Far East Journal of Mathematical Sciences (FJMS)

Volume 85, Number 2, 2014, Pages 129-149

MUTUAL INFORMATION ROUGH SETS FEATURE SELECTION AND CLASSIFICATION FOR MICROARRAY DATA ANALYSIS

Pimolrat Ounsrimuang and Veera Boonjing

Software Engineering Laboratory
Department of Computer Science
Faculty of Science
King Mongkut's Institute of Technology Ladkrabang
Bangkok 10520, Thailand
e-mail: s9062912@kmitl.ac.th
kbveera@kmitl.ac.th

Abstract

The feature selection (FS) techniques aim to reduce the subset size of an original data set, which are retained in the most useful information by selecting the most informative feature instead of irrelevant or redundant features. The benefits of FS for classification analysis can reduce the input data, improved predictive accuracy, learned knowledge is that easily understood, and reduced execution time. Many approaches based on rough set theory up to now, have operated the dependency function for measuring the goodness of the feature. However, there is not tolerance to noisy or inconsistency data, especially on high dimensional data microarray data sets. Moreover, mostly relevant information could be invisible by using only information from a positive region but neglecting a boundary region, mostly relevant may be invisible. Therefore, this paper proposes the

---

Received: November 25, 2013; Accepted: January 4, 2014

2010 Mathematics Subject Classification: 68-XX.

Keywords and phrases: feature selection, rough sets, mutual information, positive region, boundary region, classification.

maximal positive region and minimal boundary region criterion, based on rough set and mutual information, which use the different values among the information contained in the positive region, and the information contained in the boundary region. The experimental results indicate that our proposed method can increase the classification accuracy.

1. Introduction

Feature selection (FS) as known as subset selection is an important technique used in machine learning for selecting the best subset to input reducing in the preprocessing process for classification analysis. The best subset contains the least number of dimensions that remain of the most essential information of the original data set. From the good benefit of subset, it can improve the classification accuracy rate, the increasing speed of a learning algorithm, and the capability in understanding the results. Additionally, the result from an effective FS technique shows the high relevance of the subset with decision class, means the subset can predict the decision class correctly. Moreover, the attribute in the subset is not redundancy with other attributes.

The many researches in the last decade applied FS into many fields such as statistical pattern recognition [1-3], machine learning [4-6], data mining [7, 8], text classification [9], intrusion-detection [10], and gene expression analysis [11, 12] to improve the performance. Besides, many researchers [13-16] have concerned FS based on the rough set. There are shown high performance over other FS techniques.

The rough set (RS) first proposed by Pawlak [17, 18] is a mathematic model. It can solve the uncertainty, incomplete and reducing dimensional of information without the knowledge from the experts. In addition, the RS emphasizes finding minimal subset (also known as reduct), which is not only minimal redundancy of data but also remains the most information. The most existing of research rough set [19, 20, 13-16] are based on feature selection approaches that associates with data in the only positive region for fining reduct. The dependency function is used for measuring the goodness of the subset in determining an effective reduct. Their approaches are successful with numerous data sets. Nevertheless, there are neglectful on information within the boundary region. The FS process may lead to loss of much valuable information. The goodness of feature evaluation generally uses the dependency function that is not tolerant noisy data.

Especially, the determined data from only positive region has ignored important information, and noisy problem on high dimensional data such as microarray data sets [11, 12]. Some approaches [21-25] are based on rough set considering the boundary region and positive region to solve this problem, which are supposed to be conceptually separated. However, these researches using dependency value for evaluate the goodness of feature ranking that is not tolerant noisy data. Consequently, many researches showing the method based on the rough set do not tolerate noise or vagueness. The problem is mentioned that concentrates on our literature.

The RS supports the inconsistent data but ensures the value both of the condition attributes and the decision attributes are precision. In real world, the value of attributes may be imprecise. To solve noisy problem, Ziargo [26, 27] proposed variable precision rough set (VPRS) which extended from the original rough set theory as a tool for classification. VPRS deals with classification by introducing an admissible classification error (β). Many researches [22, 26-29] have been successfully applied the VPRS to solve the noisy problem. However, the feature ranking based on VPRS is calculated by dependency function that concerns with indiscernible relation and equivalence class as well as original RS theory. So, their methods based on dependency function unsuccessful applied on high relevance attributes.

The mutual information (MI) is calculated between input and class variables to solve the problem of data classification. To solve the noisy problem, the most research [30-32] applied MI instead of dependency function for feature ranking. The technique concerns the information only positive region, which it may ignore the most information. Liu [32] applied MI for subset selection getting good results. The technique gets good result; however, it spends high computation time. Battiti [31] did modify the mutual information feature selector (MIFS) using feature ranking and the largest mutual information value. Battiti showed that MI has many beneficial in feature selection and data. However, MIFS is successful with only nonlinear regression. Mutual information feature selector under uniform information distribution (MIFS-U) presenting by Kwak and Choi [30] is more effective evaluate value of MI between input and output classes. Minimal redundancy – maximum relevance (mRMR) [33] proposed by Peng at al. having two steps in feature selection. First, the mRMR selects the subset by lowest classification error value. Second, mRMR applies add or delete attribute to feature subset based on wrapper approach. This step is finding the least-feature subset from the selected feature. Technique mRMR gives the high-accuracy rate for classification system. Ideally, the good result of selection runs by complete search does not suitable because of spending high computation time. The feature sets from MIFS, MIFS-U and mRMS approaches may have many attributes that are irrelevant with decision class and redundancy with attributes. Because, measuring redundancy of their techniques causes to reduce feature, which relevant to the target class. Following, Estevez et al. [34] improving the measuring redundancy feature called *normalized mutual information feature selection* (*NMIFS*). The technique focuses the positive region only and ignoring the boundary region associating with the lost-useful information.

Recently, Parthalain et al. [21] proposed distance measure rough set attribute reduction (DMRSAR) using the information gathered from both the positive region and boundary region. DMRSAR applies a distance metric to consider the relationship between objects within the boundary region and the positive region by distance of those objects. DRMSA remains the usage of dependency value to measure feature ranking that do not tolerance with noisy data.

The main problem of RS is ignorant of the importance information in boundary region and not tolerance with noisy data in high dimensional data sets, especially microarray data sets. Therefore, we propose a new approach based on the VPRS. Moreover, the proposed technique considers managing with the noisy data or data inconsistency, relevance feature. In addition, we apply MI for measuring the goodness of feature. For experiment, running the microarray data sets compare with RSAR, and DMRSAR based RS theory. Additionally, our approach is compared with the two well-known FS techniques CNS, and ReliefF on the microarray data sets. The experiment measures classification accuracy rate, subset size, and runtimes.

This paper organizes topics as follows: Section 2 introduces the rough set theory, VPRS, and mutual information. In Section 3, we propose feature selection approach based on VPRS, mutual information and related algorithms. In the experimental Section 4, we compare our approach with four famous approaches RSAR, DMRSAR, CNS, and ReliefF. The results are compared in terms of classification accuracy, subset size, and runtimes. The classification accuracy rates are compared by four well-known classifiers SVM, C4.5, NB, and PART. Finally, Section 5 gives the conclusion.

2. Background

In this section, the basic concepts in the related theories of rough set consisting with variable precision and mutual information are described.

2.1. Rough set [17, 18]

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. V is a value of a set of attribute values in A and f is an information function $f : U, A \to V$. $IND(P)$ is called the *P-indiscernibility relation*. 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. Any subset P of attributes A there is associated an equivalence relation $IND(P)$ can define as:

$$IND(P) = \{(x_i, y_j) \in U \mid \forall a \in P,\ f_b(x_i) = f_b(x_j)\}. \tag{1}$$

For any concept $X \subseteq U$ and attribute set $P \subseteq A$, X can be approximated by the lower and upper approximations. The lower approximation of X is the set of objects of U that is exactly in X can be defined as:

$$\underline{P}(X) = \bigcup\{x_i \mid [x_i]_p \subseteq X\}. \tag{2}$$

The upper approximation of X is the set of objects of U that is possibly in X can be defined as:

$$\overline{P}(X) = \bigcup\{x_i \mid [x_i]_p \cap X \neq \varnothing\}. \tag{3}$$

The lower approximation $\underline{P}(X)$ is the union of all the elementary sets that are contained in X, and the upper approximation $\overline{P}(X)$ is the union of all the elementary sets that have a nonempty intersection with X. The rough set X is characterized by its lower and upper approximations $\underline{P}(X)$ and $\overline{P}(X)$, respectively. Here an object x_i can belong to at most one lower approximation. For any lower approximation does not contain x_i, then x_i can belong to two or more upper approximations.

2.2. Variable precision rough set

Although RST 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 RST. 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 [26, 27] introduced an extension of RST that is a 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 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

$$c(X, Y) = \begin{cases} 1 - \dfrac{|X \cap Y|}{|X|}, & \text{if } |X| > 0, \\ 0, & \text{if } |X| = 0. \end{cases} \tag{4}$$

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

$$X \subseteq_{\beta} Y \leftrightarrow c(X, Y) \leq \beta,\ 0 \leq \beta < 0.5. \tag{5}$$

Let *S* be a decision table, *X* be a nonempty subset of *U*, $0 \leq \beta < 0.5$ and $\varnothing \neq P,\ P \subseteq C$. The equivalence relation *P*, referred to as an indiscernibility relation, corresponds to a partitioning of the universe *U* into a collection of equivalence classes or elementary sets $P = \{x_1, x_2, ..., x_n\}$. The central issue of VRPS is the specification of the discernibility limits of a set in *U* by means of elementary sets of *P*. By replacing the inclusion relation with a majority inclusion relation in the original definition of the lower approximation and the upper approximation of a set they obtain the following generalized notions of β-lower approximation and β-upper approximation are defined as:

$$\underline{P}_{\beta}(X) = \{x \in U : c([x], X) \leq \beta\}, \tag{6}$$

$$\overline{P}_{\beta}(X) = \{x \in U : c([x], X) < 1 - \beta\}. \tag{7}$$

Therefore, the definitions of the positive region, the negative region and the boundary region based on VPRS are given by:

$$POS_{\beta}(X) = \{x \in U : c([x], X) \leq 1 - \beta\}, \tag{8}$$

$$BND_{\beta}(X) = \{x \in U : \beta < c([x], X) < 1 - \beta\}, \tag{9}$$

$$NEG_{\beta}(X) = \{x \in U : c([x], X) \geq 1 - \beta\}. \tag{10}$$

As well as, formally dependency can be defined based on original rough set that *X* depends on *P* in a degree $\gamma_{P\beta}(0 \leq \gamma_{P\beta} \leq 1)$. If $\gamma_{P\beta} = 1$ means *X* depends totally on *P*, and if $\gamma_{P\beta} < 1$ that *D* depends partially on *C*. The dependency is defined as:

$$\gamma_{P\beta} = \frac{|\ POS_{P\beta}(X)\ |}{U}. \tag{11}$$

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.

2.3. Mutual information based on rough set

This theory proposed by Shannon and Weaver [35] 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 MI is a measure of generalized correlation between two random variables. In addition, MI can 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 RST, 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-theoretic framework of rough sets is interpreted as follows. For any subset

$$H(P) = -\sum_{i=1}^{n} p(X_i)\log(p(X_i)), \tag{12}$$

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

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

$$H(Q|P) = -\sum_{i=1}^{n} p(X_i) \sum_{j=1}^{m} p(Y_j \,|\, X_i) \log(p(Y_j \,|\, X_i)), \tag{13}$$

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

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

$$I(Q; P) = \sum_{j=1}^{m} \sum_{i=1}^{n} p(Y_j, X_i) \log \frac{p(Y_j, X_i)}{p(Y_j)\, p(X_i)}, \tag{14}$$

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

If the mutual information between 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

$$I(P; Q) = H(Q) - H(Q|P). \tag{15}$$

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 feature selection methods based on mutual information as a measure of relevance and redundancy of features to find the most relevant features subset. In this paper, mutual information used as information measure of correlation between the lower approximation $\underline{P}(X)_\beta$ of positive region and class X. Furthermore, mutual information of the boundary region $BND_{P\beta}(X)$ with respect to the 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. Proposed Feature Selection Method

As described previously, almost all techniques for rough set attribute reduction adapt the approach to minimizing the values that determine only the information contained within the positive region of a set. Although the current mechanism (DMRSAR) [21] has been proposed on the rough-set-based method to deal with the information on the boundary region. However, DMRSAR calculates the information of the boundary region still significantly depends on the information of the lower approximation. In addition, when each lower approximation of the features is an empty set, the set of the boundary region must not be empty. DMRSAR [21] is insufficiently efficient for feature selection when applied to noisy or vague data. Furthermore, it may not be successful when applying to the data in which few equivalence classes are consistent at the first stage of the feature selection process. Therefore, the useful information of the boundary region can be used to evaluate the goodness of a feature subset when the lower approximation is an empty set.

The approach described in this section uses both the information contained in the lower approximation and the boundary region to search for the best feature subset. The calculation to approximate the sets of both lower approximation and boundary region are independent. In addition, mutual information is used as the information measure for both lower approximation and boundary region to guide the search for the best feature subset. This proposed approach selects the feature that gives the lower approximations information that is mostly relevant to class. The information of the lower approximations is subtracted by the information contained in the boundary region with respect to class.

3.1. Minimal boundary regions

As discussed above, the central problem of VPRS is the consideration in selecting a level of error in classification. Concerning the admissible classification error β, for most existing papers based on VPRS predefining is needed. Therefore, an optimal β value is taken by considering from the best results of the classification accuracy. This paper proposes a novel approach that chooses a β value automatically rather than manually predefine. This approach operates only on the information contained within the data itself.

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

$$BND_{P\beta}(D)$$

$$= \{\overline{P}(D_1)_\beta - \underline{P}(D_1)_\beta, \overline{P}(D_2)_\beta - \underline{P}(D_2)_\beta, ..., \overline{P}(D_m)_\beta - \underline{P}(D_m)_\beta\}. \quad (16)$$

Therefore, the minimal mutual information between the knowledge D and the boundary region $BNDP\_(D)$ as the criteria used to select the optimal $\widetilde{\beta}$ value is given by

$$\beta = \min_{0.0 \le \beta < 0.5} \{I(D; BND_{P\beta}(D))\}, \quad (17)$$

where β varies in the range of [0.0, 0.5) in the step of 0.05. The β value that minimizes the mutual information between the knowledge D and boundary region $BNDP\_(D)$ is selected as the $\widetilde{\beta}$ value. Besides, in this paper, the minimal mutual information can be found with varying in the range of [0.3, 0.45]. To verify that the minimal information of $I(D; BNDP\_(D))$ is equivalent to the maximal quality of classification in VPRS, equation (11) is needed to be modified. This is because the positive region in VPRS comprises not only the objects that can be classified correctly, but also all objects of elementary sets that can be classified with respect to the admissible classification error β. Therefore, we have adopted the equation in (11) to determine only the objects which can be classified correctly. The quality of the classification can be redefined as

$$QC_\beta(P) = \frac{|\,POS_{P\beta}\,| - |\,misclassify_{P\beta}(D)\,|}{|\,U\,|}, \quad (18)$$

where term $misclassify_{P\beta}(D)$ is a set of objects which cannot be classified by class categories when feature subset P and β are given.

3.2. Maximal positive region and minimal boundary region

This criterion attempts to maximize the information of the region of certainty 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 concept is expected the most relevant features obtained from the result of a feature subset. This proposed criterion is a novel concept different from most existing rough-set-based FS approaches. Besides, it is contrary to the concept of DMRSAR method [21] that uses the information gathered from both the information contained in the lower approximation and the boundary region to search for reducts. By using (17), the minimal mutual information of the boundary region with respect to knowledge *D*, for any subset of features *P* with the optimal $\widetilde{\beta}$, can be defined as

$$BndInf(P) = I(D;\ BND_{P\widetilde{\beta}}(D)). \tag{19}$$

The total information of mutual information between the lower approximation $\underline{P}(D_i)_{\widetilde{\beta}}$ and the equivalence class D_i with the optimal $\widetilde{\beta}$, denoted by $PosInf(P)$, can be defined as

$$PosInf(P) = \sum_{i=1}^{m} I(D_i,\ \underline{P}(D_i)_{\widetilde{\beta}}). \tag{20}$$

Hence, the problem of selecting feature subset *P* is equivalent to maximizing $LowerInf(P)$ and minimizing $BndInf(P)$, that is to maximize the objective function $G(P)$, where

$$G(P) = PosInf(P) - BndInf(P). \tag{21}$$

Obviously, if $PosInf(P) = H(D)$, then the objective function $G(P)$ value is maximum, it shows that the approximate information contains no uncertainty with respect to *P*. Therefore, the subset of features *P* is determined as strongly relevant features. Conversely, if $BndInf(P) = H(D)$, then *P* brings to 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 value is obtained as both operate in the range $[0, H(D)]$, and the $G(P)$ 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. PmaxBmin feature selection algorithm

Figure 1 shows a VPRS-based PmaxBminReduct algorithm based on the rough set attribute reduction (RSAR) algorithm [13]. PmaxBminReduct is similar to the RSAR algorithm but uses the maximal the objective function G value of a subset to guide the feature selection process. If value G of the current reduct candidate is greater than that of the previous, then this subset is retained and used in the next iteration of the loop. The feature selection process terminates when an addition of any remaining features results in an information function value (*PosInf*) reaching that of the unreduced data set. However, in some situations without noisy data, a value of $H(D)$ can be used as termination criterion by comparing with *PosInf* of reduct. The algorithm begins with an empty subset R. The do until loop works by determining the G 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 and an optimal $\widetilde{\beta}$ of this subset is then computed. If the difference amount of information of the current subset $(R \cup x)$ is greater the previous subset (T), then the attribute added in is retained as part of the new subset T.

The do until loop will be terminated when the amount information of the lower approximations of the current reduct candidate (*PosInf* (R)) equals the conditional attributes of the data set (*PosInf* (C)). We now analyze the time complexity of PmaxBminReduct before an empirical study of its efficiency. As we can see from Figure 1, major computation of the algorithm involves $\widetilde{\beta}$

value and *G* values for the lower approximation and boundary region, which have quadratic complexity in terms of the number of instances (*M*) in a data set. In terms of dimensionality *N*, to determine reduct, the algorithm has nonlinear complexity $O(N^2M^2)$. To determine predominant features from reduct ones (assuming all features are selected as reduct ones), it has a best-case complexity $O(NM^2)$ when only one feature is selected and all of the rest of the features are ignored, and a worse-case complexity $O(N^2M^2)$ when all features are selected.

```
PmaxBminReduct(C, D)
C is the set of all condition features.
D is the set of decision features.
$T \leftarrow \{\}, R \leftarrow \{\}$
do
    $\forall x \in (C - R)$
    compute the optimal $\tilde{\beta}$ by formula (17) for $R\ U\ \{x\}$
    if $G(R\ U\ \{x\}) > G(T)$
        $T \leftarrow R\ U\{x\}$
    $R \leftarrow T$
until $PosInf(R) == PosInf(C)$
return $R$
```

Figure 1. The PmaxBminReduct algorithm.

4. Experiment

This section presents the results of experimental studies for microarray data sets. All methods run on the platform Intel Core i7 processor. The program was implemented on Java version 7.0. After that, the result of classification accuracy was measured the accuracy with Weka version 3.6.9.

The PmaxBmin method is initially compared with the rough-set-based feature selection methods namely RSAR [13], and DMRSAR [21]. Additionally, PmaxBmin is also compared with well-known FS techniques

ReliefF [36] and the consistency based subset evaluator (CNS) [37]. For data sets containing features with continuous values, we apply the equidistance partitioning method before applying RSAR, DMRSAR, CNS and PmaxBmin to allow all methods to be compared fairly. We then apply SVM, C4.5, NB and PART classifiers on each of the newly obtained data sets and obtain overall accuracy of 10 fold cross validation. A comparison of all the feature selection techniques is made based on subset size, classification accuracy and time taken to discover subsets.

4.1. Microarray data sets

In this paper, three frequently microarray data sets are used in studies: Colon cancer [38], Leukemia [39], and Lung cancer [40]. 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 obtain the runtime and the selected genes for each algorithm. We then apply classifiers on each of the newly obtained data sets, and obtain overall classification accuracy by leave-one-out cross validation, a performance validation procedure due to a small sample size of microarray data.

Table 9. The summary of microarray data sets

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

4.2. Results and discussions on microarray data sets

The effectiveness of these five algorithms based on the number of genes selected and the leave-one-out accuracy are reported in Table 10. The classification accuracies obtained with the PmaxBmin approach are higher than all other methods for Colon Tumor data, except accuracies obtained with CNS and ReliefF that are similar to that of PmaxBmin for two and one classifiers, respectively. The results verify that the efficiency of PmaxBmin outperforms RSAR and DMRSAR in all classifiers, sometimes significantly.

Table 10. The percentage of average classification accuracy-microarray data

FS method	Classifier	Microarray datasets		
		Colon Tumor	Leukemia	Lung Cancer
RSAR	SVM	72.58	79.78	96.13
	C4.5	77.41	93.05	97.23
	NB	72.58	91.66	97.79
	PART	77.41	93.05	96.13
DMRSAR	SVM	72.58	79.78	96.13
	C4.5	77.41	93.05	97.23
	NB	72.58	91.66	97.79
	PART	77.41	93.05	96.13
CNS	SVM	79.03	84.72	88.95
	C4.5	82.25	88.88	96.13
	NB	85.48	91.66	97.79
	PART	82.25	84.72	96.68
ReliefF	SVM	79.03	90.27	97.34
	C4.5	82.25	91.66	98.94
	NB	85.40	88.88	97.79
	PART	82.25	91.66	95.58
PmaxBmin	SVM	**82.25**	79.61	95.37
	C4.5	**85.48**	**97.22**	97.23
	NB	**85.48**	**95.83**	**97.79**
	PART	**82.25**	**98.22**	**97.79**

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.

Table 11. The comparison of subset size and runtimes-microarray data

Dataset	Subset size					Time taken in locate subset				
	RSAR	DMRSAR	CNS	ReliefF	**Pmax Bmin**	RSAR	DMRSAR	CNS	ReliefF	**Pmax Bmin**
Colon Tumor	3	3	3	3	3	0.20	0.46	0.03	0.01	0.36
Leukemia	2	2	3	3	3	2.25	8.36	0.15	0.06	10.85
Lung Cancer	3	3	2	3	3	27.10	106.98	0.36	1.18	18.65

Although the number of genes selected on Leukemia data set by PmaxBmin slightly increases in subset size when compared with RSAR and DMRSAR, PmaxBmin can lead to the highest accuracy by C4.5, NB and PART classifiers shown in Table 10. It is notable for Leukemia data that PmaxBmin shows an increase of up to **9.38%** with C4.5 and more than **15%** with PART when comparing CNS with PmaxBmin. For Lung cancer data, we can see that the PmaxBmin and ReliefF approaches achieve an accuracy of classification with two and three classifiers, respectively, except the RSAR, the DMRSAR, and the CNS approaches that are with only one classifier achieved. Although the subset size obtained with CNS is smaller than all other methods, accuracy obtained with CNS of SVM, C4.5 and PART is the smallest.

As in Table 11, it is also clear from the runtime that CNS and ReliefF perform faster than all methods based on the data partition (RSAR, DMRSAR and PmaxBmin). However, the data containing a very small number of training samples and a large number of genes (thousands or tens of thousands of genes) in which PmaxBmin also runs faster than both RSAR and DMRSAR when the subset size is the same. Clearly, DMRSAR demonstrates a large increase in runtime for Lung cancer data when compared with PmaxBmin. The reason lies in the searching of the best subset in DMRSAR involves the calculation of the distance of objects in the boundary that is more costly for data containing a large number and high uncertainty of the attributes.

5. Conclusion

The comparison of PmaxBmin with RSAR and DMRSAR has shown that the PmaxBmin method is a good starting point for further work based on the information measure both certainty and boundary region for exploring the variable precision rough set.

The paper proposes a new PmaxBmin feature selection for high dimensional data such as microarry data. The new feature selection is a rough-set-based technique. It optimally combines both certainty and boundary region to be mutual information for measuring goodness of set of attributes obtained from modified VPRS training method. Instead of using a predefined admissible classification error as in previous techniques, the new method determines this value automatically using information of the boundary region. The experimental evaluation confirms that the PmaxBmin feature selection, compared with RSAR and DMRSAR, gives higher classification accuracy.

In this paper, we have used mutual information that evaluates the goodness of a subset on the training data partitioned by using VPRS. As previously discussed, proposed by many papers based on the VPRS model, the admissible classification error (β) does need to be predefined. However, it does not need to be predefined for our method. This threshold value is automatically selected by determining from the minimal information of the boundary region or the region of uncertainty. The experimental evaluation emphasizing on much valuable information is extracted by maximizing information of the lower approximation and simultaneously minimizing information contained in the boundary region of a rough set. However, it is clear from the results obtained in the previous section that an increase in the accuracy of the PmaxBmin algorithm is highly desirable and will lead to further increase in efficiency of dimensionality reduction.

References

[1] M. Arif, Evaluation of discrimination power of features in the pattern classification problem using arif index and its application to physiological datasets, Int'l J. Innovative Computing, Information and Control 7(2) (2011), 525-536.

[2] A. Jain and D. Zongker, Feature selection: evaluation, application, and small sample performance, IEEE Trans. Pattern Analysis and Machine Intelligence 19(2) (1997), 153-158.

[3] X. Wei, C. Zhou and Q. Zhang, ICA-based features fusion for face recognition, Int'l J. Innovative Computing, Information and Control 6(10) (2010), 4651-4661.

[4] A. L. Blum and P. Langley, Selection of relevant features and examples in machine learning, Artificial Intelligence 97 (1997), 245-271.

[5] R. Kohavi and G. H. John, Wrappers for feature subset selection, Artificial Intelligence 97(1-2) (1997), 273-324.

[6] X. Song, S. K. Halgamuge, D. Chen, S. Hu and B. Jiang, The optimized support vector machine with correlative features for classification of natural spearmint essence, Int'l J. Innovative Computing, Information and Control 6(3(A)) (2010), 1089-1099.

[7] M. Dash and H. Liu, Feature selection for classification, Intelligent Data Analysis: An Int'l J. 1(3) (1997), 131-156.

[8] 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, 2000, pp. 365-369.

[9] 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 33 (2007), 1-5.

[10] W. Lee, S. J. Stolfo and K. W. Mok, Adaptive intrusion detection: a data mining approach, AI Rev. 14(6) (2000), 533-567.

[11] J. C. Patra, G. P. Lim, P. K. Meher and E. L. Ang, DNA microarray data analysis: effective feature selection for accurate cancer classification, Proceedings of International Joint Conference on Neural Networks, Orlando, Florida, USA, August 12-17, 2007.

[12] S. Hengpraprohm and P. Chongstitvatana, Feature selection by weighted-snr for cancer microarray data classification, Int'l J. Innovative Computing, Information and Control 5(12(A)) (2009), 4627-4635.

[13] R. Jensen and Q. Shen, Semantics-preserving dimensionality reduction: rough and fuzzy-rough-based approaches, IEEE Transactions on Knowledge and Data Engineering 16(12) (2004), 1457-1471.

[14] Y. Chen, D. Miao and R. Wang, A rough set approach to feature selection based on ant colony optimization, Pattern Recognition Letters 31(3) (2010), 226-233.

[15] 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.

[16] Q. Hu, D. Yu, J. Liu and C. Wu, Neighborhood rough set based heterogeneous feature subset selection, Inform. Sci. 178(15) (2008), 3577-3594.

[17] Z. Pawlak, Rough sets, Int. J. Inf. Computer Science 11 (1982), 314-356.

[18] Z. Pawlak, Rough Sets: Theoretical Aspects of Reasoning About Data, Kluwer Academic Publishing, Dordrecht, 1991.

[19] 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 55(11) (2004), 954-962.

[20] Y. H. Hung, A neural network classifier with rough set-based feature selection to classify multiclass IC package products, Advanced Engineering Informatics 23 (2009), 348-357.

[21] N. M. Parthalain, Q. Shen and R. Jensen, A distance measure approach to exploring the rough set boundary region for attribute reduction, IEEE Transactions on Knowledge and Data Engineering 22(3) (2010), 305-317.

[22] D. Miao, Q. Duan, H. Zhang and N. Jiao, Rough set based hybrid algorithm for text classification, Expert Systems with Applications 36 (2009), 9168-9174.

[23] 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 Transactions on Nanobioscience 7(1) (2008), 91-99.

[24] 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, 1995, pp. 1-10.

[25] 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 2(1) (2006), 1-12.

[26] W. Ziarko, Variable precision rough set model, J. Comput. Syst. Sci. 46(1) (1993), 44-54.

[27] W. Ziarko, Probabilistic approach to rough sets, Internat. J. Approx. Reason. 49(2) (2008), 272-284.

[28] M. Beynon, Reducts within the variable precision rough sets model: a further investigation, European J. Oper. Res. 134(3) (2001), 592-605.

[29] M. Ningler, G. Stockmanns, G. Schneider, H. D. Kochs and E. Kochs, Adapted variable precision rough set approach for EEG analysis, Artificial Intelligence in Medicine 47(3) (2009), 239-261.

[30] N. Kwak and C. H. Choi, Input feature selection for classification problems, IEEE Trans. Neural Networks 13(1) (2002), 143-159.

[31] R. Battiti, Using mutual information for selecting features in supervised neural net learning, IEEE Trans. Neural Networks 5(4) (1994), 537-550.

[32] H. Liu, J. Sun, L. Liu and H. Zhang, Feature selection with dynamic mutual information, Pattern Recognition 42(7) (2009), 1330-1339.

[33] H. Peng, F. Long and C. Ding, Feature selection based on mutual information: criteria of max-dependency, max-relevance and min-redundancy, IEEE Transactions on Pattern Analysis and Machine Intelligence 27(8) (2005), 1226-1238.

[34] P. A. Estevez, M. Tesmer, C. A. Perez and J. M. Zurada, Normalized mutual information feature selection, IEEE Trans. Neural Networks 20(2) (2009), 1045-9227.

[35] C. E. Shannon and W. Weaver, The mathematical theory of communication, University of Illinois Press, Urbana, Israel, 1949.

[36] I. Kononenko, Estimating attributes: analysis and extensions of Relief, Proc. Seventh European Conf. Machine Learning, 1994, pp. 171-182.

[37] H. Liu and R. Setiono, A probabilistic approach to feature selection: a filter solution, Proc. 13th Int'l Conf. Machine Learning, 1996, pp. 319-327.

[38] 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 (1999), 6745-6750.

[39] T. R. Golub, D. K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. P. Mesirov, H. Coller, M. L. Loh, J. R. Downing, M. A. Caligiuri, C. D. Bloomfield and E. S. Lander, Molecular classification of cancer: class discovery and class prediction by gene expression monitoring, Science 286 (1999), 531-537.

[40] G. J. Gordon, R. V. Jensen, L. L. Hsiao, S. R. Gullans, J. E. Blumenstock, S. Ramaswamy, W. G. Richards, D. J. Sugarbaker and R. Bueno, Translation of microarray data into clinically relevant cancer diagnostic tests using gene expression ratios in lung cancer and mesothelioma, Cancer Research 62 (2002), 4963-4967.