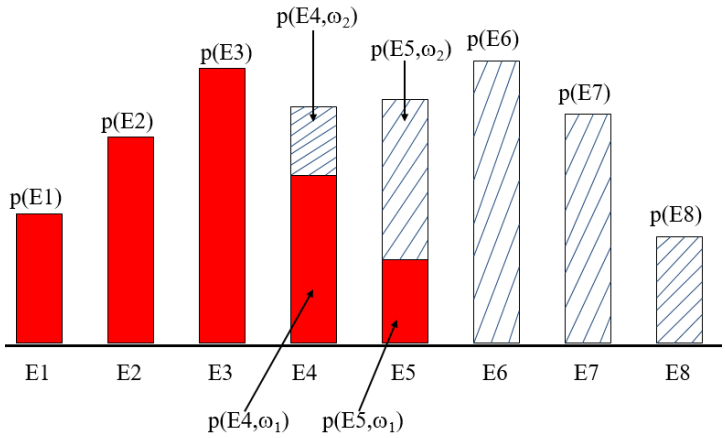

Figure 2 Equivalence classes in a 1-D discrete feature space.

Figure 2 shows a similar case in discrete spaces, where the samples are divided into a set of equivalence classes {*E1, E2, . . . , E8*} based on their feature values. Samples with the same feature values are grouped into one equivalence class. The height of the rectangles in **Figure 2** denotes the probability $p(Ei)$ of the equivalence class, and $p(\omega_i,Ej)$ is the joint probability of ω_i and Ej. We can observe that the equivalence classes are consistent, because each equivalence class is composed of samples from the same class, e.g., *E1*, *E2*, *E3*, *E6*, *E7*, and *E8*. However, some equivalence classes are inconsistent, like *E4* and *E5*, where samples with the same feature values are assigned to different classes. Therefore, from **Figure 2**, the RS dependency-based approaches yield a nonempty set of reducts when they are applied to data in which some equivalence class is consistent in terms of a single feature, because the dependency of some single features is nonzero.

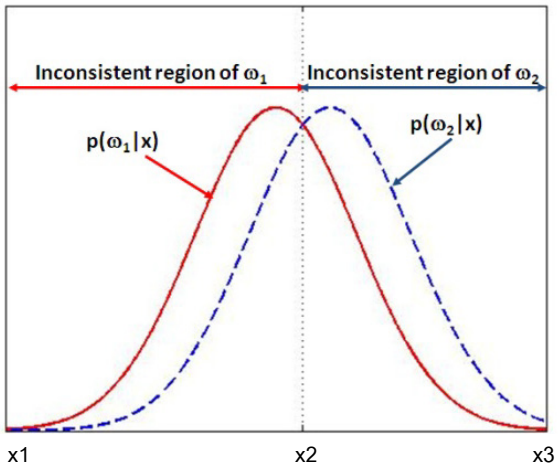

Figure 3 Inconsistent in a 1-D numerical feature space.

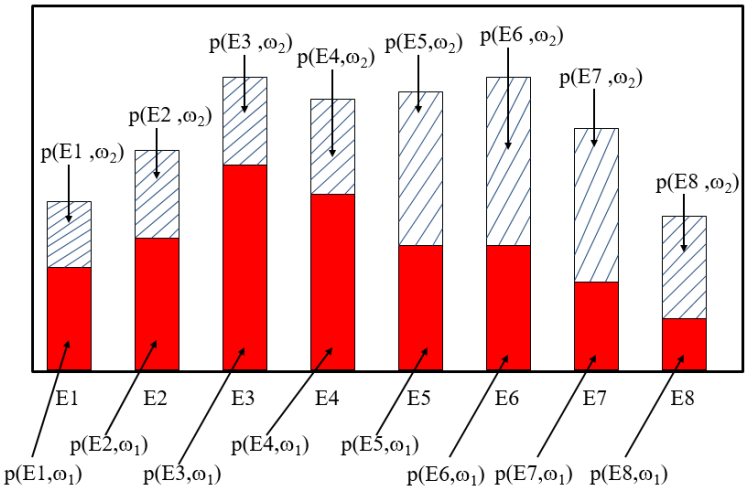

Figure 4 Inconsistent in a 1-D discrete feature space.

However, we may face problems when we apply the RS dependency-based approach to noisy data or combined features (e.g., XOR). These problems may affect the equivalence classes that lead to inconsistence, as shown in **Figures 3** and **4**. Therefore, in the case of all equivalence classes for each feature discrete space being inconsistent, the RS dependency-based approaches yield an empty set of positive regions because the dependency of each single feature is zero.

Hu [42] described the problem of feature selection based on RS-dependency in the case of numerical feature spaces, in that it consists of only inconsistent regions of class, as shown in **Figure 3**. In addition, the author described that the inconsistent feature spaces that lead to the neighborhood of any sample would not be "pure" (homogeneous), and the samples in it would come from 2 classes. Furthermore, in this case, the dependency of neighborhood is zero, whereas the probabilities of Bayes errors are less than 1. Dependency cannot present the differences in information between these classes. Therefore, the author used the concept of the Bayes error rate for considering the neighborhood of sample between the positive region and boundary region. Subsequently, he defined the neighborhood decision function ND(x) as follows:

Given NDT = $<U, C, D>$, $x_i \ U$, $\delta(x_i)$ 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 $ND(x_i) = \omega_l$ 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 $\delta(x_i)$.

From the above definition, Hu [42] used the ND(x) for dividing the sample of neighborhood into decision positive regions and decision boundary regions. Furthermore, Hu introduced the neighborhood decision error rate (NDER) to compute the averages of samples which were in the decision boundary regions. In addition, he proposed the Neighborhood Decision Error Minimization (NDEM) which is a procedure for feature selection involving minimizing the NDER or maximizing the 1-NDER in different feature subsets. Hu demonstrated that the NDEM can tolerate noisy data better than RS-based dependency, and can also be directly applicable to numerical data without data discretization.

B: Certainty set and uncertainty set based on neighborhood decision

As described previously, the RS dependency-based approaches raise problems with the discrete feature space, in that the equivalence classes are inconsistent. Therefore, in this paper, we propose a feature selection method based on the neighborhood error rate and the mutual information for measuring the goodness of feature subset. Evaluating the goodness of feature subset by using mutual information can provide a finer determination than the 0-1 loss function that is used in the method of NDEM [42]. This is because the NDEM method determines the loss function with value 1 for misclassified (boundary) samples and 0 for classified (positive) samples. Furthermore, the NDEM considers only information in the boundary region, but the positive region is neglected. In this paper, we use both the information contained in the lower approximation and the boundary region in the feature selection strategy. This proposed approach selects the feature that gives the lower approximation information that is mostly relevant to class.

In this paper, we have defined the decision positive region (*DPR*) and decision boundary region (*DBR*) based on the neighborhood decision function. *DPR* is a set of samples determined to belong to the lower approximation, while *DBR* is a set of samples that is determined to belong to the boundary region. Both *DPR* and *DBR* are based on the concept of neighborhood decision, and are defined as follows.

Given NDT = $<U, C, D>$, $U/IND(D) = \{X_1, X_2, , X_l\}$, $\{\omega_1, \omega_2, .., \omega_l\}$ is a set of decision values, δ is neighborhood size , and the *DPR* (Certainty set) of decision value ω_i with respect to conditional attribute $B \subseteq C$ and with δ value is defined as;

$$DPR_{B\delta}(\omega_i) = \left\{ x_j \mid x_j \in U \wedge ND(x_j) = \omega(x_j) \ where \ \omega(x_j) = \omega_i \right\}, \tag{12}$$

and the *DBR* (Uncertainty set) of decision attribute D with respect to B and with δ value is defined as;

$$DBR_{B\delta}(D) = \left\{ x_j \mid x_j \in U \wedge ND(x_j) \neq \omega(x_j) \right\}. \quad (13)$$

For a subset of features B and δ value, the mutual information of the Uncertainty set $DBR_{B\delta}(D)$ with respect to knowledge D can be defined as;

$$UI(B,\delta) = I(D; DBR_{B\delta}(D)) \quad (14)$$

The total information of mutual information between the Certainty set $DPR_{B\delta}(\omega_i)$ and the equivalence class X_i with respect to B and δ value, denoted by $CI(B, \delta)$, can be defined as;

$$CI(B,\delta) = \sum_{i=1}^{l} I(X_i, DPR_{B\delta}(\omega_i)) \quad (15)$$

Hence, the problem of selecting feature subset B is equivalent to the maximizing of $CI(B, \delta)$ and the minimizing of $UI(B, \delta)$, that is, to maximize the objective function $E(B, \delta)$, where;

$$E(B, \delta) = CI(B, \delta) - UI(B, \delta). \quad (16)$$

Obviously, if $CI(B, \delta) = H(D)$, and the objective function $E(B, \delta)$ value is maximum, it shows that the approximate information contains no uncertainty with respect to B and δ. Therefore, a subset of features B is determined as being strongly relevant features. Conversely, if $UI(B, \delta) = H(D)$, then B and δ bring about the approximating of information that has the highest uncertainty. Consequently, a subset of features B is determined as being irrelevant features that has no useful information related to decision attribute D. Different amounts of both values are obtained as both operate in the range of $[0; H(D)]$, and the $E(B, \delta)$ has a value in the range of $[-H(D); H(D)]$. A new feature selection mechanism can be constructed by using the different amount of information between the certainty value and uncertainty value to guide the search for the best feature subset.

C: mUMCNR feature selection algorithm

In this section, we will present an algorithm for feature selection using the objective function E, as defined above, to evaluate the goodness of feature subset. **Figure 5** shows the *mUMCNRREDUCT* algorithm. *mUMCNRREDUCT* is based on the idea of maximum certainty and minimum uncertainty. The proposed method is a searching scheme to find a superset for all candidates reducting with the value of δ, which varies from 0.02 to 0.2 in the step of 0.02. Here, the parameter δ is the size of the neighborhood of sample in a numerical feature space. Therefore, δ is used as a parameter for controlling the number of samples in the boundary and the effect of noise.

Algorithm: mUMCNRREDUCT (C,D)
Input: decision table <U, C, D>;
delta //control the size of the neighborhood
Output: feature subset R.
1: $R \leftarrow \phi, T \leftarrow \phi$
2: do while $C - R \neq \phi$
3: for each $f_i \in C - R$
4: compute $E(R \cup \{f_i\}) = CI(R \cup \{f_i\} - UI(R \cup \{f_i\})$
5: end
6: select the attribute f_k that satisfies the condition:
7: $E(R \cup \{f_k\}) = \max_i(E(R \cup \{f_i\}))$
8: if $E(R \cup \{f_k\}) > E(R)$
6: $R \leftarrow R \cup \{f_k\}$
7: else break
8: end
9: return R

Figure 5 The mUMCNRREDUCT algorithm.

Each candidate reduct is calculated by considering the δ value. Therefore, the maximum number of a candidate reduct equals the number in step of the divided δ interval.

The *mUMCNRREDUCT* algorithm uses the maximum value of objective function *E* value of a subset to guide a candidate reduct selection process. If the *E* value of the current reduct is greater than that of the previous one, then this subset is retained and used in the next iteration of the loop. A candidate reduct selection process terminates when an addition of any remaining features results in the value of the objective function *E* reaching the information entropy of the decision classes. In addition, if the *E* value of the current candidate reduct is not better than the previous one, then the *mUMCNRREDUCT* algorithm will be terminated as well.

The proposed *mUMCNRREDUCT* algorithm works on the idea of greedy search for the feature selection process. The algorithm begins with an empty subset *R*. The do while loop works by calculating the *E* value of a subset and incrementally adding a single conditional attribute at a time. For each iteration, a conditional attribute a_k that does not belong to *R* will be temporarily added to subset *R* to compute the value of the objective function *E* (line 4). At the same time, the attribute a_k that yields the maximum *E* value will be selected to compare with the previous subset *R* (line 7). If the information of the current subset *R* $\{a_k\}$ is greater the previous subset (*R*), then the attribute added in (line 8) is retained as part of the new subset *R*.

We now analyze time complexity of *mUMCNRREDUCT*, before an empirical study of its efficiency is done. There are several main steps in this proposed algorithm. However, the proposed algorithm can calculate time complexity the same as the NDEM [42]. We can summarize the steps to calculate time complexity of the algorithm is as follows. First, the sorting technique and a sliding windows technique are used to find the neighborhood of each sample. Second, the neighborhood of a sample in a multidimensional space with the intersection of the neighborhoods of a sample in each feature space is computed. Then, the class probability of the neighborhood of each sample is calculated. Subsequently, the class probability of the neighborhood of each sample is calculated. Finally, the goodness of the remaining attributes is evaluated and the attributes are added into the *R* one by one. Therefore, the overall time complexity is Nm ($n \log n + kn + n$), where N is candidate attributes, m is selected attributes, and k is a constant value for searching the neighborhood of each sample.

Experimental results and discussion

In this section, we first test the influence of parameter δ on estimation for all candidate reducts, with the value of δ varying from 0.02 to 0.2 in the step of 0.02 increments. An optimal reduct of each classifier is selected from the candidate reducts with the highest predictive accuracy. Then, the results of mUMCNR-based feature selection are compared to some existing techniques.

A: The influence of the size of the neighborhood δ on mUMCNR-based feature selection

In this section, we show the influence of the neighborhoods size on the number of the selected features and an optimal subset of features for the learning algorithm on 15 data sets from the UCI Machine Learning Repository (see **Table 1**) [43]. We also consider 3 well-known learning algorithms, named SVM, C4.5, and PART, and estimate an optimal subset and classification accuracy based on a tenfold cross validation.

To show the influence of the sizes of parameter δ, we consider a series of numeric values varying from 0.02 to 0.2 in the step of 0.02. For each value of δ, we are able to get a candidate reduct. Therefore, from the size of the neighborhoods from 0.02 to 0.2, we are able to get not more than 10 candidate reducts. However, there may be some values of δ in which the reduct will be empty, because all equivalence classes have an information of uncertainty set that is greater than the certainty set for each single feature. When considering both the subset size and classification accuracy, the δ in the range of 0.02 - 0.2 is the best range. Besides, in the case where δ is greater than 0.2, each neighborhood increases the number of samples that come from different classes. Therefore, the uncertainty set of samples is greater than the certainty set, and leads to the problem of local maximal.

Table 1 Description of UCI benchmark data sets.

No.	Data	Type	Samples	Numerical	Categorical	Class
1	Wine	numerical	178	13	0	3
2	Sonar	numerical	208	60	0	2
3	Ionos	numerical	351	34	0	2
4	Wdbc	numerical	569	31	0	2
5	Parkinsons	numerical	195	22	0	2
6	Cleveland	numerical	297	13	0	2
7	Glass	numerical	214	9	0	6
8	Votes	category	690	0	15	2
9	Soybean	category	683	0	35	19
10	Lymphography	category	148	0	18	4
11	Promoters	category	106	0	57	2
12	Ecoli	mixed	336	5	2	8
13	Heart	mixed	270	7	6	2
14	Hepatitis	mixed	155	6	13	2
15	German	mixed	1000	7	13	2

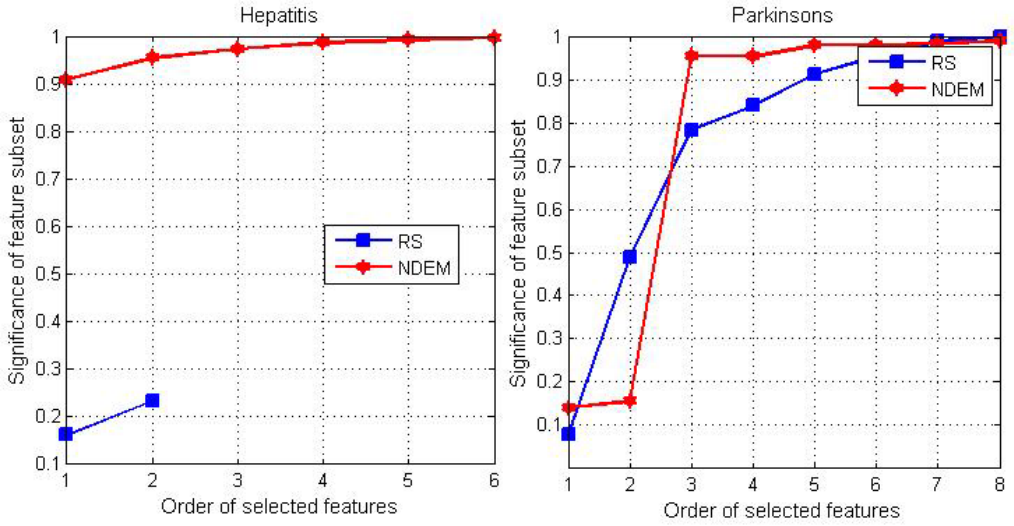

Figure 6 Attribute significance versus the number of selected features.

Figures 6 - 8 show changes of attribute significance with the number of selected features obtained for Hepatitis and Parkinsons data. The significance of the feature subset is computed with the RS-based dependency, NDEM-based NDER [42], and mUMCNR, respectively. The values of mUMCNR rapidly grow with relatively low size of δ before the set of features is formed. Then, the growth slows down until it completely stops when the value of significance has equaled the entropy of decision classes. The value of RS stops early at 2 features, with a dependency value of less than 0.4 on the Hepatitis data. Therefore, RS encounters the local maximum and is unsuccessful when applied to Hepatitis data. Meanwhile, the values of mUMCNR and NDEM proceeded to rapidly increase on Hepatitis data. mUMCNR can be successful when applied to Hepatitis data that finishes with the entropy of decision classes in every value of δ . However, NDEM terminates at 6 features, with the significant value of features subset as 0.99, which is an incomplete value.

With regard to features selection of Parkinsons data, the values of NDEM proceeded to gradually increase. NDEM finishes at 8 features, with the significant value of 0.98. At the same time, RS stops at 8 features, with the dependency value equal 1.0 on the Parkinsons data. The values of mUMCNR rapidly grow and it completely finishes at the entropy of the decision classes. The set of features is constituted with 4 features on δ = 0.02, 0.04, 0.06, 0.08, and 0.1. On other values of δ, mUMCNR yields an empty set, where the information of uncertainty set is greater than the certainty set. We can observe that NDEM encounters the local maximum on both Hepatitis data and Parkinsons data. However, mUMCNR can be successful when applied to Hepatitis data and Parkinsons data that finish with the entropy of decision classes. Therefore, we can observe that the efficiency of mUMCNR for selecting the subset of features in the phenomenon is greater than RS and NDEM on the Hepatitis data and Parkinsons data.

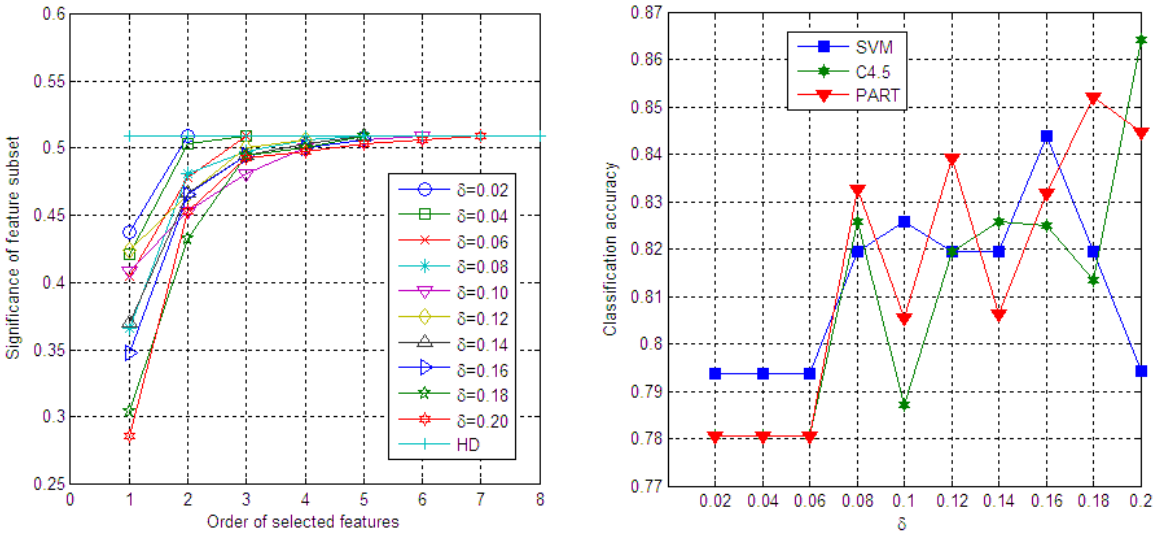

Figure 7 Attribute significance and classification accuracy on Hepatitis data with different sizes of neighborhoods.

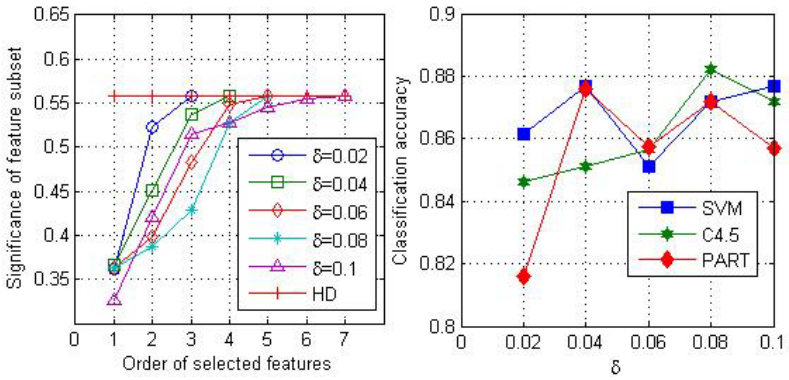

Figure 8 Attribute significance and classification accuracy on Parkinsons data with different sizes of neighborhoods.

The classification accuracy of the candidate subsets with the values of δ on both Hepatitis and Parkinsons data is shown in **Figures 7** and **8**. The subset of features with the highest accuracy of classifier is chosen as an optimal subset, to compare performance with other FS techniques. We can see that the highest accuracy on Hepatitis data of SVM, C4.5, and PART is given when $\delta = 0.16$, 0.2, and 0.18, respectively. While dealing with the Parkinsons data, the highest accuracies of SVM, C4.5, and PART is given with values of δ are 0.04, 0.08, and 0.04, respectively. The optimal subset of features and the values of parameter with the highest predictive accuracy of each learning algorithm are shown in **Table 2**. From **Table 2**, we see that each data set of the discrete data, including the Votes, Soybean, Lymphography, and Promoters, has only a single candidate reduct. This is because the size of 0.02 to 0.2 gets the same reduct of every value δ. In comparison, the performances of the proposed method with the RS-based attribute reduction method are compared to some existing classical techniques. The highest predictive accuracy on a learning algorithm is selected to compare its performance with these methods.

Table 2 The optimal subset and the δ values for classifier.

No.	Data	SVM		C4.5		PART	
		size	δ	size	δ	size	δ
1	Wine	6	0.2	3	0.02	3	0.02
2	Sonar	10	0.18	4	0.06	6	0.12
3	Ionos	11	0.2	10	0.1	10	0.1
4	Wdbc	11	0.14	4	0.02	4	0.02
5	Parkinsons	4	0.04	5	0.08	4	0.04
6	Cleveland	10	0.12, 0.18	11	0.16	11	0.18
7	Glass	6	0.16	8	0.12	8	0.2
8	Votes	7	-	7	-	7	-
9	Soybean	14	-	14	-	14	-
10	Lymphography	8	-	8	-	8	-
11	Promoters	4	-	4	-	4	-
12	Ecoli	7	0.18 - 0.2	6	0.04 - 0.08	6	0.04 - 0.08
13	Heart	6	0.02	6	0.02	6	0.02
14	Hepatitis	5	0.16	7	0.2	5	0.18
15	German	8	0.02	10	0.06	12	0.1-0.16

B: Comparison of feature selection algorithms on UCI benchmark data sets

In this paper, we experiment with different algorithms of feature selection using fifteen data sets, as detailed in **Table 1**, where 4 data sets come with discrete features (i.e., Votes, Soybean, Lymphography, and Promoters), 7 data sets come with numerical features (i.e., Wine, Sonar, Ionos, WDBC, Parkinsons, Cleveland, and Glass), and the rest of the data sets come with mixed numerical and categorical features. Before applying all of the feature selection techniques, the numerical features are required to be discretized by the minimum descriptive length (MDL) discretization [28], in order to segment the numerical features into several intervals and form the discretized data sets. Meanwhile, we also apply ReliefF, NDEM, and mUMCNR to directly select the continuous features which are normalized into [0, 1]. We set $\delta = 0.14$ for experiments of the NDEM [42]. Next, we use sequentially greedy forward search to form the best features when comparing the algorithms that evaluate features based on RS-dependency function (RS), consistency-based subset (CNS) [37], and correlation-based feature selection (CFS) [36], respectively. In addition, the proposed method is compared with ReliefF [40] which has special searching strategies.

We first show the results of discrete, numerical, and mixed feature selection. The number of selected features of the data is given in **Table 3**, where the last 3 columns are the average number of features on 3 classifiers of the mUMCNR method. For each data set, mUMCNR gives not more than 10 subsets, with δ in the range of 0.02 - 0.2. Subsequently, a subset of features that has the highest predictive accuracy for SVM, C4.5, or PART is chosen as the optimal subset of classifier (as described in **Table 2**). For example, on Wine data, 10 candidate subsets are reduced to 3 subsets which yield the highest predictive accuracy for 3 classifiers. On SVM, we select $\delta = 0.2$, which gives the highest predictive accuracy, with subset size as 6 features. Meanwhile, on C4.5 and PART, we select $\delta = 0.02$, which achieves with the highest predictive accuracy, and subset size as 3 features for both classifiers. Therefore, the subset size and classification accuracy of SVM, C4.5, and PART will compare favorably to the performance to other methods.

Conversely, the NDEM method [42] achieves only one features subset on each data set with the neighborhood size valued 0.14. Then, the selected subset is applied to SVM, C4.5, and PART in order to compare the classification accuracies with mUMCNR and other methods. Meanwhile, RS, CNS, CFS, and ReliefF come with one feature subset on each data set, the same as NDEM. Furthermore, the classification accuracies for SVM, C4.5, and PART of the selected subset with these methods are compared to mUMCNR and NDEM.

Table 3 Number of selected features with different techniques.

No.	Data	unselect	RS	NDEM	CNS	CFS	ReliefF	mUMCNR (SVM)	mUMCNR (C4.5)	mUMCNR (PART)
1	Wine	13	5	5	4	7	4	6	3	3
2	Sonar	60	-	7	4	12	4	10	4	6
3	Ionos	34	8	9	5	5	8	11	10	10
4	Wdbc	31	7	6	6	8	8	5	4	4
5	Parkinsons	22	8	7	5	7	5	4	5	4
6	Cleveland	13	-	8	7	5	9	10	11	10
7	Glass	9	-	5	7	7	6	6	8	8
8	Votes	15	-	8	8	3	11	7	7	7
9	Soybean	35	13	11	11	21	19	14	14	14
10	Lymphography	18	6	7	7	9	9	8	8	8
11	Promoters	57	4	4	4	6	5	4	4	4
12	Ecoli	7	-	7	7	6	3	7	6	6
13	Heart	13	-	11	11	8	7	6	6	6
14	Hepatitis	19	-	6	12	10	8	5	7	5
15	German	20	-	11	14	3	9	8	10	12
	Average	**24.40**	**N/A**	**7.47**	**7.47**	**7.80**	**7.67**	**7.40**	**7.13**	**7.13**

Table 4 Classification accuracy of SVM classifier.

No.	Data	unselect	RS	NDEM	CNS	CFS	ReliefF	mUMCNR
1	Wine	0.9775	0.9831	0.9775	0.9325	0.9775	0.9550	**0.9944**
2	Sonar	0.7596	N/A	0.7933	0.7211	0.7644	0.7548	**0.8125**
3	Ionos	0.8860	0.8291	0.9402	0.8148	0.8689	0.8262	**0.9487**
4	Wdbc	**0.9772**	0.9543	0.9666	0.9648	0.9630	0.9402	0.9736
5	Parkinsons	0.8718	**0.8769**	0.8718	0.8666	0.8512	0.8512	**0.8769**
6	Cleveland	0.8283	N/A	0.8047	0.8215	**0.8383**	**0.8383**	0.8215
7	Glass	0.5748	N/A	0.5697	0.5748	0.5841	0.5280	**0.5939**
8	Votes	0.9433	N/A	0.9400	0.9333	0.9433	0.9433	**0.9467**
9	Soybean	**0.9385**	0.8306	0.8608	0.8436	0.9218	0.9250	0.9136
10	Lymphography	0.8311	0.8175	0.7971	0.7972	0.8243	**0.8445**	0.7771
11	Promoters	0.9340	0.8584	0.8582	0.8584	0.9150	**0.9433**	0.8773
12	Ecoli	**0.8393**	N/A	**0.8393**	**0.8393**	0.8333	0.7560	**0.8393**
13	Heart	0.8296	N/A	0.8370	0.8333	0.8296	**0.8444**	**0.8444**
14	Hepatitis	**0.8516**	0.7935	0.8196	0.8323	0.8323	0.8323	0.8438
15	German	0.751	N/A	**0.758**	0.756	0.717	0.747	0.756
	Average	**0.8529**	**N/A**	**0.8423**	**0.8260**	**0.8443**	**0.8353**	**0.8546**

In **Table 3**, we observe that most of the features in the raw data have been deleted by all the feature selection algorithms. At the same time, we then apply SVM, C4.5, and PART classifiers to each of the newly obtained data sets (with only selected features), and obtain the average accuracy of 10-fold cross validation. The results show that these algorithms are effective in retaining the classification ability. The RS algorithm yields an empty set when it is applied to the "Sonar", "Cleveland", "Glass", "Vote", "Ecoli" "Heart", "Hepatitis", and "German" datasets, because all equivalence classes are inconsistent at the first stage. In this case, the positive region of each single feature is an empty set. However, all the other feature

selection algorithms can determine the subset of features. We can also find that the subset contains different features when applying different algorithms.

Noisy data had a great influence on the results that were produced by the RS algorithms. mUMCNR is based on the idea of dealing with noise with a parameter that controls the noise effect. The noisy instance has little influence on the feature selection process on both mUMCNR and NDEM. However, NDEM considers only the number of samples in the decision boundary region alone. Therefore, measuring the goodness of feature subset is not accurate enough. Besides, considering only samples in the boundary region and neglecting the positive region may lead to losing valuable attributes. At the same time, the mUMCNR method determines both information of the decision positive region and the decision boundary region simultaneously. Therefore, mUMCNR is able to create a subset of features which contains more valuable information than those obtained using NDEM. We can see that mUMCNR demonstrated better performance than NDEM, as shown by the experimental results.

Table 5 Classification accuracy of C4.5 classifier.

No.	Data	unselect	RS	NDEM	CNS	CFS	ReliefF	mUMCNR
1	Wine	0.9382	0.9494	0.9213	0.9662	0.9438	0.9550	**0.9719**
2	Sonar	0.7115	N/A	0.7644	0.7500	0.7163	0.7211	**0.7692**
3	Ionos	0.9145	0.9202	0.9288	0.9031	0.9088	0.8774	**0.9316**
4	Wdbc	0.9332	0.9525	0.9473	0.9420	0.9332	0.9297	**0.9684**
5	Parkinsons	0.8000	0.8718	0.8513	0.8820	0.8307	0.8461	**0.8821**
6	Cleveland	0.7778	N/A	0.8182	0.7811	**0.8249**	0.7878	0.8182
7	Glass	0.6729	N/A	0.6591	0.6449	0.6916	**0.7290**	0.6955
8	Votes	0.9367	N/A	0.9367	0.9367	**0.9400**	0.9333	0.9367
9	Soybean	**0.9151**	0.8175	0.8185	0.8045	0.8208	0.8436	0.9034
10	Lymphography	0.7635	0.7364	0.7433	0.7432	0.7637	0.7740	**0.7762**
11	Promoters	0.8113	0.8490	0.8491	0.8490	0.8301	0.8301	**0.8773**
12	Ecoli	**0.8423**	N/A	**0.8423**	**0.8423**	**0.8423**	0.7649	**0.8423**
13	Heart	0.8000	N/A	0.8111	0.7926	0.8037	0.7963	**0.8185**
14	Hepatitis	0.8387	0.7935	0.8454	0.8323	0.8129	0.8258	**0.8642**
15	German	0.705	N/A	0.7070	0.7260	0.7050	0.7290	**0.7320**
	Average	**0.8240**	**N/A**	**0.8296**	**0.8264**	**0.8245**	**0.8229**	**0.8525**

Among the fifteen data sets and 6 algorithms of feature selection, mUMCNR on PART comes with the minimal number of features, with 6 data sets; meanwhile, mUMCNR on SVM and C4.5 obtains the minimal number of features, with 4 data sets and 5 data sets, respectively. On average, mUMCNR on SVM, C4.5, and PART selects 7.40, 7.13, and 7.13, respectively, features for dimensionality reduction, which are the least 3 values among the size of the features that the 6 algorithms are applied to. With regard to the performance of SVM-based classification, as shown in **Table 4**, mUMCNR results in the highest predictive accuracy in 8 cases. At the same time, in **Tables 5 - 6**, the performance of mUMCNR achieved with the highest predictive accuracy are 11 and 9 cases, with regard to C4.5 and PART, respectively. On the average, mUMCNR on C4.5 and PART classifiers comes with the highest classification accuracy when comparing it with all other methods. By investigating the results in **Tables 4 - 6**, we conclude that the efficiency and capability of mUMCNR can be achieved impressively with the maximal number of the highest accuracy for all classifiers when comparing it with all other methods. Meanwhile, the average dimensionality reduction is still lower than all of the methods as reported in **Table 3**.

As described above, mUMCNR uses the neighborhood size, with numeric values varying from 0.02 to 0.2 in the step of 0.02. Obviously, mUMCNR achieves a subset of features of not more than 10 subsets on each data set. At the same time, NDEM gets one subset of features, with the neighborhood size defined as 0.14 [42], while other methods yield one subset of features for each data set. Therefore, mUMCNR takes a much longer computation time to search for feature subsets than other methods. From the experiment, we can observe that the computation time the mUMCNR method uses is approximately 4 times that of NDEM. In addition, we can illustrate the relation of using the computation time between mUMCNR and other methods to search for a feature subset as mUMCNR > NDEM > ReliefF > RS > CNS > CFS.

Table 6 Classification accuracy of PART classifier.

No.	Data	unselect	RS	NDEM	CNS	CFS	ReliefF	mUMCNR
1	Wine	0.9326	0.9438	0.9163	0.9438	0.9269	0.9438	**0.9719**
2	Sonar	**0.8029**	N/A	0.7490	0.7740	0.7596	0.7211	0.7874
3	Ionos	0.9174	**0.9231**	0.9204	0.8888	0.9088	0.8803	0.9203
4	Wdbc	0.9332	0.9420	0.9561	0.9420	0.9455	0.9297	**0.9632**
5	Parkinsons	0.8154	0.8410	0.8263	0.8564	0.8153	0.8205	**0.8758**
6	Cleveland	0.7744	N/A	0.8113	0.7777	0.8013	0.8148	**0.8172**
7	Glass	0.6776	N/A	0.6820	0.7196	0.6869	0.6869	**0.7290**
8	Votes	0.9267	N/A	0.9367	0.9333	**0.9433**	0.9367	**0.9433**
9	Soybean	**0.9195**	0.7850	0.7907	0.7622	0.8795	0.8762	0.8623
10	Lymphography	**0.8176**	0.7770	0.7648	0.7635	0.7567	0.8040	0.7643
11	Promoters	0.8491	**0.9339**	0.9336	**0.9339**	0.8490	0.8773	0.7927
12	Ecoli	**0.8363**	N/A	0.8360	**0.8363**	**0.8363**	0.7619	0.8360
13	Heart	0.7481	N/A	0.7741	0.7778	0.7815	0.7926	**0.8111**
14	Hepatitis	0.8452	0.7935	**0.8521**	0.8000	0.8516	0.7935	**0.8521**
15	German	0.702	N/A	0.699	0.702	0.715	0.706	**0.729**
	Average	**0.8332**	**N/A**	**0.8299**	**0.8274**	**0.8305**	**0.8230**	**0.8437**

The advantages of the mUMCNR method are that it can produce more than one features subset with δ value in the range of 0.02 - 0.2. Moreover, each value of δ may begin searching in the feature space with a different starting feature. Therefore, these subsets are expected to increase the opportunities that lead to a near-optimal subset or a globally optimal subset. In addition, mUMCNR is suitable for numerical and mixed features. This is because mUMCNR uses the Euclidean distance for distance measuring between samples, which is more suitable for numerical than for discrete features. However, the disadvantages of mUMCNR are that it takes a much longer computation time than NDEM and other methods. Furthermore, mUMCNR may be unsuitable with discrete features, because it computes the distance by using Euclidean distance.

Generally, a learning algorithm selects the optimal subset that is suitable for it. The optimal subset often measures from the classification accuracy of the learning algorithm. The optimal subset of features varies when changing from one learning algorithm to another. However, feature selection based on mUMCNR can create candidate reducts of more than one candidate reduct with a value of δ in the range of [0.02, 0.2]. Therefore, mUMCNR has the advantage over other FS methods, because it gives an opportunity to find a candidate reduct that is appropriate with SVM, C4.5, and PART classifiers as shown in **Tables 4 - 6**.

C: Comparison of feature selection algorithms with wrapper subset evaluation

In many applications in machine learning, wrapper-based subset evaluation is necessary for pressing to select an optimal subset of features. The number of features in the optimal subset is greatly reduced with regard to a classifier. The optimal subset often measures based on the classification accuracy of the classifier. The optimal subset of features may differ in each learning algorithm. Therefore, it is difficult, or impossible, for one feature selection algorithm to choose one subset of features that is suitable for all learning algorithms. We will compare the proposed method with NDEM and wrapper subset evaluation, in order to illustrate the efficiency and effectiveness of mUMCNR, in this section.

Tables 7 - 8 show the number of selected features and the corresponding classification performance based on NDEM, mUMCNR, and wrapper, on SVM, C4.5, and PART, respectively. Among the fifteen data sets, on average, the wrapper selects 3.87, 4.33, and 4.80 on SVM, C4.5, and PART, respectively, which are the minimal number of averages in all classifiers. Meanwhile, NDEM comes with the maximal number of averages in 2 classifiers. When considering the classification accuracy of classifiers, as shown in **Table 8**, wrapper on SVM and PART comes with the highest accuracy in 9 cases and 10 cases, respectively. Meanwhile, mUMCNR on SVM and mUMCNR on PART achieve the highest accuracy in 7 cases and 4 cases, respectively. At the same time, mUMCNR on C4.5 achieves the highest accuracy in 8 cases, which is more outstanding than wrapper on C4.5, at 7 cases. On average, mUMCNR on SVM and C4.5 classifiers comes with the highest classification accuracy when compared with all other methods. However, wrapper receives an accuracy of classification on PART that is higher than that of mUMCNR and NDEM, as shown in **Table 8**.

Table 7 Number of selected features when compared to wrapper subset evaluation.

No.	Data	SVM			C4.5			Part		
		Wrapper	**NDEM**	**mUMCNR**	**Wrapper**	**NDEM**	**mUMCNR**	**Wrapper**	**NDEM**	**mUMCNR**
1	Wine	5	5	6	4	5	3	4	5	3
2	Sonar	3	7	10	3	7	4	4	7	6
3	Ionos	3	9	11	5	9	10	5	9	10
4	Wdbc	8	6	5	6	6	4	5	6	4
5	Parkinsons	2	7	4	4	7	5	6	7	4
6	Cleveland	1	8	10	2	8	11	3	8	10
7	Glass	4	5	6	4	5	8	4	5	8
8	Votes	1	8	8	1	8	8	2	8	8
9	Soybean	10	11	14	14	11	14	16	11	14
10	Lymphography	5	7	8	3	7	8	3	7	8
11	Promoters	6	4	4	1	4	4	3	4	4
12	Ecoli	6	7	7	5	7	6	6	7	6
13	Heart	1	11	6	5	11	6	4	11	6
14	Hepatitis	1	6	5	2	6	7	2	6	5
15	German	2	11	8	6	11	10	5	11	12
	Average	**3.87**	**7.47**	**7.47**	**4.33**	**7.47**	**7.20**	**4.80**	**7.47**	**7.20**

Table 8 Comparison of classification accuracy on classifiers.

No.	Data	SVM			C4.5			PART		
		Wrapper	**NDEM**	**mUMCNR**	**Wrapper**	**NDEM**	**mUMCNR**	**Wrapper**	**NDEM**	**mUMCNR**
1	Wine	0.9831	0.9775	**0.9944**	**0.9775**	0.9213	0.9719	0.9607	0.9163	**0.9719**
2	Sonar	0.7788	0.7933	**0.8125**	0.7356	0.7644	**0.7692**	0.7115	0.7490	**0.7874**
3	Ionos	0.8917	0.9402	**0.9487**	**0.9402**	0.9288	0.9316	**0.9373**	0.9204	0.9203
4	Wdbc	**0.9719**	0.9666	**0.9719**	0.9543	0.9473	**0.9684**	0.9543	0.9561	**0.9632**
5	Parkinsons	**0.8769**	0.8718	**0.8769**	**0.9333**	0.8513	0.8821	**0.8974**	0.8263	0.8758
6	Cleveland	0.7508	0.8047	**0.8215**	0.7744	**0.8182**	**0.8182**	**0.8418**	0.8113	0.8172
7	Glass	**0.6215**	0.5697	0.5939	0.6729	0.6591	**0.6955**	0.6776	0.6820	**0.7290**
8	Votes	**0.9500**	0.9400	0.9400	**0.9500**	0.9367	0.9367	**0.9467**	0.9367	0.9367
9	Soybean	**0.9429**	0.8608	0.9136	**0.9297**	0.8185	0.9034	**0.9283**	0.7907	0.8623
10	Lymphography	**0.8378**	0.7971	0.7771	**0.8108**	0.7433	0.7762	**0.8311**	0.7648	0.7643
11	Promoters	**0.9717**	0.8582	0.8773	0.8019	0.8491	**0.8773**	0.8208	**0.9336**	0.7927
12	Ecoli	**0.8452**	0.8393	0.8393	0.8304	**0.8423**	**0.8423**	**0.8363**	0.8360	0.8360
13	Heart	0.7519	0.8370	**0.8444**	0.8000	0.8111	**0.8185**	**0.8556**	0.7741	0.8111
14	Hepatitis	**0.8452**	0.8196	0.8438	0.8516	0.8454	**0.8642**	**0.8581**	0.8521	0.8521
15	German	0.7200	**0.7580**	0.7560	**0.7430**	0.7070	0.7320	**0.7430**	0.6990	0.7290
	Average	**0.8493**	**0.8423**	**0.8541**	**0.8470**	**0.8296**	**0.8525**	**0.8534**	**0.8299**	**0.8433**

As described above, mUMCNR is effective in selecting an appropriate subset of learning algorithms, and can also select a subset that contains information that is valuable to the learning algorithm. Here, we will show the efficiency and effectiveness of mUMCNR when compared to NDEM, by combining the concept of filter with wrapper together. That is, we first select the relevant features that were evaluated with mUMCNR and NDEM. We then use 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, by the order of selection. The results that are evaluated with SVM, C4.5, and PART are shown in **Tables 9** and **10**. The results also show that wrapper-based postpruning is essential for feature selection. The numbers of features in optimal subsets are greatly reduced in most of the cases. Also, each learning algorithm comes with an optimal number of features that is different. No feature selection algorithm is applicable to various learning algorithms. It is efficient to use mUMCNR to select a candidate subset that is suitable for a learning algorithm, and then use wrapper to select the optimal subset.

Table 9 Subset size after postpruning with wrapper.

No.	Data	SVM			C4.5			Part		
		Wrapper	**NDEM**	**mUMCNR**	**Wrapper**	**NDEM**	**mUMCNR**	**Wrapper**	**NDEM**	**mUMCNR**
1	Wine	5	4	5	4	3	3	4	4	3
2	Sonar	3	7	10	3	6	3	4	5	6
3	Ionos	3	9	11	5	5	4	5	4	3
4	Wdbc	8	3	5	6	3	3	5	4	3
5	Parkinsons	2	2	2	4	6	3	6	5	3
6	Cleveland	1	5	6	2	8	7	3	3	3
7	Glass	4	4	5	4	4	5	4	5	8
8	Votes	1	1	1	1	1	1	2	2	2
9	Soybean	10	11	10	14	9	13	16	11	12
10	Lymphography	5	5	4	3	4	2	3	3	2
11	Promoters	6	3	4	1	3	4	3	4	2
12	Ecoli	6	6	6	5	5	6	6	6	6
13	Heart	1	11	6	5	4	4	4	4	3
14	Hepatitis	1	6	1	2	6	6	2	6	5
15	German	2	11	2	6	6	7	5	10	10
	Average	**3.87**	**5.87**	**5.20**	**4.33**	**4.87**	**4.73**	**4.80**	**5.07**	**4.73**

Table 10 Comparison of classification accuracy on classifiers.

No.	Data	SVM			C4.5			PART		
		Wrapper	**NDEM**	**mUMCNR**	**Wrapper**	**NDEM**	**mUMCNR**	**Wrapper**	**NDEM**	**mUMCNR**
1	Wine	**0.9831**	0.9719	**0.9831**	**0.9775**	0.9157	0.9719	0.9607	0.9213	**0.9719**
2	Sonar	0.7788	0.7933	**0.8125**	0.7356	0.7644	**0.7885**	0.7115	0.7548	**0.7874**
3	Ionos	0.8917	0.9402	**0.9487**	**0.9402**	0.9316	0.9316	**0.9373**	0.9204	0.9231
4	Wdbc	**0.9719**	0.9666	**0.9719**	0.9543	0.9525	**0.9684**	0.9543	0.9561	**0.9632**
5	Parkinsons	**0.8769**	**0.8769**	**0.8769**	**0.9333**	0.8615	0.8923	0.8974	0.8513	**0.9026**
6	Cleveland	0.7508	0.8249	**0.8451**	0.7744	0.8182	**0.8215**	**0.8418**	**0.8418**	**0.8418**
7	Glass	**0.6215**	0.5981	0.6028	0.6729	0.6822	**0.7430**	0.6776	0.6820	**0.7290**
8	Votes	**0.9500**	**0.9500**	**0.9500**	**0.9500**	**0.9500**	**0.9500**	**0.9467**	**0.9467**	**0.9467**
9	Soybean	**0.9429**	0.8608	0.9268	**0.9297**	0.8243	0.9034	**0.9283**	0.7907	0.8624
10	Lymphography	**0.8378**	0.8041	0.8243	**0.8108**	0.7770	0.8041	**0.8311**	0.7905	0.8041
11	Promoters	**0.9717**	0.9057	0.8773	0.8019	0.8585	**0.8773**	0.8208	**0.9336**	0.8113
12	Ecoli	**0.8452**	**0.8452**	**0.8452**	0.8304	0.8304	**0.8423**	**0.8363**	**0.8363**	0.8360
13	Heart	0.7519	0.8370	**0.8444**	0.8000	0.8370	**0.8444**	0.8556	0.8556	**0.8593**
14	Hepatitis	**0.8452**	0.8196	**0.8452**	0.8516	0.8454	**0.8710**	**0.8581**	**0.8521**	**0.8521**
15	German	0.7200	**0.7580**	0.7560	**0.7430**	0.7350	0.7390	**0.7430**	0.7240	0.7390
	Average	**0.8493**	**0.8502**	**0.8607**	**0.8470**	**0.8389**	**0.8632**	**0.8534**	**0.8438**	**0.8553**

Tables 9 - 10 show the number of selected features and the classification accuracy based on the wrapper, NDEM+wrapper, and mUMCNR+wrapper. From the 15 data, wrapper on SVM and C4.5 achieves a minimal number of features in 10 data sets and 8 data sets, respectively. Meanwhile, MUMCNR+wrapper achieves a minimal number of features in 7 data sets and 8 data sets on SVM and C4.5, respectively. However, on the PART classifier, MUMCNR+wrapper can achieve the minimal number of features in 10 data sets, which is more outstanding than wrapper on PART and NDEM+wrapper. On average, MUMCNR+wrapper selects 5.20 features on SVM which have a value less than NDEM+wrapper. In addition, mUMCNR+wrapper selects a size of features that is smaller than NDEM+wrapper on both C4.5 and PART classifiers.

With regard to the performance of classification based on SVM, C4.5, and PART, as shown in **Table 10**, mUMCNR+SVM achieves the highest accuracy in 10 cases. At the same time, mUMCNR+C4.5 and mUMCNR+PART achieve the highest accuracy in 9 cases on 2 classifiers. In addition, on average in **Table 10**, mUMCNR+wrapper can achieve impressive results by producing accuracies that are the highest 3 values among the performance of classifiers on the e algorithms offered. By observing the results in **Tables 9 - 10**, we conclude that the efficiency and capability of mUMCNR can be achieved with the maximal number of the highest accuracy for all classifiers, when compared with all other methods. Therefore, it is an indication that information considering both the positive region and boundary region of mUMCNR can extract a subset of features which contain much more valuable information to the learning algorithm than NDEM.

Conclusions

Selecting a subset of attributes with mUMCNR, and maximizing information of the certainty region while minimizing that of the uncertainty region, leads to an impressive improvement in accuracy over various data sets when compared with the RS and NDEM approaches. Therefore, it is clear that a subset of attributes obtained from mUMCNR contains much more valuable information than those obtained using the dependency function alone, and also that using this idea for feature selection by minimizing the number of samples in the boundary region (the idea of NDEM) is beneficial.

In this paper, we proposed a feature evaluation measure strategy called mUMCNR, which can be directly applied to both discrete and continuous features. In addition, a search algorithm based on mUMCNR can be used to deal with heterogeneous features without the discretizing of numerical features. We used the neighborhood decision to define and compute a decision positive region and a decision boundary region in metric spaces. The difference of information between the decision positive region and

the decision boundary region is used to measure the goodness of feature subset. The NDEM method focuses on the number of samples in the boundary region, so that measuring the goodness of feature subset is not elaborated on. Meanwhile, the mUMCNR method considers the information of the samples that belongs to both the positive region and boundary region, and is more accurate in measuring the goodness of feature subset. For this reason, from the experiments, we can see that mUMCNR yields results of higher average classification accuracy than NDEM on SVM and C4.5.

We have presented a forward greedy strategy for searching feature subsets to minimize the information of the positive region and maximize the information contained in the boundary region. We compared the proposed method with some classical algorithms, e.g., CFS, CNS, and ReliefF. The results have shown that the proposed algorithm is effective when dealing with discrete data, numerical data, and mixed data. We have demonstrated the phenomenon of effectiveness on classification accuracy and efficiency of subset size which occurs in selecting the optimal subset of features by using the wrapper technique. Combining mUMCNR with the wrapper technique can bring improvement to classification performance that is more notable than NDEM on SVM, C4.5, and PART. Although mUMCNR does need to predefine the value of δ that is suitable for the classifier, the value of δ is specified without using domain knowledge. Furthermore, it can be seen that the idea of mUMCNR leads to an impressive improvement of the classification accuracy over various data sets when compared with the RS and NDEM approaches. To increase the efficiency and effectiveness on the selected features, the wrapper-based postpruning method is necessary for optimal subset selection. When considering the number of selected features and the corresponding classification performance on classifiers, mUMCNR is better suited for applying to wrapper-based postpruning.

References

[1] A Jain and D Zongker. Feature Selection: Evaluation, application, and small sample performance. *IEEE Trans. Pattern Anal. Mach. Intell.* 1997; **19**, 153-8.
[2] S Mitra. An evolutionary rough partitive clustering. *Pattern Recogn. Lett.* 2004; **25**, 1439-49.
[3] AL Blum and P Langley. Selection of relevant features and examples in machine learning. *Artif. Intell*. 1997; **97**, 245-71.
[4] R Kohavi and GH John. Wrappers for feature subset selection. *Artif. Intell.* 1997; **97**, 273-324.
[5] R Ruiz, JC Riquelme and JS Aguilar-Ruiz. Incremental wrapperbased gene selection from microarray data for cancer classification. *Pattern Recogn.* 2006; **39**, 2383-92.
[6] M Dash and H Liu. Feature selection for classification. *Intell. Data Anal.* 1997; **1**, 131-56.
[7] Y Kim, W Street and F Menczer. Feature selection for unsupervised learning via evolutionary search. *In*: Proceeding of the 6th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Boston, Massachusetts, USA, 2000, p. 365-9.
[8] MH Aghdam, N Ghasem-Aghaee and ME Basiri. Text feature selection using ant colony optimization. *Expert Syst. Appl.* 2009; **36**, 6843-53.
[9] W Shang, H Huang, H Zhu, Y Lin, Y Qu and Z Wang. A novel feature selection algorithm for text categorization. *Expert Syst. Appl*. 2007; **33**, 1-5.
[10] W Lee, SJ Stolfo and KW Mok. Adaptive intrusion detection: A data mining approach. *Artif. Intell. Rev*. 2000; **14**, 533-67.
[11] GJ Mun, BN Noh and YM Kim. Enhanced stochastic learning for feature selection in intrusion classification. *Int. J. Innov. Comput. Inform. Contr.* 2009; **5**, 3625-35.
[12] JC Patra, GP Lim, PK Meher and EL Ang. DNA microarray data analysis: Effective feature selection for accurate cancer classification. *In*: Proceedings of the International Joint Conference on Neural Networks, Orlando, Florida, USA, 2007, p. 260-5.
[13] E Xing, M Jordan and R Karp. Feature selection for high-dimensional genomic microarray data. *In*: Proceedings of the 18th International Conference on Machine Learning. San Francisco, USA, 2001, p. 601-8.
[14] Z Pawlak. Rough sets. *Int. J. Inform. Comput. Sci.* 1982; **11**, 314-56.

[15] Z Pawlak. *Rough Sets: Theoretical Aspects of Reasoning about Data.* Kluwer Academic Publishing, Dordrecht, 1991.
[16] TM Cover and JA Thomas. *Elements of Information Theory*. Wiley, New York, 1991.
[17] Y Chen, D Miao and R Wang. A rough set approach to feature selection based on ant colony optimization. *Pattern Recogn. Lett.* 2010; **31**, 226-33.
[18] JH Chiang and SH Ho. A combination of rough-based feature selection and RBF neural network for classification using gene expression data. *IEEE Trans. Nanobiosci*. 2008; **7**, 91-9.
[19] A Hassanien. Rough set approach for attribute reduction and rule generation: A case of patients with suspected breast cancer. *J. Am. Soc. Inform. Sci. Tech.* 2004; **55**, 954-62.
[20] AR Hedar, J Wang and M Fukushima. Tabu search for attribute reduction in rough set theory. *Soft Comput*. 2008; **12**, 909-18.
[21] YH Hung. A neural network classifier with rough set-based feature selection to classify multiclass IC package products. *Adv. Eng. Inform*. 2009; **23**, 348-57.
[22] R Jensen and Q Shen. Semantics-preserving dimensionality reduction: Rough and fuzzy-rough-based approaches. *IEEE Trans. Knowl. Data Eng*. 2004; **16**, 1457-71.
[23] Y Li, SCK Shiu, SK Pal and JNK Liu. A rough setbased case-based reasoner for text categorization. *Int. J. Approx. Reason*. 2006; **41**, 229-55.
[24] P Maji and S Paul. Rough sets for selection of molecular descriptors to predict biological activity of molecules. *IEEE Trans. Syst. Man Cybern. Syst. C: Appl. Rev*. 2010; **40**, 639-48.
[25] RW Swiniarski and A Skowron. Rough set methods in feature selection and recognition. *Pattern Recogn. Lett*. 2003; **24**, 833-49.
[26] R Battiti. Using mutual information for selecting features in supervised neural net learning. *IEEE Trans. Neural Netw.* 1994; **5**, 537-50.
[27] PA Estevez, M Tesmer, CA Perez and JM Zurada. Normalized mutual information feature selection. *IEEE Trans. Neural Netw.* 2009; **20**, 189-201.
[28] N Kwak and CH Choi. Input feature selection for classification problems. *IEEE Trans. Neural Netw*. 2002; **13**, 143-59.
[29] 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.* 2005; **27**, 1226-38.
[30] Q Hu, D Yu, J Liu and C Wu. Neighborhood rough set based heterogeneous feature subset selection. *Inform. Sci.* 2008; **178**, 3577-94.
[31] JS Deogun, VV Raghavan and H Sever. Exploiting upper approximation in the rough set methodology. *In*: Proceedings of the 1st International Conference Knowledge Discovery and Data Mining. Montréal, Québec, Canada, 1995, p. 69-74.
[32] M Inuiguchi and M Tsurumi. Measures based on upper approximations of rough sets for analysis of attribute importance and interaction. *Int. J. Innov. Comput. Inform. Contr* 2006; **2**, 1-12.
[33] D Miao, Q Duan, H Zhang and N Jiao. Rough set based hybrid algorithm for text classification. *Expert Syst. Appl.* 2009; **36**, 9168-74.
[34] S Foitong, B Attachoo and O Pinngern. Feature subset selection wrapper based on mutual information and rough sets. *Expert Syst. Appl.* 2012; **39**, 574-84.
[35] H Liu and L Yu. Toward integrating feature selection algorithms for classification and clustering. *IEEE Trans. Knowl. Data Eng.* 2005; **17**, 1-12.
[36] MA Hall. Correlation-based feature selection for discrete and numeric class machine learning. *In:* Proceedings of the 17th International Conference on Machine Learning. San Francisco, CA, USA, 2000, p. 359-66.
[37] H Liu and R Setiono. A probabilistic approach to feature selection: A filter solution. *In*: Proceedings of the 13th International Conference on Machine Learning. San Francisco, CA, USA, 1996, p. 319-27.
[38] N Parthalain, Q Shen and R Jensen. A distance measure approach to exploring the rough set boundary region for attribute reduction. *IEEE Trans. Knowl. Data Eng.* 2010; **22**, 305-17.

[39] L Yu and H Liu. Efficient feature selection via analysis of relevance and redundancy. *J. Mach. Learn. Res*. 2004; **5**, 1205-24.
[40] MR Sikonja and I Kononenko. Theoretical and empirical analysis of ReliefF and RReliefF. *Mach. Learn*. 2003; **53**, 23-69.
[41] CE Shannon and W Weaver. *The Mathematical Theory of Communication*. University of Illinois Press, Urbana, Israel, 1994.
[42] Q Hu, W Pedrycz, D Yu and J Lang. Selecting discrete and continuous features based on neighborhood decision error minimization. *IEEE Trans. Syst. Man Cybern. Syst. B* 2010; **40**, 137-50.
[43] CJ Merz and PM Murphy. *UCI Repository of Machine Learning Databases, Irvine*. Department of Information and Computer Science, University of California, 1996.