IEICE TRANS. INF. & SYST., VOL.E95–D, NO.4 APRIL 2012

PAPER Special Section on Knowledge-Based Software Engineering

Rough-Mutual Feature Selection Based on Min-Uncertainty and Max-Certainty

Sombut FOITONG[†a)], ***Student Member***, **Ouen PINNGERN**[††], ***and*** **Boonwat ATTACHOO**[†], ***Nonmembers***

SUMMARY Feature selection (FS) plays an important role in pattern recognition and machine learning. FS is applied to dimensionality reduction and its purpose is to select a subset of the original features of a data set which is rich in the most useful information. Most existing FS methods based on rough set theory focus on dependency function, which is based on lower approximation as for evaluating the goodness of a feature subset. However, by determining only information from a positive region but neglecting a boundary region, most relevant information could be invisible. This paper, the maximal lower approximation (Max-Certainty) – minimal boundary region (Min-Uncertainty) criterion, focuses on feature selection methods based on rough set and mutual information which use different values among the lower approximation information and the information contained in the boundary region. The use of this idea can result in higher predictive accuracy than those obtained using the measure based on the positive region (certainty region) alone. This demonstrates that much valuable information can be extracted by using this idea. Experimental results are illustrated for discrete, continuous, and microarray data and compared with other FS methods in terms of subset size and classification accuracy.

key words: *rough sets, mutual information, feature selection, boundary region, classification*

1. Introduction

In many fields of artificial intelligence such as machine learning, pattern recognition, text categorization, and data mining, an essential technique used in data preprocessing is Feature Selection (FS). Feature selection technique is applied to reduce the number of features, remove irrelevant, redundant, or noisy data, and bring about important effects for applications, namely the increasing speed of a learning algorithm, the improved predictive accuracy and the capability in understanding the results. Feature selection is a process which selects a subset of the original features of a data set while the most essential information of the data set should still be preserved. Feature selection has been expanded into many other fields of research. It has also been developed for decades as in the statistical pattern recognition [20], [30], machine learning [5], [23], [41], and data mining [9], [22]. At the same time, it was widely applied in a number of fields such as text classification [1], [42], intrusion detection [25], [31], and gene expression analysis [36], [48].

Over the past ten years, a large number of feature selection methods have been proposed. The widely used methods for filter-feature selection method are rough set [37], [38] and mutual information [8]. Most existing FS approaches [6], [7], [15], [16], [18], [21], [26], [28], [46] based on the rough set method take 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. Mutual information approach is widely used for features ranking [3], [11], [24], [39] 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 to remove redundant features because features are likely to have a similar ranking. Besides, 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 [37], [38] provided a new mathematic model for dealing with imprecise, uncertain, and incomplete information. The rough set approach analyzes data relying on two important concepts, namely the lower and upper approximation of a set. The theory of RSs has been applied successfully in many fields of research [6], [7], [15], [18], [21], [26], [28], [29], [35] and is currently one of the most developed techniques in intelligent data analysis. Unlike other intelligent methods, such as fuzzy set theory and statistical methods, rough sets analysis requires no human input or domain knowledge and uses only the information presented in the given data. However, in some situations, the theory of RSs may not be able to effectively analyze the data sets with noise or vagueness. Therefore, many papers have solved those problems by applying to the variable precision rough set [4], [33], [51], [52] which uses a parameter to control noise effect in data.

Finding a subset (reduct) of an information system is a key problem in RS Theory. We desire to get reducts of an information system in order to extract rule-like knowledge from an information system. 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 [6], [16], [17], [21], [28] rely on the key concept of the lower approximation or region of certainty as for evaluating the goodness of a feature subset in determining an optimal reduct

Manuscript received June 30, 2011.
Manuscript revised October 29, 2011.

[†]The authors are with the Department of Computer Engineering, Faculty of Engineering, King Mongkut's Institute of Technology Ladkrabang, Thailand.
[††]The author is with the Department of Computer Science, Faculty of Science, Ramkhamhaeng University, Thailand.
a) E-mail: s8060058@kmitl.ac.th

DOI: 10.1587/transinf.E95.D.970

such as dependency function [16], [21] and the significance of attributes [6], [28]. Although this concept has successfully been 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 dealing with high-dimensional or highly-noisy data. 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 RSs which determine the boundary region information [10], [19], 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 been successfully applied the method to solve several problems [29], [30] which consider the lower approximation and the boundary region separately.

Recently, Parthaláin et al. [35] proposed the feature selection (DMRSAR) algorithm, an extension of the rough set attribute reduction approach (RSAR) [21]. In this method, the information contained both in the lower approximation and the boundary region is used to search for the best feature subset. The DMRSAR algorithm used a distance measure to determine the proximity of objects in the boundary region and those in the lower approximation and assign a significant value to these distances. The measure of the goodness of a feature subset is the combination of equal participation of the dependency value and the significance value. Obviously, in some situations, this method can obtain a subset that is smaller than those selected by using the information gathered from the lower approximation alone. However, determining to select a feature subset still significantly relies on the lower approximation information. Therefore, both RSAR and DMRSAR approaches relying on the dependency function as an evaluation measure are unsuccessful when applied to an inconsistent data.

The central problem of RS theory is classification analysis. It uses the available information to completely perform classification of the objects that belong to a specified class. Although it is able to handle an inconsistent granular data, it is not tolerate to noise or inexact attribute values. Quite frequently, the available information allows only for partial classification. The theory of RSs can be used as a model which is a kind of classification while the classification must be fully correct or certain. Therefore, classifying by controlling the degree of uncertainty or the error of misclassification is outside the realm of RS theory. Furthermore, acceptance for some certain level of uncertainty in practice could result in improved performance of the learning algorithm. To overcome these drawbacks, Ziarko [51], [52] introduced variable precision rough set (VPRS), extended from theory of RSs, which determines some objects of the given data set as misclassified or uncertain objects. Therefore, in this paper, VPRS is used to partition feature space of a data set. In addition, the lower and upper VPRS approximations can be calculated based on of the majority inclusion relation where the degree of inclusion is obtained by allowing a predefined level of an error.

This paper presents a novel feature selection method which is based on the VPRS and mutual information. This method determines the different amount 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 were produced by the proposed method. It can also result in outperformance of the classification accuracies compared to those obtained by using the RS dependency-based approaches.

The remainder of this paper is structured as follows. Section 2 summarizes the theoretical background of VPRS and mutual information. In Sect. 3 we propose the novel approach for feature selection based on VPRS 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 of University of California, Irvine (UCI). Section 5 concludes the method proposed in this paper and points out some future research tasks.

2. Background

In this section, the basic concepts in the theories of variable precision rough set and mutual information based on rough sets are described.

2.1 Variable Precision Rough Set

Although the theory of RSs is able to handle inconsistencies in data, the values of condition or decision attributes are expected to be exact and accurate. Noisy or vague data are outside the scope of RS theory. In the application of many real data sets, the assumption of exact data is not fulfilled and some objects are misclassified or condition attribute values are corrupted. To overcome these drawbacks, Ziarko [51], [52] introduced an extension of RS theory that is an variable precision rough set. The principal idea of VPRS is to allow objects to be classified with an error smaller than a certain predefined level. Some fundamentals of VPRS are introduced in the following part.

Let $IS = (U, A)$ be an information system, where U is a finite nonempty set of N objects $\{x_1, x_2, \ldots, x_N\}$, A is a finite nonempty set of attributes. V is a value of a set of attribute values in A and f is an information function $f : UxA \longrightarrow V$.

Any subset P of attributes A, the equivalence (also called indiscernibility) relation $IND(P)$ on U:

$$IND(P) = \left\{ \left(x_i, x_j\right) \in U \times U | \forall a \in P, f_a(x_i) = f_a(x_j) \right\}. \quad (1)$$

If $(x_i, x_j) \in IND(P)$, then x_i and x_j are indiscernible with respect to P. The equivalence classes of the P-indiscernibility relation are denoted by $[x_i]_P$. Therefore, the elements in $[x_i]_P$ are indiscernible by attributes from P.

Let X and Y be the nonempty subsets of a finite universe U. The relative degree of misclassification of set X

with respect to set Y is defined as

$$c(X,Y) = 1 - \frac{|X \cap Y|}{|X|}, if\ |X| > 0 \quad (2)$$

$$= 0, if\ |X| = 0 \quad (3)$$

It is important to note that $c(X,Y) = 0$ if and only if $X \subseteq Y$.

The majority inclusion relation which is the degree of inclusion obtained by allowing an admissible classification error (β), can be defined as

$$X \subseteq_\beta Y \Longleftrightarrow c(X,Y) \leq \beta, 0 \leq \beta < 0.5 \quad (4)$$

The *P-lower* approximation and *P-upper* approximation of X can be defined as

$$\underline{P}(X)_\beta = \bigcup \left\{ [x_i]_P \,|\, [x_i]_P \subseteq_\beta X \right\}, \quad (5)$$

$$\overline{P}(X)_\beta = \bigcup \left\{ [x_i]_P \,|c([x_i]_P, X) < 1 - \beta \right\}. \quad (6)$$

Let Q be equivalence relations over U. Therefore, the definition of the positive region, the negative region and the boundary region based on VPRS is given by

$$POS_{P\beta}(Q) = \bigcup_{X \in U/Q} \underline{P}(X)_\beta, \quad (7)$$

$$NEG_{P\beta}(Q) = U - \bigcup_{X \in U/Q} \overline{P}(X)_\beta, \quad (8)$$

$$BND_{P\beta}(Q) = \bigcup_{X \in U/Q} (\overline{P}(X)_\beta - \underline{P}(X)_\beta), \quad (9)$$

and also the degree of dependency (or the quality of classification) as:

$$\gamma_{P\beta}(Q) = \frac{\left|POS_{P\beta}(Q)\right|}{|U|}. \quad (10)$$

Note that, according to the above definitions of set approximations, the lower approximation of set X can be interpreted as the collection of all the elementary sets which can be classified into X with the classification error not greater than β. The upper approximation of X includes all the elementary sets that cannot be classified into $-X$ with the error not greater than β. Finally, the boundary region of X consists of all the elementary sets that cannot be classified either into X or into $-X$ with the classification error that is not greater than β. Note also that $\underline{P}(X)_\beta = \underline{P}(X)$ for $\beta = 0$, therefore, the traditional rough set becomes a special case of VPRS.

Figure 1 shows set approximations based on VPRS of set X. The rectangular grid demonstrates the information granules of U induced by the equivalence relations $IND(P)$. In VPRS, we can see the granules which include X can be classified into the positive region depending on the value of the specified β. Therefore, we can imagine the trade-off of information granules which consist of X between the positive region and the boundary region controlled by β value.

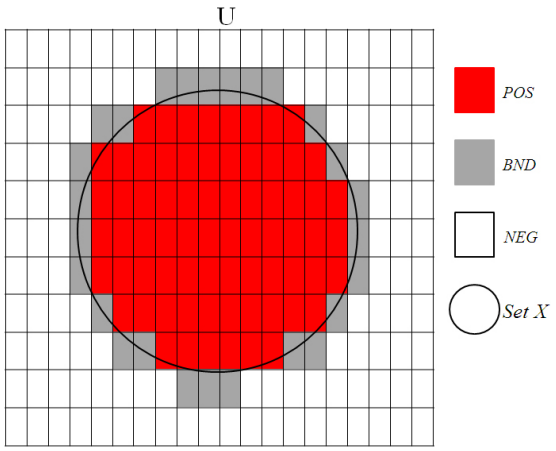

Fig. 1 Variable precision rough set in discrete feature space.

2.2 Mutual Information Based on Rough Sets

The information theory proposed by Shannon [43] provides useful tools to measure the information of a data set with entropy and mutual information. The entropy can be interpreted as an estimation of the quantity of information represented in random variables. The mutual information is a measure of generalized correlation between two random variables, and can also be interpreted as the amount of information shared by two 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 $P \subseteq A$ of features, let $U/IND(P) = \{X_1, X_2, \ldots, X_n\}$ denote the partition induced by the equivalence relation $IND(P)$. The information entropy $H(P)$ of knowledge P is defined as

$$H(P) = -\sum_{i=1}^{n} p(X_i) \log(p(X_i)), \quad (11)$$

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

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

$$H(Q|P) = -\sum_{i=1}^{n} p(X_i) \sum_{j=1}^{m} p(Y_j|X_i) \log(p(Y_j|X_i)), \quad (12)$$

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

The mutual information is a measure of the amount of information that knowledge P contains about knowledge Q which is defined as

$$I(Q;P) = \sum_{j=1}^{m} \sum_{i=1}^{n} p(Y_j, X_i) \log \frac{p(Y_j, X_i)}{p(Y_j)p(X_i)}, \quad (13)$$

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

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

$$I(P;Q) = H(Q) - H(Q|P). \tag{14}$$

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 transformations. We focus on the feature selection methods based on mutual information as a measure of relevance and redundancy of features to find the most relevant feature subset. In this paper, mutual information is used as information measure of correlation between the lower approximation $\underline{P}(X)_\beta$ and class X. Furthermore, mutual information of the boundary region $BND_{P\beta}(X)$ with respect to class X is measured. More details on information measuring of the lower approximation and the boundary region can be seen in the next section.

3. Feature Selection Based on Min-Uncertainty and Max-Certainty

3.1 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 as for evaluating the goodness of a feature subset in determining an optimal subset. Many approaches based on the theory of RSs have employed the dependency function which is based on the lower approximation as an evaluation step in the FS process. Recently, the DMRSAR approach [35] has been proposed on the RS-based FS method which uses information of both the boundary region (uncertainty region) and the positive region (certainty region) to guide a search for the best feature subset. Unfortunately, these RS-based approaches yield an empty set of reduct 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 2 shows a discrete feature space in one dimension, where the samples are divided into a set of equivalence classes $\{E1, E2, \ldots, E6\}$ based on their feature values. Samples with the same feature values are grouped into one equivalence class. The height of the rectangles in Fig. 2 denotes the number of samples of the equivalence class. We can see that all of the equivalence classes are inconsistent because their samples belong to more than one of the decision classes. Therefore, both the RSAR and DMRSAR methods based on dependency function yield an empty set for the data sets which no equivalence class is consistent at the first stage.

In this section, we will present a strategy for feature subset selection based on the idea of uncertainty information minimization and certainty information maximization.

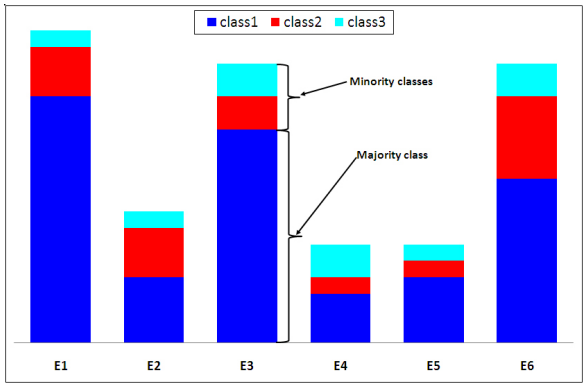

Fig. 2 A discrete feature space in one dimension.

This idea yields a nonempty set of reduct when it is applied to the data sets which all equivalence classes are inconsistent in terms of a single feature. We use both the information contained in the lower approximation and the boundary region for feature selection. In addition, mutual information is used as the information measure for both lower approximation and boundary region to guide the search for the optimal feature subset. This proposed approach selects the feature that gives the lower approximation information that is mostly relevant to class. The total information of all lower approximation is subtracted by the information contained in the boundary region with respect to classes.

3.2 Min-Uncertainty and Max-Certainty

The idea of Min-Uncertainty and Max-Certainty attempts to maximize the information of the certainty region while minimizing those of uncertainty. The evaluation of the goodness of a feature subset can be done by selecting the features that contain most different amount of information calculated by subtracting the information of the boundary region from the information of the lower approximation. This proposed criterion is a novel concept different from most existing rough-set-based FS approaches. Besides, it is contrary to the concept of the DMRSAR method [35] which uses the information gathered from both the information contained in the lower approximation and the boundary region to search for reducts.

Let D be a decision attribute while universe U can be partitioned into a collection of equivalence classes $U/IND(D) = \{D_1, D_2, \ldots, D_m\}$. Then the boundary region of $U/IND(D)$ with respect to the set of attributes P and with β value is defined as

$$BND_{P\beta}(D) = \bigcup_{D_i \in U/IND(D)} (\overline{P}(D_i)_\beta - \underline{P}(D_i)_\beta), \tag{15}$$

For a subset of features P and β value, the mutual information of the boundary region $BND_{P\beta}(D)$ with respect to knowledge D can be defined as

$$BI(P,\beta) = I(U/IND(D); BND_{P\beta}(D)/IND(P)). \tag{16}$$

The total information of mutual information between

the lower approximation $\underline{P}(D_i)_\beta$ and the equivalence class D_i with β value, denoted by $LI(P,\beta)$, can be defined as

$$LI(P,\beta) = \sum_{i=1}^{m} I(D_i; \underline{P}(D_i)_\beta). \quad (17)$$

Hence, the problem of selecting feature subset P is equivalent to the maximizing of $LI(P,\beta)$ and the minimizing of $BI(P,\beta)$, that is to maximize the objective function $E(P,\beta)$, where

$$E(P,\beta) = LI(P,\beta) - BI(P,\beta), 0 < \beta < 0.5. \quad (18)$$

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

3.3 mUMC Feature Selection Algorithm

Figure 3 shows a VPRS-based mUMCREDUCT algorithm with the idea of maximum certainty and minimum uncertainty. The proposed method is the calculation searching for a superset for all candidate reducts with the value of β varies from 0.05 to 0.45 in the step of 0.05. We can consider parameter β as the level of uncertainty or the admissible classification error of each equivalence class in a feature space. Therefore, β is used as a parameter for controlling the ratio of the samples in the minority classes (misclassified) and the majority class (classified) in each equivalence class.

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

```
mUMCREDUCT(C, D, β)
C : the set of all conditional features
D : the set of decision features
β : an admissible classification error

1)  T ← {}, R ← {}
2)  Do
3)      ∀x ∈ (C − R)
4)          if E(R ∪ {x}, β) > E(T, β)
5)              T ← R ∪ {x}
6)      R ← T
7)  Until E(R, β) = H(D) OR no more feature can be added
8)  Return R
```

Fig. 3 The mUMCREDUCT algorithm.

The mUMCREDUCT algorithm uses the maximum value of objective function E value of a subset to guide a candidate reduct selection process. If E value of the current reduct is greater than that of the previous, 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.

The proposed mUMCREDUCT algorithm works on the idea of greedy search for the feature selection process. The algorithm begins with an empty subset R. The do until loop works by calculating E value of a subset and incrementally adding a single conditional feature at a time. For each iteration, a conditional feature that has not already been evaluated will be temporarily added to subset R (line 4). If the difference amount of information of the current subset $R \cup \{x\}$ is greater the previous subset (T), then the attribute added in (line 5) is retained as part of the new subset T (line 6).

The do until loop is terminated when the difference amount of information of the current candidate reduct ($E(R,\beta)$) equals the information entropy of decision classes ($H(D)$). However, the algorithm can be terminated when the addition of all remaining feature does not improve the value of the evaluation function E.

We now analyze time complexity of mUMCREDUCT before an empirical study of its efficiency is done. As we can see from Fig. 3, major computation of the algorithm involves E values for the certainty and uncertainty region which have quadratic complexity in terms of the number of instances (M) in a data set. In terms of dimensionality N, to determine a candidate reduct, the algorithm has the best-case complexity $O(N)$ only when one feature is selected and the rest of the features are all neglected, and the worst-case complexity $O(N^2)$ when all features are selected.

4. Experimental Results and Discussion

In this section, we first test the influence of parameter β on estimation for all candidate reducts with the value of β varies from 0.05 to 0.45 in the step of 0.05. An optimal reduct of each classifier is selected from the candidate reducts with the highest predictive accuracy. Then, the results of mUMC-based feature selection will be compared to some existing techniques.

4.1 The Influence of β on mUMC-Based Feature Selection

In this section, we show the influence of parameter β on the number of selected features and an optimal subset of features for the learning algorithm on eighty data sets from the UCI Machine Learning Repository (see Table 1) [47]. We also consider the four well-known learning algorithms, namely SVM, C4.5, NB and PART, to estimate an optimal subset and classification accuracy based on a tenfold cross validation.

Table 1 Description of UCI benchmark data sets.

	Dataset	Number of features	Number of instances	Attribute types	Class
1	credit	21	1000	discrete	2
2	heart	14	294	discrete	2
3	votes	17	300	discrete	2
4	soybean	36	307	discrete	19
5	lymp	19	148	discrete	4
6	promoters	58	106	discrete	2
7	splice	61	3190	discrete	3
8	derm	35	358	discrete	6
9	dna	58	318	discrete	2
10	ionos	34	351	continuous	2
11	wine	14	178	continuous	3
12	sonar	61	208	continuous	2
13	landsat	37	2000	continuous	6
14	wdbc	31	569	continuous	2
15	parkinsons	23	195	continuous	2
16	water2	39	521	continuous	3
17	spectf	45	267	continuous	2
18	vehicle	19	846	continuous	4

Table 2 Number of selected features with the value of β.

Data	β									Average
	0.05	0.1	0.15	0.2	0.25	0.3	0.35	0.4	0.45	
credit	-	-	11	9	10	10	10	10	11	10.1
heart	7	7	8	7	7	7	7	8	8	7.3
votes	-	-	11	10	9	9	9	9	9	9.4
soybean	-	-	-	-	-	12	19	16	16	15.8
lymp	8	8	8	7	7	6	9	9	7	7.7
promoters	5	5	5	5	5	5	5	5	5	5
splice	-	-	-	-	-	-	11	11	11	11
derm	10	10	12	9	11	9	10	7	7	9.4
dna	4	5	5	4	4	4	4	4	4	4.2
ionos	6	7	6	8	7	7	8	6	6	6.8
wine	4	4	4	4	4	4	4	4	4	4
sonar	4	4	4	4	4	4	4	4	4	4
landsat	-	-	-	12	13	12	14	14	15	13.3
wdbc	8	8	8	8	6	6	6	7	7	7.1
parkinsons	5	5	5	5	5	5	5	6	6	5.2
water2	8	8	7	7	6	7	7	7	7	7.1
spectf	7	7	7	7	7	7	6	6	7	6.8
vehicle	-	-	-	-	-	-	9	8	8	8.3

To show the influence of the values of parameter β, we consider a series of numeric values varies from 0.05 to 0.45 in the step of 0.05. Table 2 shows the number of features selected with β value in the range of 0.05–0.45. It can be seen that some data yield empty subsets when calculating with some β values, e.g., $\beta = 0.05, 0.1, 0.15, 0.2, 0.25, 0.3$ for the splice and vehicle data. In this situation, the ratio of the samples in all minority classes over the whole set of the samples in the equivalence class is not less than the specified β value of each single feature. However, we can calculate to find candidate feature subsets with β value that is specified to be higher than the ratio of the samples in all minority classes over the whole sample set of the equivalence class. Therefore, the ratio of samples in the minority classes for each single feature does not exceed the specified β value.

Figure 4, 5 and 6 show changes of attribute significance with the number of selected features obtained for soybean and splice data. The significance of the feature subset is computed with the RSAR-based dependency, DMRSAR-based dependency and significance, and mUMC, respectively. The values of mUMC rapidly grow relatively high size of β before the set of features is formed by eight features. Then, the growth slows down until it completely stops when the value of features has equaled the entropy of decision classes. Meanwhile, the values of RSAR proceeded gradually increase on both data. For splice data which is inconsistent data, the value of RSAR stops early with the dependency value less than 1. However, the feature significance of DMRSAR stops very early on the soybean and splice data because it encounters the local maximum. Therefore, we can observe that the relationship mUMC $\geq$ RSAR $\geq$ DMRSAR in the phenomenon in the soybean and splice data.

The classification accuracy of candidate subsets with the values of β on both soybean and splice data is shown in Fig. 5 and 6. 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 of SVM, C4.5, NB, and PART when β = 0.4 or 0.45 for the splice data. While dealing with the soybean data, the highest accuracies of SVM, C4.5, NB, and PART with values of β are 0.4, 0.45, 0.4 or 0.45, and 0.45, respectively.

The optimal subset of features and the values of parameter with the highest predictive accuracy of each learning algorithm are shown in Table 3. Then, the highest predictive accuracy on a learning algorithm is selected to compare its performance with the RS-based attribute reduction methods and some existing classical techniques.

4.2 Comparison of Feature Selection Algorithms on UCI Benchmark Data Sets

In this section, we experiment with different algorithms of feature selection using 18 data sets, as described in Table 1, where nine data sets are the discrete features and nine data sets are the numerical features. Before applying to all feature selection techniques the numerical features are required to be discretized by the equidistance partitioning method [24] in order to segment the numerical features into several intervals and form the discretized data sets. Meanwhile, we also apply ReliefF to directly select the continuous features which are normalized into $[0, 1]$. Next, we use sequentially greedy forward search to form the best features when we compare the algorithms that evaluate features based on RS-dependency function (RSAR and DMRSAR), consistency-based subset (CNS) [27], and correlation-based feature selection (CFS) [14], respectively. In addition, the proposed method is compared with ReliefF [44] and FCFB [50] which have special searching strategies.

We first show the results of both discrete and numerical feature selection. The number of selected features of the data is given in Table 4, where the last column is the average number of features on four classifiers of the mUMC

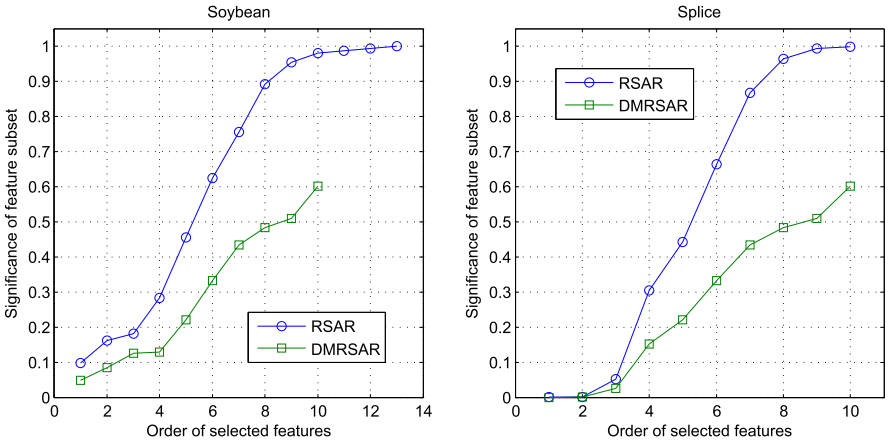

Fig. 4 Attribute significance versus the number of selected features.

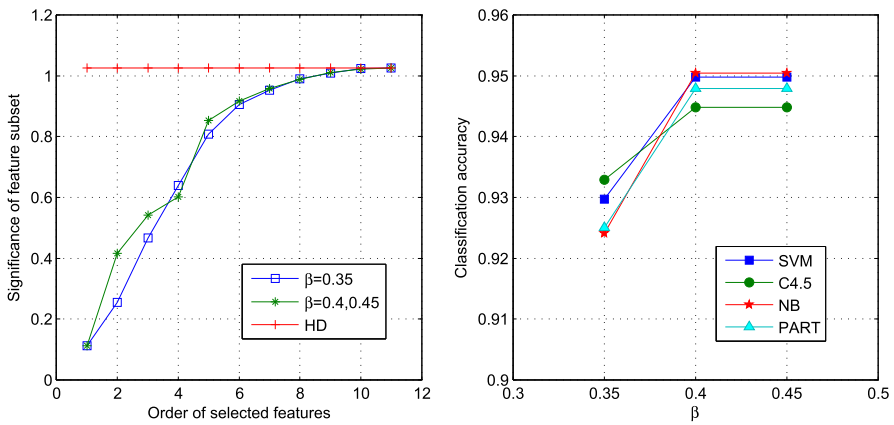

Fig. 5 Attribute significance and classification accuracy on splice data with values of β.

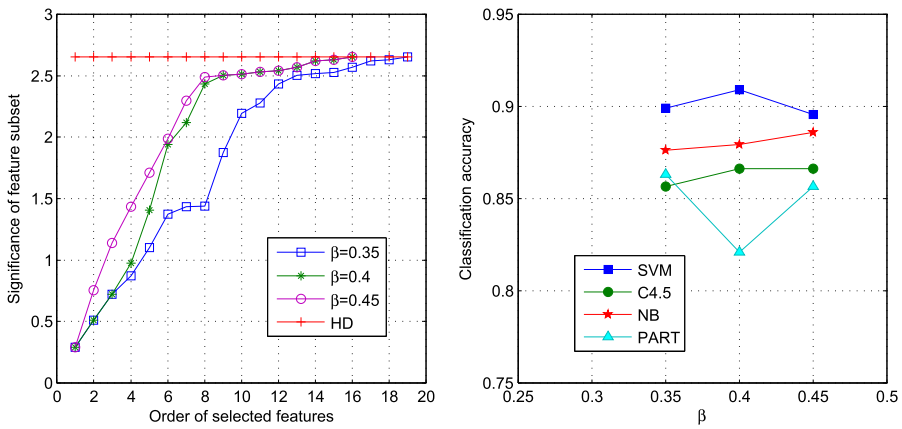

Fig. 6 Attribute significance and classification accuracy on soybean data with values of β.

method. In Tables 4–8, we observe that most of the features in raw data have been deleted by all the feature selection algorithms. At the same time, we then apply SVM, C4.5, NB and PART classifiers on each of the newly obtained data sets (with only selected features), and obtained the average accuracy of 10-fold cross validation. The results show that these algorithms are effective in retaining the classification ability. The RSAR and DMRSAR algorithms yield an empty set when it is applied to the "votes" and "credit" data, 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 to find the subset of features. We can also find that the subset contains different features

Table 3 The optimal subset and the β values for classifier.

Data	SVM		C4.5		NB		PART	
	Size	β	Size	β	Size	β	Size	β
credit	11	0.15	11	0.15	11	0.45	10	0.35–0.4
heart	8	0.4–0.45	7	0.25–0.35	8	0.4–0.45	7	0.2
votes	9	0.25–0.45	10	0.2	9	0.25–0.45	9	0.25–0.45
soybean	16	0.4	16	0.4–0.45	16	0.45	19	0.35
lymp	7	0.25	9	0.35–0.4	9	0.35–0.4	6	0.3
promoters	5	0.2,0.45	5	0.15	5	0.2,0.45	5	0.2,0.45
splice	11	0.4–0.45	11	0.4–0.45	11	0.4–0.45	11	0.4–0.45
derm	11	0.25	11	0.25	11	0.25	11	0.25
dna	4	0.05	4	0.2	4	0.05	5	0.15
ionos	7	0.25	7	0.1	6	0.15	8	0.2
wine	4	0.05–0.15	4	0.2–0.3	4	0.35–0.45	4	0.2–0.45
sonar	4	0.15	4	0.15	4	0.05	4	0.15
landsat	14	0.35	15	0.45	14	0.35	13	0.25
wdbc	6	0.25–0.3	8	0.1	6	0.25–0.3	8	0.1
parkinsons	5	0.15–0.2	5	0.05–0.1	5	0.05–0.1	5	0.05–0.1
water2	7	0.3	7	0.3	8	0.1	7	0.3
spectf	6	0.35–0.4	7	0.2	7	0.45	6	0.35
vehicle	9	0.35	9	0.35	9	0.35	9	0.35

Table 4 Number of selected features with different techniques.

Data	Feature Selection Algorithm						
	RSAR	DMRSAR	CFS	CNS	FCFB	ReliefF	mUMC
credit	-	-	3	10	3	11	10.7
heart	7	6	6	8	5	4	7.5
votes	-	-	3	8	3	6	9.2
soybean	13	4	21	11	15	15	16.7
lymp	6	1	9	7	8	7	7.7
promoters	4	4	6	4	6	4	5
splice	10	10	22	10	22	11	11
derm	7	2	20	6	15	15	11
dna	4	4	4	4	2	6	4.2
ionos	5	5	5	5	3	8	7
wine	4	4	7	4	7	4	4
sonar	4	4	12	4	1	14	4
landsat	11	12	22	13	3	13	14
wdbc	7	6	8	6	5	10	7
parkinsons	4	4	7	5	1	7	5
water2	7	2	7	6	3	9	7.2
spectf	7	6	8	7	1	6	6.5
vehicle	8	9	5	7	4	6	9
Average	**-**	**-**	**9.72**	**6.99**	**5.94**	**8.66**	**8.16**

when applying different algorithms.

Noisy data has great influence on the results that were produced by the RSAR and DMRSAR algorithms. ReliefF introduces a number of the nearest neighbors to control the noise effect. mUMC is created 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 of mUMC. However, ReliefF does need to be predefined the number of the nearest neighbors and the number of the sampling instances which both require domain knowledge. Moreover, ReliefF is feature selection based on the feature ranking method and it needs to be chosen for the optimal number of features.

Among the 18 data and seven algorithms of feature selection, CFS comes with the maximal number of features with nine data sets; meanwhile, ReliefF obtains the maximal number of features as seven data sets. On the average, CFS and ReliefF select 9.72 and 8.66 features for dimensionality reduction which are the highest two average values among the size of features that seven algorithms are applied. Although FCFB obtains with the highest feature reduction on eleven data sets, it produces the number of the highest accuracy that is less than that of mUMC with regard to all classifiers. In many applications on machine learning the wrapper-based postpruning is necessary for pressing a subset of features. The number of features in the optimal subset is greatly reduced with regard to a classifier. The optimal number of features varies from one learning algorithm to another. It is efficient to use a filter in selecting a candidate subset then wrapper is used to select an optimal subset. However, a few of features (e.g., one or two features) in the candidate subset is difficult or impossible to improve classification accuracy with wrapper-based reduction because the candidate subset contains insufficient information for the classifier.

Table 5 Classification accuracy of SVM classifier (In Percent).

Data	RSAR	DMRSAR	CFS	CNS	FCFB	ReliefF	mUMC
credit	-	-	71.20	74.20	71.20	75.70	**75.90**
heart	77.21	77.50	79.93	80.61	79.93	**82.31**	80.61
votes	-	-	94.33	93.33	94.33	93.66	**95.33**
soybean	83.06	85.34	**92.18**	84.36	90.55	88.27	90.90
lymp	81.75	81.75	82.43	79.72	81.08	**86.48**	83.10
promoters	85.84	85.84	91.50	85.84	91.50	94.02	**94.33**
splice	81.66	83.00	**95.95**	94.26	**95.95**	94.54	94.98
derm	81.00	55.02	**97.76**	75.69	95.81	94.41	**97.76**
dna	93.71	93.71	**96.22**	93.71	89.62	95.59	95.28
ionos	83.19	83.13	86.89	81.48	86.46	84.61	**87.03**
wine	93.82	93.82	**97.75**	93.25	**97.75**	95.50	95.50
sonar	76.44	76.44	76.44	72.11	74.03	78.80	**78.88**
landsat	83.70	83.70	84.40	83.50	82.60	81.50	**83.75**
wdbc	95.43	94.90	96.30	96.48	95.43	96.48	**97.01**
parkinsons	86.15	86.15	85.12	86.66	75.38	84.61	**88.20**
water2	79.65	75.81	80.80	78.31	78.50	**82.14**	80.99
spectf	**79.40**	**79.40**	**79.40**	**79.40**	**79.40**	**79.40**	**79.40**
vehicle	68.20	60.75	48.93	56.97	49.29	60.63	**69.14**

With regard to the performance of SVM-based classification, as shown in Table 5, mUMC results in the highest predictive accuracy in eleven cases. At the same time, in Tables 6–8, the performance of mUMC achieves with the highest predictive accuracy are twelve, eleven, and thirteen cases with regard to C4.5, NB, and PART, respectively. By scanning the results in Tables 5–8, we conclude that the efficiency and capability of mUMC can achieve impressively with the maximal number of the highest accuracy for all classifiers when comparing with all other methods. Meanwhile, the average of dimensionality reduction is still higher than both CFS and ReliefF as reported in Table 5.

4.3 Performance Comparison on Gene Expression Data Sets

One important application of gene expression data is the classification of samples into different categories, such as types of tumor. Gene expression data are characterized by many variables on only a few observations. In most gene expression data, the number of training samples is very small

Table 6 Classification accuracy of C4.5 classifier (In Percent).

Data	RSAR	DMRSAR	CFS	CNS	FCFB	ReliefF	mUMC
credit	-	-	71.20	72.20	71.20	**73.60**	73.00
heart	**81.97**	**81.97**	78.91	79.25	80.27	80.61	**81.97**
votes	-	-	**94.00**	93.66	**94.00**	93.66	**94.00**
soybean	81.75	83.38	82.08	80.45	83.38	85.99	**86.64**
lymp	73.64	76.35	**78.37**	74.32	75.67	76.35	77.70
promoters	84.90	84.90	83.01	84.90	83.01	83.96	**86.79**
splice	81.91	82.35	**94.48**	93.82	**94.48**	93.98	**94.48**
derm	69.83	55.30	93.01	72.06	92.73	87.15	**94.69**
dna	83.64	83.64	84.59	83.64	**85.84**	84.90	84.27
ionos	89.74	89.45	90.88	90.31	89.45	**92.59**	91.45
wine	89.88	89.88	94.38	**96.62**	94.38	95.50	**96.62**
sonar	70.67	70.67	71.63	75.00	71.63	76.92	**83.65**
landsat	**85.75**	**85.75**	83.25	81.30	82.45	81.70	84.05
wdbc	95.25	92.72	93.32	94.20	**97.01**	94.02	96.48
parkinsons	87.69	87.69	83.07	88.20	87.69	86.15	**89.74**
water2	81.95	75.43	82.34	80.23	81.38	81.95	**85.60**
spectf	80.14	79.02	76.77	79.40	79.40	79.40	**82.77**
vehicle	72.34	67.73	69.26	65.36	70.21	67.25	**72.63**

Table 7 Classification accuracy of NB classifier (In Percent).

data	RSAR	DMRSAR	CFS	CNS	FCFB	ReliefF	mUMC
credit	-	-	74.40	75.80	74.40	75.30	**76.00**
heart	79.32	79.59	**84.69**	83.67	80.61	84.01	83.67
votes	-	-	**95.66**	93.00	**95.66**	93.66	90.66
soybean	82.41	82.41	87.62	81.10	85.99	83.38	**88.59**
lymp	80.40	83.10	83.10	83.10	82.43	84.45	**84.88**
promoters	93.39	93.39	**95.28**	93.39	**95.28**	93.39	93.39
splice	82.47	83.51	**96.14**	94.32	**96.14**	95.07	95.04
derm	82.12	55.86	**97.76**	76.25	96.08	95.81	96.64
dna	93.39	93.39	**94.33**	93.39	87.73	93.71	93.39
ionos	87.17	86.32	90.31	87.74	90.31	84.61	**90.45**
wine	93.25	93.25	96.62	93.82	96.62	95.50	**97.19**
sonar	**75.48**	**75.48**	66.82	73.07	73.55	71.15	**75.48**
landsat	78.40	78.40	79.65	78.75	78.90	75.30	**79.70**
wdbc	92.97	90.15	94.20	94.55	94.02	94.37	**95.60**
parkinsons	82.56	82.56	80.51	73.33	**83.58**	75.89	**83.58**
water2	80.99	74.66	**84.64**	81.76	83.87	**84.64**	**84.64**
spectf	68.91	70.41	71.53	78.27	**79.02**	64.41	77.53
vehicle	46.21	44.08	43.73	43.02	49.29	38.06	**51.18**

(often less than a hundred) compared to a large number of genes (thousands or tens of thousands of genes) involved in the experiments. However, among a large amount of genes, only a small fraction is effective for performing a certain task. Therefore, selecting a small number of discriminative genes from thousands of genes is essential for successful sample classification.

In this paper, we select three microarray data sets which are frequently used in the studies: Colon cancer [2], Leukemia [12], and Lung cancer [13]. The details of these data sets are summarized in Table 9. For each data set, we first apply all the above feature selection algorithms in comparison and the selected genes for each algorithm. We then apply classifiers on each of the newly obtained data sets (with only the selected genes), and obtain overall classification accuracy by leave-one-out cross-validation, a performance validation procedure due to a small sample size of the microarray data.

Table 10 shows the number of genes selected for mUMC with β values in the range of 0.05–0.45. All microarray data in Table 10 can select a gene subset with any value of β in the range of 0.05–0.45. The subset of genes with β values that obtain the highest predictive accuracy of each classifier is selected to compare the performance with other techniques. Table 11 records the number of genes selected by each feature selection algorithm. We can see that all these algorithms achieve significant reduction of dimensionality by selecting only a small portion of the original genes.

The effectiveness of these seven algorithms based on the number of genes selected and the leave-one-out cross-validation accuracy are reported in Table 12. For Colon data, the classification accuracy obtained with the mUMC approach produces more numbers of the highest accuracy than all other methods. Meanwhile, CFS, CNS, and Reli-

Table 8 Classification accuracy of PART classifier (In Percent).

data	RSAR	DMRSAR	CFS	CNS	FCFB	ReliefF	mUMC
credit	-	-	71.3	**74.1**	71.3	72.6	**74.1**
heart	80.61	81.97	79.93	80.95	79.59	79.59	**82.99**
votes	-	-	94.33	93.33	94.33	94.06	**94.78**
soybean	78.5	78.5	**87.947**	76.22	83.38	86.97	86.31
lymp	77.7	77.02	75.67	76.35	76.35	**81.75**	77.7
promoters	**93.39**	**93.39**	84.9	**93.39**	84.9	86.79	**93.39**
splice	81.84	81.92	93.38	93.13	93.38	92.72	**94.79**
derm	73.46	55.58	95.53	63.4	94.97	90.22	**96.08**
dna	85.84	85.84	87.1	85.84	**88.05**	87.1	**88.05**
ionos	86.6	87.74	90.88	88.88	88.6	**93.16**	90.31
wine	90.44	90.44	92.69	**94.38**	92.69	**94.38**	**94.38**
sonar	75	75	75.96	77.4	71.63	77.88	**80.76**
landsat	**84.2**	**84.2**	83.25	83.55	81.85	80.1	83.6
wdbc	94.2	93.84	94.55	94.2	95.07	94.55	**95.78**
parkinsons	86.66	86.66	81.53	85.64	**87.69**	86.15	86.15
water2	81.38	75.23	82.14	78.88	80.42	80.23	**85.22**
spectf	78.27	79.77	77.15	79.40	79.4	79.77	**82.64**
vehicle	70.68	65.48	66.43	66.31	66.43	68.32	**71.8**

Table 9 Summary of microarray data sets.

Dataset	Number of genes	Number of samples	Number of sample per class	
			tumor	normal
Colon Tumor	2000	62	40	22
			ALL	AML
Leukemia	7129	72	47	25
			MPM	ADCA
Lung Cancer	12533	181	31	150

Table 10 Number of selected genes with the value of β.

Data	β									Average
	0.05	0.1	0.15	0.2	0.25	0.3	0.35	0.4	0.45	
Colon Tumor	3	3	3	3	3	3	3	3	3	3
Leukemia	3	3	3	3	3	3	3	3	3	3
Lung Cancer	3	3	3	3	3	3	3	3	3	3

Table 11 Number of selected genes with different techniques.

data	Feature Selection Method						
	RSAR	DMRSAR	CFS	CNS	ReliefF	FCFB	mUMC
Colon Tumor	3	3	8	3	5	2	3
Leukemia	2	2	10	3	5	6	3
Lung Cancer	3	3	N/A	2	3	12	3

Table 12 Performance of classifiers on selected genes.

data	classifiers	RSAR	DMRSAR	CFS	CNS	FCFB	ReliefF	mUMC
Colon Tumor	SVM	72.58	72.58	**83.87**	79.03	80.64	79.03	82.25
	C4.5	77.41	77.41	75.80	82.25	85.48	85.48	**87.09**
	NB	72.58	72.58	83.87	**85.48**	83.87	**85.48**	**85.48**
	PART	77.41	77.41	82.25	82.25	85.48	85.48	**88.70**
Leukemia	SVM	77.78	77.78	**94.44**	84.72	88.88	91.66	**94.44**
	C4.5	93.05	93.05	84.72	88.88	95.83	93.05	**97.22**
	NB	91.66	91.66	**98.61**	91.66	95.83	94.44	**98.61**
	PART	93.05	93.05	84.72	84.72	95.83	93.05	**97.22**
Lung Cancer	SVM	96.13	96.13	-	88.95	**98.89**	98.34	96.13
	C4.5	97.23	97.23	-	96.13	96.13	**98.34**	**98.34**
	NB	97.79	97.79	-	97.79	**99.44**	97.79	98.34
	PART	96.13	96.13	-	96.68	96.13	95.58	**97.79**

efF come with the highest accuracy in one case on SVM, NB, and NB, respectively. However, both CFS and ReliefF come with the subset of genes that is larger than mUMC.

It is notable for Leukemia data that mUMC achieves with the highest accuracy in all classifiers. At the same time, CFS obtains with the highest accuracy in two cases. However, the mUMC approach can discover the gene subset with size smaller than the CFS methods. For Lung Cancer data, both FCFB and mUMC achieve with the highest accuracy in two cases whereas mUMC obtains a subset of genes smaller than FCFB. Meanwhile, ReliefF achieves in one case only on C4.5. However, CFS fails on the Lung Cancer data as the program ran out of memory after a period of time due to its $O(N^2)$ space complexity in terms of the number of gene N.

It can be seen from the results reported in Table 12 that the efficiency of mUMC outperforms RSAR and DMRSAR in almost all classifiers for any gene expression data. In addition, the paired two-tailed t-test is used to evaluate the statistical significance (at 0.1 level) of the difference between two average accuracy values: one resulted from mUMC and the other resulted from one of RSAR and DMRSAR. The results obtained from mUMC are statistically better than those obtained from both RSAR and DMRSAR in all classifiers for Colon Tumor data. Meanwhile, on Leukemia data the results of mUMC are statistically better with regard to SVM and NB. It is important to demonstrate that mUMC does tolerant to noise in data and performs with noise better than both the RSAR and DMRSAR approaches. Furthermore, mUMC demonstrates that the subset of genes selected is much valuable information than those extracted from the gene expression data set by considering the information contained both in the certainty and uncertainty region.

5. Conclusions

The comparison of mUMC with the RS dependency-based methods has shown that the mUMC method is a good starting point for further work based on information measure both in the lower approximation (certainty region) and the boundary region (uncertainty region) for exploring the variable precision rough set. The classification accuracy results have shown to be very well to those of the RS dependency-based approaches which are outstanding for noisy data and data in which a granular of each attribute is inconsistent. When applied to high-dimensional data such as microarray data, mUMC proved the efficiency of classification accuracy in most cases compared with other algorithms. Selecting a subset of attributes with mUMC, maximizing information of the certainty region while minimizing that of the uncertainty region at the same time, leads to an impressive improvement of the accuracy over various data sets when compared with the RS dependency-based approaches. Therefore, it is clear that a subset of attributes obtained from mUMC contains much valuable information than those obtained using the dependency function alone and also using the information gathered from both lower approximation dependency value and distance metric which is a distance of the objects in the boundary region to those in the lower approximation (the idea of DMRSAR).

In this paper, we have used mutual information to evaluate the goodness of a subset on the training data partitioned by using VPRS. The difference amount of information between the information contained in the certainty region and uncertainty region is used to guide the search for the best feature subset. The value of parameter β varying from 0.05 to 0.45 is the parameter used to control the noise effect. In addition, an outlier in data is easy to handle with parameter β. Therefore, noisy information has little influence on the subset of features that were produced by mUMC algorithm. However, both noisy sample and outliers have great influence on the RS dependency-based FS methods. Those approaches based on the RS dependency are unsuccessful when applied to the inconsistent data, e.g., splice data, whereas the results obtained by mUMC have shown to be very well. The proposed method, differing form the existing RS-based FS, presents a novel feature selection method which leads to the power of extracting much valuable subset in data. In addition, parameter β is used to control the ratio of samples in all minority classes over the whole sample set of the equivalence class in the feature spaces.

We have presented a forward greedy strategy for searching feature subsets to minimize the information of minority classes and maximize the information contained in the majority class. We compared the proposed method with some classical algorithms, e.g., CFS, CNS, FCFB, and ReliefF. The results show that the proposed algorithm is effective when dealing with discrete data and numerical data. We have shown the phenomenon of effectiveness on classification accuracy and efficiency of subset size which occur in gene selection on gene expression data. Although mUMC does need to be predefined the value of β that is suitable with the classifier, the value of β is specified without using domain knowledge. Furthermore, it can be seen that the idea of mUMC leads to an impressive improvement of the classification accuracy over various data sets when compared with the RS dependency-based FS approaches. To increase the efficiency and effectiveness on the selected features, the wrapper-based postpruning method is necessary for an optimal subset selection. When considering the number of selected features and the corresponding classification

performance on classifiers, mUMC suits better for applying to wrapper-based postpruning.

It is more difficult or impossible to apply an exhaustive search due to time complexity when dealing with huge feature data such as microarray data. Greedy search guided by some heuristics is more feasible because a suboptimal feature subset could be archived with reasonable computational costs. However, the mUMCREDUCT algorithm that works on the idea of greedy search can produce more than one suboptimal subset with β value in the range of 0.05–0.45. Moreover, each value of β may begin searching in feature space with a different starting feature. Therefore, these suboptimal subsets are expected to increase the opportunities that lead to near-optimal subset or a globally optimal subset. In addition, from our experiments on all data sets we discover that the feature evaluation function E is monotonic. To avoid a suboptimal subset of greedy search and an optimal subset with running exhaustive search, other search techniques can be used with the feature evaluation function E (such as branch and bound (B&B) [32], [45], floating search [40], or GAs [34], [49] etc.).

References

[1] M.H. Aghdam, N. Ghasem-Aghaee, and M.E. Basiri, "Text feature selection using ant colony optimization," Expert Systems with Applications, vol.36, pp.6843–6853, 2009.
[2] U. Alon, N. Barkai, D. Notterman, K. Gish, S. Ybarra, D. Mack, and A. Levine, "Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays," Proc. Natl. Acad. Sci. USA 96, pp.6745–6750, 1999.
[3] R. Battiti, "Using mutual information for selecting features in supervised neural net learning," IEEE Trans. Neural Netw., vol.5, no.4, pp.537–550, 1994.
[4] M. Beynon, "Reducts within the variable precision rough sets model: A further investigation," Eur. J. Oper. Res., vol.134, no.3, pp.592–605, 2001.
[5] A.L. Blum and P. Langley, "Selection of relevant features and examples in machine learning," Artif. Intell., vol.97, pp.245–271, 1997.
[6] Y. Chen, D. Miao, and R. Wang, "A rough set approach to feature selection based on ant colony optimization," Pattern Recognit. Lett., vol.31, no.3, pp.226–233, 2010.
[7] J.H. Chiang and S.H. Ho, "A combination of rough-based feature selection and rbf neural network for classification using gene expression data," IEEE Trans. Nanobioscience, vol.7, no.1, pp.91–99, 2008.
[8] T.M. Cover and J.A. Thomas, Elements of information theory, Wiley, New York, 1991.
[9] M. Dash and H. Liu, "Feature selection for classification," Intelligent Data Analysis: An Int'l J., vol.1, no.3, pp.131–156, 1997.
[10] J.S. Deogun, V.V. Raghavan, and H. Sever, "Exploiting upper approximation in the rough set methodology," Proc. First Int'l Conf. Knowledge Discovery and Data Mining, pp.1–10, 1995.
[11] P.A. Estevez, M. Tesmer, C.A. Perez, and J.M. Zurada, "Normalized mutual information feature selection," IEEE Trans. Neural Netw., vol.20, no.2, pp.1045–9227, 2009.
[12] T. Golub et al., "Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring," Science, vol.286, pp.531–537, 1999.
[13] G.J. Gordon et al., "Translation of microarray data into clinically relevant cancer diagnostic tests using gene expression ratios in lung cancer and mesothelioma," Cancer Res., vol.62, no.17, pp.4963–4967, 2002.
[14] M.A. Hall, "Correlation-based feature selection for discrete and numeric class machine learning," Proc. 17th Int. Conf. Mach. Learn., pp.359–366, 2000.
[15] A. Hassanien, "Rough set approach for attribute reduction and rule generation: A case of patients with suspected breast cancer," J. Am. Soc. Information Science and Technology, vol.55, no.11, pp.954–962, 2004.
[16] A. Hedar, J. Wang, and M. Fukushima, "Tabu search for attribute reduction in rough set theory," Technical Report 2006-008, Dept. of Applied Mathematics and Physics, Kyoto Univ., 2006.
[17] Q. Hu, D. Yu, J. Liu, and C. Wu, "Neighborhood rough set based heterogeneous feature subset selection," Information Sciences, vol.178, no.15, pp.3577–3594, 2008.
[18] Y.H. Hung, "A neural network classifier with rough set-based feature selection to classify multiclass IC package products," Advanced Engineering Informatics, vol.23 pp.348–357, 2009.
[19] M. Inuiguchi and M. Tsurumi, "Measures based on upper approximations of rough sets for analysis of attribute importance and interaction," Int'l J. Innovative Computing, Information and Control, vol.2, no.1, pp.1–12, 2006.
[20] A. Jain and D. Zongker, "Feature Selection: evaluation, application, and small sample performance," IEEE Trans. Pattern Anal. Mach. Intell., vol.19, no.2, pp.153–158, 1997.
[21] R. Jensen and Q. Shen, "Semantics-preserving dimensionality reduction: rough and fuzzy-rough-based approaches," IEEE Trans. Knowledge Data Eng, vol.16, no.12, pp.1457–1471, 2004.
[22] Y. Kim, W. Street, and F. Menczer, "Feature selection for unsupervised learning via evolutionary search," Proc. Sixth ACM SIGKDD: Int'l Conf. Knowledge Discovery and Data Mining, pp.365–369, 2000.
[23] R. Kohavi and G.H. John, "Wrappers for feature subset selection," Artif. Intell., vol.97, nos.1–2, pp.273–324, 1997.
[24] N. Kwak and C.H. Choi, "Input feature selection for classification problems," IEEE Trans. Neural Netw., vol.13 no.1, pp.143–159, 2002.
[25] W. Lee, S.J. Stolfo, and K.W. Mok, "Adaptive intrusion detection: a data mining approach," AI Rev., vol.14, no.6, pp.533–567, 2000.
[26] Y. Li, S.C.K. Shiu, S.K. Pal, and J.N.K. Liu, "A rough set-based case-based reasoner for text categorization," International Journal of Approximate Reasoning, vol.41, no.2, pp.229–255, 2006.
[27] H. Liu and R. Setiono, "A probabilistic approach to feature selection: a filter solution," Proc. 13th Int'l Conf. Machine Learning, pp.319–327, 1996.
[28] P. Maji and S. Paul, "Rough sets for selection of molecular descriptors to predict biological activity of molecules," IEEE Trans. Syst. Man, Cybern. C: Appl. Rev., vol.40, no.6, pp.639–648, 2010.
[29] D. Miao, Q. Duan, H. Zhang, and N. Jiao, "Rough set based hybrid algorithm for text classification," Expert Systems with Applications, vol.36, pp.9168–9174, 2009.
[30] S. Mitra, "An evolutionary rough partitive clustering," Pattern Recognit. Lett., vol.25, no.12, pp.1439–1449, 2004.
[31] G.J. Mun, B.N. Noh, and Y.M. Kim, "Enhanced stochastic learning for feature selection in intrusion classification," Int'l J. Innovative Computing, Information and Control, vol.5, no.11(A), pp.3625–3635, 2009.
[32] P.M. Narendra and K. Fukunaga, "A branch-and-bound algorithm for feature subset selection," IEEE Trans. Comput., vol.C–26, no.9, pp.917–922, 1977.
[33] M. Ningler, G. Stockmanns, G. Schneider, H.D. Kochs, and E. Kochs, "Adapted variable precision rough set approach for EEG analysis," Artif. Intell. Medicine, vol.47, no.3, pp.239–261, 2009.
[34] I.-S. Oh, J.-S. Lee, and B.-R. Moon, "Hybrid genetic algorithms for feature selection," IEEE Trans. Pattern Anal. Mach. Intell., vol.26, no.11, pp.1424–1437, 2004.
[35] N.M. Parthaláin, R. Jensen, and Q. Shen, "A distance measure approach to exploring the rough set boundary region for attribute re-

duction," IEEE Trans. Knowl. Data Eng., vol.22, no.3, pp.305–317, 2010.
[36] J.C. Patra, G.P. Lim, P.K. Meher, and E.L. Ang, "DNA microarray data analysis: Effective feature selection for accurate cancer classification," Proc. International Joint Conference on Neural Networks, Orlando, Florida, USA, Aug. 2007.
[37] Z. Pawlak, "Rough sets," Int. J. Inf. Comput. Sci., vol.11, pp.314–356, 1982.
[38] Z. Pawlak, Rough sets: Theoretical aspects of reasoning about data, Kluwer Academic Publishing, Dordrecht, 1991.
[39] H. Peng, F. Long, and C. Ding, "Feature selection based on mutual information: Criteria of max-dependency, max-relevance and min-redundancy," IEEE Trans. Pattern Anal. Mach. Intell., vol.27, no.8, pp.1226–1238, 2005.
[40] P. Pudil, J. Novovičová, and J. Kittler, "Floating search methods in feature selection," Pattern Recognit. Lett., vol.15, no.11, pp.1119–1125, 1994.
[41] R. Ruiz, J.C. Riquelme, and J.S. Aguilar-Ruiz, "Incremental wrapper- based gene selection from microarray data for cancer classification," Pattern Recognit., vol.39, no.12, pp.2383–2392, 2006.
[42] W. Shang, H. Huang, H. Zhu, Y. Lin, Y. Qu, and Z. Wang, "A novel feature selection algorithm for text categorization," Expert Systems with Applications, vol.33, pp.1–5, 2007.
[43] C.E. Shannon and W. Weaver, The mathematical theory of communication, University of Illinois Press, Urbana, Israel, 1949.
[44] M.R. Sikonja and I. Kononenko, "Theoretical and empirical analysis of ReliefF and RReliefF," Mach. Learn., vol.53, no.1/2, pp.23–69, Oct./Nov. 2003.
[45] P. Somol, P. Pudil, and J. Kittler, "Fast branch-and-bound algorithms for optimal feature selection," IEEE Trans. Pattern Anal. Mach. Intell., vol.26, no.7, pp.900–912, 2004.
[46] R.W. Swiniarski and A. Skowron, "Rough set methods in feature selection and recognition," Pattern Recognit. Lett., vol.24, no.6, pp.833–849, 2003.
[47] D.J. Newman, S. Hettich, C.L. Blake, and C.J. Merz, UCI repository of machine learning databases, dept. of information and computer science, Univ. of California, http://www.ics.uci.edu/mlearn/MLRepository.html, 1998.
[48] E. Xing, M. Jordan, and R. Karp, "Feature Selection for High-Dimensional Genomic Microarray Data," Proc. 15th Intl Conf. Machine Learning, pp.601–608, 2001.
[49] J. Yang and H. Vasant, "Feature subset selection using a genetic algorithm," IEEE Intell. Syst., vol.13, no.2, pp.44–49, 1998.
[50] L. Yu and H. Liu, "Efficient feature selection via analysis of relevance and redundancy," J. Mach. Learn. Res., vol.5, pp.1205–1224, 2004.
[51] W. Ziarko, "Variable precision rough set model," J. Comput. Syst. Sci., vol.46, no.1, pp.44–54, 1993.
[52] W. Ziarko, "Probabilistic approach to rough sets," International Journal of Approximate Reasoning, vol.49, no.2, pp.272–284, 2008.

Sombut Foitong received the B.Sc. degree in mathematics-computer from Ramkhamhaeng University, Thailand, in 1997 and the M.Eng. degree in computer engineering from King Mongkuts Institute of Technology Ladkrabang, Bangkok, Thailand, in 2004, where he is currently working towards the Ph.D. degree in electrical engineering. His research interests include data mining and knowledge discovery with fuzzy and rough techniques, neural networks, pattern recognition, and machine learning.

Ouen Pinngern received the M.S. degree from Oregon State University, Corvallis, OR, in 1983 and the Ph.D. degree from University of Nebraska at Lincoln, Lincoln, in 1986, both in computer science. He is an Associated Professor in the Department of Computer Science, Faculty of Science, Ramkhamhaeng University, Bangkok, Thailand. His current research interests lie in the areas of soft computing, machine learning, data mining, information retrieval, and model of computation.

Boonwat Attachoo received the M.Eng. degree from King Mongkut's Institute of Technology Ladkrabang, in 1982 and the D.Eng. degree from Tokai University, Japan, in 1986, both in Electrical Engineering. He is an Associated Professor in the Department of Computer Engineering, Faculty of Engineering, King Mongkut's Insititute of Technology Ladkrabang, Bangkok, Thailand. His main research interests are image processing, pattern recognition, image retrieval and machine learning.