Efficient selection of discriminative genes from microarray gene expression data for cancer diagnosis

D. Huang, Tommy W. S. Chow, Eden W. M. Ma, Jinyan Li*

Abstract – A new mutual information (MI) based feature selection method to solve the so-called *large p & small n* problem experienced in a microarray gene expression based data is presented. First, a grid based feature clustering algorithm is introduced to eliminate redundant features. A huge gene set is then greatly reduced in a very efficient way. As a result, the computational efficiency of the whole feature selection process is substantially enhanced. Second, MI is directly estimated using quadratic MI together with Parzen window density estimators. This approach is able to deliver reliable results even when only a small pattern set is available. Also, a new MI based criterion is proposed to avoid the highly redundant selection results in a systematic way. At last, attributed to the direct estimation of MI, the appropriate selected feature subsets can be reasonably determined.

Keywords – Cancer diagnosis, Feature selection, Grid based redundancy elimination, Microarray gene expression data, Quadratic mutual information.

1. INTRODUCTION

Microarrays are a powerful biotechnological means because they are able to record the expression levels of thousands of genes simultaneously. Through hybridizing the fluorescent DNA probe of an examined sample with that of a reference cell, the mRNA levels of the genes in the examined sample are obtained. Since the mRNA levels are roughly related to the amount of protein product, the obtained microarray result can be used to express the "state" of the examined sample. Details about the

hybridization of cDNA and monitoring of gene expression can be referred in [25]. Generally, different cells or a cell under different conditions yield different microarray results. The comparisons of microarray results between normal and cancer cells can thus provide the important information of cancer diagnosis and treatment [1-6]. Among a large amount of genes encoded in the microarray gene expression data, only a very small fraction of them are informative for a certain task. A very challenging task arises as a result – how to select the most useful features (genes) for performing data analysis such as diagnosis, prognosis, subtype classification of a heterogeneous disease and understanding of a gene network [1, 6]. This selection procedure is important and sometimes necessary because of two main reasons. First, it is impossible for biologists or physicians to examine the whole feature space (e.g. the genes in human genome) in the laboratory experiments at one time. Thus, it is necessary to recommend a small fraction of the features by using computational algorithms. Second, it is widely known that taking many irrelevant features into account amid the course of classification will increase the dimensionality of the problem, and thus results an unnecessary computational difficulties and additional noise.

Currently, various statistic or machine learning methods have been developed for gene selection [1, 4-8]. For example, in [1], Golub et al proposed a classical scheme in which features were ranked according to their linear discriminant ability with an assumption that features are orthogonal to each other. In [6], several developed techniques, including correlation, principle component analysis (PCA), discriminant analysis (DAV) and self-organizing maps (SOMs), are employed to reduce the feature set. In [7], SVM RFE recursively reduces a given feature set according to the parameters of the built linear SVM models. In [8], based on the concept of information gain and Markov Blanket filter, a given feature set is reduced gradually. All these methods are able to determine the feature subsets that can produce promising classification results. But in these methods, users have to determine the number of the selected features in advance. There is, however, no theoretical guideline for this type of determination, and the whole process is rather subjective and experience dependent. For example, in [6],

all feature selection methods were required to select 50 features. For different datasets, SVM RFE [7] recursively eliminates half of the remaining features. In [8], the top 600 features were left behind for Markov Blanket filter based process in which the size of the finial gene selection results has to be predetermined. All these size-determination strategies appeared to be random and may be wide off the target. Also, different clustering methods are developed for gene selection [26]. This type of methods is generally very computationally demanding because they require determining a large amount of similarity between genes. The determination of the number of clusters (i.e., the number of the selected genes) is also problem dependent and has always been computationally tough. In our study, a mutual information (MI for short) based forward feature selection scheme is considered. Compared with the statistical approaches, MI can reflect the arbitrary relationship between variables. Detailed discussion on MI, especially the advantages of MI over other feature selection criteria, such as consistency and correlation-based feature selection, has been thoroughly discussed and can be referred to [11, 12, 15, 18, 21, 27]. The other reason for us employing MI is that a forward MI based feature selection process can estimate the appropriate number of the selected features, which will be detailed later.

Despite these merits, the computational difficulties and complexity prevent the typical MI based forward feature selection methods [16, 18, 20] from being widely used for handling gene expression data. The computation of MI poses a great difficulty because the conditional probability density functions (pdf) should be correctly estimated prior the integration of these functions. Researchers have tried different ways to address this well-known difficulty [16, 20]. But these methods fell short when one is handling small sample sets that are generally experienced in the microarray type data. In this aspect, theoretical analysis and experimental results are detailed in [19]. Also, it is worth noting that these MI based methods employ a forward feature searching process, in which the computational complexity is $O(M^2)$, where M is the number of the given features. Obviously, it is very difficult to apply these methods directly to explore the huge gene space in a microarray type dataset. Thus, there is

a need to develop a special strategy to improve the computational efficiency. Until now, no strategy of this type has been developed in the MI based feature selection schemes.

In this paper, the above difficulties are addressed. As shown in Fig. 1, the proposed methodology, called the quadratic MI based feature selection method using a grid based clustering algorithm (QMIFS-GC for short), consists of two sequential parts:

1) A new supervised grid based algorithm is designed to sort out and discard the highly redundant features. As a result, the computational efficiency of the whole feature selection process is greatly improved without reducing the quality of the selection results.

2) In the MI based forward selection stage, the quadratic MI estimation [12, 19] and Gaussian based probability estimators [13] is employed. With them, MI can be estimated effectively even when only a small pattern set is available. Also, a new MI based criterion is introduced to filter out the redundant features in a fine way. Finally, the direct MI estimation enables us to terminate the selection process at an appropriate point where the selected feature subset has preserved the most essential information of a given feature set.

This paper is organized as follows. The next section gives the background of our study. Then, the feature cluster algorithm and MI based feature selection process of the proposed QMIFS-GC are detailed in section 3 and section 4, respectively. And in Section 5, results are listed. The discussions of these results are given. Finally, the conclusion is drawn in Section 6.

2. BACKGROUND

2.1 Mutual information

Shannon's information theory states that the uncertainty of a random variable C can be measured by entropy $H(C)$. For two variables X and C, the conditional entropy $H(C \mid X)$ measures the uncertainty about C when X is known. The mutual information $I(X;C)$ measures the certainty about C that is resolved by X. The relation of $H(C)$, $H(C|X)$ and $I(X;C)$ is

$H(C) = H(C \mid X) + I(X;C)$, or, equivalently, $I(X;C) = H(C) - H(C \mid X)$

The objective of training classification model is to reduce the uncertainty about predictions on class labels C for the known observations X as much as possible. That is, training classifier increases MI $I(X;C)$ as much as possible. In a feature selection process for classification, the goal is naturally to achieve higher values of $I(X;C)$ with the possible smallest feature subset. In fact, the MI between X and C is the distance between the joint probability density $p(c, x)$ and the product of the priori probability density functions of X and C. The commonly used MI definition is the Shannon's one (2.1) [9], which is consistent with Kullback-Leibler divergence of probability density functions.

$$I_s(X;C) = \sum_{c \in C} \int_x p(c,x) \log \frac{p(c,x)}{P(c)p(x)} dx \,. \tag{2.1}$$

Also, in [12], the quadratic MI definitions (one of them is (2.2)) were proposed based on the Euclidean distance [22] of two density functions.

$$I_{CS}(X;C) = \log \frac{(\sum_c \int p(x,c)^2 dx)(\sum_c P(c)^2)(\int p(x)^2 dx)}{(\sum_c P(c) \int p(c,x) p(x) dx)^2} \,. \tag{2.2}$$

As mentioned above, estimating MI (especially high-dimensional MI) poses a great challenge in any MI based process. The simplest way for MI estimation is based on histogram, in which through discretizing continuous probability density functions with histogram the integration operation of MI is simplified to a summation operation [15]. But this type of approach is not suitable for a high-dimensional data space. Generally, in order to guarantee the accuracy of histogram, the size of X is required to exponentially increase with the dimensionality of data space [10]. This requirement is rather difficult to be satisfied in most real world applications, especially for biomedical applications where the number of patients are usually around hundreds. Alternatively, the continuous kernel based probability density estimator is considered in our study. Actually, it has been suggested that continuous density estimators are more accurate than histograms [10, 24].

2.2 Parzen window probability density estimator

Assume that in an L-class classification dataset $X=\{x_1, x_2, \ldots\ldots, x_{Nx}\}$ and $C=\{c_1,c_2,\ldots\ldots,c_{Nx}\}$ be input variables and output class labels, respectively. The class labels are modeled as $c_i = l_k$ ($1 \le i \le Nx, 1 \le k \le Nl$). Based on Parzen window estimators [13], $p(x)$ and $p(x|c)$ are,

$$p(x) = \frac{1}{Nx}\sum_{i=1}^{Nx} p(x \mid x_i) = \frac{1}{Nx}\sum_{i=1}^{Nx} \kappa(x - x_i, \Sigma_i), \tag{2.3}$$

$$p(x \mid l_k) = \frac{\sum_{x_i \in class\, l_k} p(x \mid x_i)}{\text{Number of patterns in class } l_k} = \frac{\sum_{i} \kappa(x - x_i, \Sigma_i)}{\text{Number of patterns in class } l_k}, \tag{2.4}$$

$$P(l_k) = \frac{\text{Number of patterns in class } l_k}{Nx}, \tag{2.5}$$

where $\kappa(\bullet)$ is the kernel function of Parzen window and is set with the Gaussian function. That is,

$$\kappa(x - x_i, \Sigma_i) = G(x - x_i, \Sigma_i) = \frac{1}{(2\pi h)^{M/2}\left|\Sigma_i\right|^{1/2}} \exp\left(-\frac{(x - x_i)^T \Sigma_i^{-1}(x - x_i)}{2h^2}\right), \tag{2.6}$$

where M is the dimension of x, Σ_i is determined from the variance matrix of the overall data, and h is the bandwidth of kernel function. In our study, all the input variables are proceeded to have zero means and unit variances. Thus it is set that $\Sigma_i = I$. h is determined in a way developed in [14], i.e., $h = \left\{\frac{4}{(M+2)}\right\}^{1/(M+4)} Nx^{-1/(M+4)}$.

2.3 MI BASED FORWARD FEATURE SELECTION SCHEME

A forward searching strategy is commonly used because of its simple implementation and the relative high efficiency. Also, the difficulty of estimating high-dimensional MIs motives the adoption of a forward searching strategy in a MI based feature selection scheme. Given a classification dataset $\{X; C\}$, in which the feature set is denoted by F, a MI based forward feature selection algorithm [15, 16, 20] is generally realised as follows.

1. (Initialization) Set the selected feature set S empty.
2. For any feature (say, f_i) in F, compute the MI $I(f_i ;C)$.
3. Determine the feature that maximizes $I(f_i ;C)$. Add hat feature into S, and delete it from F.
4. Repeat the following two steps until stopping criterion is met
    a) Calculate $I(S+f_i ;C)$, for any feature (say, f_i) remaining in F.
    b) Choose the feature that maximizes $I(S+f_i ;C)$. And put that feature into S, and eliminate it from F.
5. Output the selected feature subset S.

In the above process, the MI between S and the output variable C increases gradually because the adding of input variables cannot decrease MI [9]. The incremental MI gradually decreases to zero when all the relevant features have been selected. Assume that f_a is the selected feature at certain iteration (say ith iteration), S is the selected feature subset before this iteration. The incremental MI at ith iteration is the conditional MI $I(f_a;C|S)$. Assume that f_b is the next selected feature, the incremental MI of the $(i+1)$th iteration is $I(f_b;C|S+f_a)$. In the above process, it must be that

$$I(f_a;C \mid S) > I(f_b;C \mid S). \tag{2.7}$$

Based on the definition of the conditional MI, we have

$$I(f_b;C \mid S) = I(f_b;C \mid S,f_a) + I(f_a;C \mid S) \geq I(f_b;C \mid S,f_a). \tag{2.8}$$

With (2.7) and (2.8), we have $I(f_a;C|S) > I(f_b;C|S+f_a)$. This inequality suggests that the incremental MI decreases in the above searching process. As a result, the forward process can be reliably terminated when the incremental MI is small enough that implies the unselected features at that point contain little additional information for classification. The small sample set and huge feature (gene) set of a microarray gene expression data poses two main challenges to the above MI based feature selection scheme. First, a relatively small sample set makes the estimation of high-dimensional MIs much harder. Second, a large amount of genes leads to a remarkably huge computational burden. In our study, these

difficulties are addressed. In the next section, the strategy for reducing the computational burden is described. As for the estimation of MI, it will be detailed in Section 4.

3. THE SUPERVISED GRID BASED REDUNDANCY ELIMINATION

In a microarray type data, many redundant genes exist. Filtering out those redundant genes efficiently before performing feature selection will greatly enhance the computational efficiency. For this task, The existing gene clustering methods are too computationally demanding to be employed. Alternatively, based on the concept of grid [17], a simple and fast algorithm is developed to identify the redundant genes. In details, the basic concept of the proposed grid-based redundancy elimination algorithm is that objects in a grid must be similar to each other when the size of that grid is small enough. Due to the sparsity in the high-dimensional spaces, the size of the grid becomes critical to the proposed algorithm. In order to enhance the performance robustness, an adaptive grid size, rather than a fixed grid size, is used. Using the property of MI ranking, only the features with close MI values are checked if they are within the grid. For a considered feature *f*, if a feature has similar MI value to *f* and falls within the grid around *f*, it will be removed as a redundant feature.

At the beginning of the clustering process, the MI of each gene with output variable is estimated (MI estimation approach is to be detailed in section 4.1). With these estimates, the discrimination abilities of genes are evaluated. Clustering is performed on each gene in a descending order of the MI estimates unless the ones are marked as redundant one. In each iteration, an adaptive grid is generated around the considered gene and only genes with the acceptable MI values are checked. The grid size starts at the maximum distance different in a dimension and changes until the number of redundant genes is within the pre-defined range, *GridNumRange*. The number of genes is defined by the user or determined according to the MI estimate differences. There are two types of input parameters: 1) the number of genes within the grid, *GridNumRange*, and 2) the number of genes for checking redundancy, *RedNum*. The number of genes within the grid should be given to guide the changing of the grid. The

number of genes for checking redundancy could be defined by fixing the number of genes directly or input the acceptable MI difference, which is used to determine the number of genes for redundancy check. The changing of the grid size depends upon the number of redundant genes. All the MI used in this section are estimated by using the approach mentioned in the next section. This supervised grid-based algorithm is realized as follows.

```
01: Given: {F}, which rank by I(f_i ;C), GridNumRange, RedNum
02: DO {
03:       f_c = the gene in F with highest I(f_i ;C),
04:       uppGrid = data space
05:       lowGrid = 0
06:       curGrid = data space
07:       DO {
08:                   IF number of genes within curGrid >GridNumRange THEN{
09:             curGrid = (uppGrid+lowGrid)/2 + lowGrid
10:             uppGrid = curGrid
11:            }ELSEIF number of genes within curGrid <GridNumRange {
12:             currGrid = (uppGrid+lowGrid)/2 + lowGrid
13:               lowGrid = curGrid
14:            }
15:       } While number of genes within curGrid <> GridNumRange
16:         {F_CR} = f_c , f_c+1 , … , f_c+RedNum
17:         Mark genes in F_CR lies within curGrid as redundant one and remove from {F}
18:         remove f_c from {F}
19: } While {F} is not an empty set
```

4. QMI BASED FEATURE SELECTION PROCESS

4.1 The feature selection criteria -- *MIIO* and *MISF*

The MI between the selected input variable and output, called *MIIO* below, is employed to evaluate the relevancy of *S*. That is, *MIIO(S)* denotes $I(S;C)$. To measure the similarity between *S* and a single gene f_a ($f_a \notin S$), a new MI based criterion is proposed. This criterion is called *MISF*, with which the

redundancy is handled further. In order to avoid the computational difficulties, we do not directly explore $I(S; f_a)$ to evaluate the similarity between S and f_a. Instead, *MISF* is defined as

$$MISF(f_m; S) = \arg\max_{f_i \in S}\left(\frac{I(f_m; f_i)}{H(f_i)}\right). \tag{4.1}$$

According to the property of MI, variable *a* must be redundant to a variable set having variable *b* when *a* is similar to *b*. This is the rationale behind *MISF* (4.1). A large value of *MISF* (f_m;*S*) indicates that at least one element in *S* is very similar to f_m. In that case, f_m cannot be chosen into *S*. The maximum of *MISF*(f_m;*S*) is 1 because we have $\frac{I(f_i; f_m)}{H(f_i)} = \frac{I(f_i; f_m)}{H(f_i \mid f_m) + I(f_i; f_m)} \le 1$. In our study, when *MISF*(f_m;*S*) ≥ 0.9, f_m is considered to be redundant to *S*, and can be safely rejected from *S*. Below, our MI estimation approach is detailed, in which QMI (2.2) and the probability density functions (2.3 – 2.6) are employed. Using the property of Guassian function

$$\int G(x - x_1, \Sigma_1) G(x - x_2, \Sigma_2) dx = G(x_1 - x_2, \Sigma_1 + \Sigma_2),$$

the computational difficulty of estimating MI can be overcome easily. In details, we have

$$MIIO(S) = I_{cs}(S; C) = \log \frac{V_{(c,x)^2} V_{(c)^2} V_{(x)^2}}{V_{(cx)}{}^2}, \tag{4.2}$$

where

$$V_{(c,x)^2} = \sum_c \int p(c,x)^2 dx = \left(\frac{1}{Nx}\right)^2 \sum_{k=1}^{Nl} \sum_{x_i \in class\ l_k} \sum_{x_i \in class\ l_k} G(x_i - x_j, 2I),$$

$$V_{(c)^2} = \sum_{k=1}^{Nl} P(l_k)^2 = \sum_{k=1}^{Nl} \left(\frac{\text{Number of patterns in class } l_k}{Nx}\right)^2,$$

$$V_{(x)^2} = \int p(x)^2 dx = \left(\frac{1}{Nx}\right)^2 \sum_{i=1}^{Nx} \sum_{j=1}^{Nx} G(x_i - x_j, 2I),$$

$$V_{(cx)} = \sum_c \int p(c,x) P(c) p(x) dx = \left(\frac{1}{Nx}\right)^2 \sum_{k=1}^{Nl} \left(\frac{\mid \text{Number of patterns in class } l_k}{Nx} \sum_{j=1}^{Nx} \sum_{x_i \in \text{class } l_k} G(x_j - x_i, 2I) \right).$$

Similarly, $I(f_m; f_i)$ in *MISF* (4.1) can be estimated as follows.

$$I(f_m; f_i) = \log \frac{V_{(f_m,f_i)^2} V_{(f_m)^2} V_{(f_i)^2}}{V_{(f_{m,i})}^2}, \tag{4.3}$$

where

$$V_{(f_m,f_i)^2} = \left(\frac{1}{Nx}\right)^2 \sum_{k=1}^{Nx} \sum_{j=1}^{Nx} \prod_{q=m,i} G(x_{qj} - x_{qk}, 2I),$$

$$V_{(f_q)^2} = \left(\frac{1}{Nx}\right)^2 \sum_{k=1}^{Nx} \sum_{j=1}^{Nx} G(x_{qk} - x_{qj}, 2I),$$

$$V_{(f_{m,i})} = \left(\frac{1}{Nx}\right)^2 \sum_{k=1}^{Nx} \prod_{q=m,i} \left(\sum_{j=1}^{Nx} G(x_{qk} - x_{qj}, 2I) \right), \qquad q = m, i .$$

As for $H(f_i)$ in *MISF* (4.1), it can be estimated in the same way with $I(f_m;f_i)$ in that $H(f_i) = I(f_i; f_i)$.

4.2 The forward feature selection process using *MIIO* and *MISF*

As illustrated in Fig. 1, the proposed QMIFS-GC consists of two sequential processes – the supervised grid based feature clustering process, which has been detailed in section 3, and the MI based forward feature selection process, which is to be described. Suppose that *R* is the result of the grid-based redundancy elimination. In the MI based forward process, the features in *R* are firstly ranked in a descend order of *MIIO*. The feature satisfying two constraints – having as the large *MIIO* as possible and not being redundant to the selected feature subset (determined by using *MISF*) – is identified and placed into the selected feature subset *S*. This process repeats until it is determined that there are no important features unselected. Using *R*, the forward selection process can be stated as follows.

Step 1. R is the result of the above clustering process. And the selected feature set (*S*) is set empty.

Step 2. Calculate *MIIO*(*f*) for each feature *f* in R. According to *MIIO(f)*, sort out the most important feature, f_k. Put f_k into S, delete f_k from R, and set $MIIO_1 = MIIO(f_k)$.

Step 3. Estimate $MIIO(S + f)$ for each feature *f* remaining in R.

Step 4. Identify f_k. having $MIIO(S + f_k) = \arg\max_i (MIIO(S + f_i))$, and delete f_k. from *R*.

Step 5. If the candidate feature f_k is not redundant to S, i.e., $MISF(f_k; S) \leq 0.9$, put f_k into S, set $MIIO_j = MIIO(S)$ (j is the number of the features in S), otherwise, goto Step 4.

Step 6. If $(MIIO_j - MIIO_{j-1}) / MIIO_1 \leq \gamma$, goto Step 7, otherwise, goto Step 3.

Step 7. Output the feature subset S.

In this study, the threshold in the stopping criterion γ is set with 0.05. With $\gamma = 0.05$, we know that the information beyond the selected feature subset is small enough to be ignored.

5. EXPERIMENTAL RESULTS

The proposed QMIFS-GC is compared with other gene selection methods, such as the classical gene ranking method described in [1] (FR for short in this paper) and SVM RFE [7]. Assuming that all genes are independent to each other, FR ranks genes according to the individual linear discriminant ability. To rank the genes, SVM RFE depends on SVM, a state-of-art classification model: SVM RFE firstly builds a linear SVM model using all the genes, and then according to the parameters of the built SVM model it ranks genes in a descending order of classification importance. Through discarding low-ranked ones, the current gene set is reduced by half. The process of building-SVM-discarding-half-of-genes repeats until no gene remains. We also implemented the QMI based feature forward selection method (QMIFS). QMIFS, a conventional MI based scheme, does not include the proposed feature clustering process. That is, QMIFS conducts the forward feature selection on the whole gene set. It is the single difference between QMIFS and QMIFS-GC. The comparisons between QMIFS and QMIFS-GC thus emphasis the contributions of the proposed grid-based redundancy elimination algorithm.

The four different types of classifiers are employed to evaluate the feature (gene) selection results. They are two types of support vector machine models (SVM), decision tree (DT) and k-NN rule. Decision tree and k-NN rule are available in the Weka software package (available at http://www.cs.waikato.ac.nz/~ml/weka). And the default setting was used throughout our study. Following Guyon et al in [7], we downloaded SVM model from http://www.isis.ecs.soton.ac.uk

/resources/svminfo, and used two types of SVM models – the linear SVM model (SVM-L) and the RBF SVM model (SVM-R). In this study, classification results are also employed to evaluate if the stopping point of QMIFS-GC is appropriate or not. When all these classifiers are able to achieve the best or the near-best performance before the stopping points, it can be naturally concluded that the selection results of QMIFS-GC have covered most of the important information, i.e., the stopping criterion in QMIFS-GC is reliable for the subsequent data analysis processes.

Similar to other studies, each input variable was firstly normalized to have zero mean and unit standard deviation. And with a selected feature subset, the classification models were built. Based on the results of these models performing on the testing data, the quality of that feature subset was then evaluated. Primarily because of the limitation of SVM model, SVM RFE can not be directly applied to the multi-class problems. And all our studies were conducted by using Matlab 6.1 on a PC with 1.3GHz P4 CPU and 128MB memory.

5.1 Prostate cancer classification dataset

The objective of this task is to distinguish prostate cancer cases from non-cancer cases [5]. The original raw data are published at http://www.genome.wi.mit.edu/mpr/prostate. This dataset consists of 102 samples from the same experimental conditions. And each sample is described by using 12600 features. We split the 102 samples into two disjoint groups – one group with 60 samples for training and the other one with 42 samples for testing.

First, QMIFS and QMIFS-GC were required to select 50 features. They are compared with FR and SVM RFE in terms of efficiency (in Fig. 2) and effectiveness (in Table 1). These results show that FR and SVM RFE are much faster than QMIFS and QMIFS-GC. This is because the searching strategies in FR and SVM-RFE are very simple – FR only ranks features individually, and SVM RFE reduces the remaining features in an exponential rate. The comparisons between QMIFS and QMIFS-GC clearly suggest the huge computational savings caused by the proposed redundancy elimination algorithm. In

practice, this process could reduce the number of features from 12,600 to 872 with less than 4 minutes. The results listed in Table 1 indicate that QMIFS and QMIFS-GC have the substantially similar feature selection effectiveness, and outperform SVM RFE and FR in most cases. The searching strategy in FR and SVM RFE may be too simply to consistently guarantee the better feature selection performance. With the (near) best effectiveness and the better efficiency, QMIFS-GC is the better choice for this microarray data.

In Fig. 3, the changes of *MIIO* and the incremental *MIIO* of QMIFS-GC are illustrated. They imply that QMIFS-GC stopped when 25 features had been selected. And all classifiers are able to deliver their best or near-best performance before the stopping point, as illustrated in Fig. 4. Thus, it can be asserted that QMIFS-GC is able to obtain the reliable feature selection results in this example. In Table 2, the top 8 genes selected by QMIFS-GC are briefly described. Each gene basically carries different biological meaning and exhibits different biological function. For example, 37639_at, which is also determined as one of the genes for prostate cancer classification in [5], is for human hepatoma mRNA for serine protease and it plays an essential role in cell growth and maintenance of cell morphology (referred to http://www.rzpd.de/cgi-bin/cards/). Further details on these genes can be found in the websites about genomics, such as, http://expression.gnf.org/cgi-bin/index.cgi.

5.2 Subtype of ALL classification dataset

The pediatric acute lymphoblastic leukemia (ALL) is a heterogeneous disease [6]. The correct diagnosis of the subtypes for a patient is crucial because different subtypes have different treatment plan. Over-treated or less-treated therapy could lead to serious consequences to the patient. The subtype classification of this disease has been comprehensively studied previously using gene expression profiling and supervised machine learning methods [3, 6]. The original data has been divided into six diagnostic groups (BCR-ABL, E2A-PBX1, Hyperdiploid>50, MLL, T-ALL and TEL-AML1), and a miscellaneous class that contains diagnostic samples that did not fit into any one of the above groups

(thus labeled as "Others"). There are total of 12558 features and 327 samples in this dataset. This dataset has been partitioned into two disjoint subsets, in which 215 samples were used for training and 112 were used for testing [6].

Comparative results are presented in Fig. 5 and Table 3. The running time of QMIFS-GC and QMIFS presented in Fig. 5 is the time required for selecting 150 features. Similar conclusions can be drawn – QMIFS-GC and QMIFS can deliver better feature selection results compared with FR, and QMIFS-GC is shown to be much faster than QMIFS. The change of *MIIO* is shown in Fig. 6, which shows that QMIFS-GC stops when 95 genes have been selected. In Fig. 7, it indicates that the best or the near best classification results could be obtained before this stopping point. Thus, it can be concluded that the stopping criterion of QMIFS-GC is reliable in this case. Also, by using the classification schemes adopted in [6], QMIFS-GC was compared with the gene selection methods used in [6]. These results are summarized in Table 4. In Table 5, the top 20 genes selected by QMIFS-GC are listed. Interestingly, all these top-ranked genes except the 12th one, 38596_i_at, are also reported and studied in [6] in which many different statistical feature selection methods are used to determine distinguishable genes for the ALL subtype classification.

Also, these feature selection methodologies were applied to other microarray type data, such as the colon cancer classification data and the ovarian cancer classification data (the proteomic data of this application were treated in the same way with the microarray data). In the ovarian cancer classification [4], there are 253 data samples and 15154 genes. Among these samples, 91 are control samples (non-cancer) while 162 are cancer samples. We randomly selected 150 samples for training, and the others for testing. The colon cancer classification dataset [23] consists of 62 samples and 2000 genes. The 62 samples were randomly split into two disjoint parts – one part of 40 samples for training and the other of 22 samples for testing. The results obtained in these datasets are summarized in Table 6. These results

lead to the similar conclusions: QMIFS-GC is much more efficient than QMIFS, and in most cases, QMIFS-GC and QMIFS outperformed FR and SVM RFE in terms of the gene selection result quality.

6. CONCLUSIONS

A new mutual information based feature (gene) selection scheme is proposed. In the proposed methodology, mutual information is employed for three purposes. First, with the guidance of mutual information, the newly introduced grid based approach can greatly eliminate the redundancy in a huge feature set. As a result, it is able to enhance the efficiency of the whole feature selection immensely. Second, based on mutual information, the salient features are identified gradually. The computational difficulty of estimating the high dimensional MI is addressed. Also, attributed to the characteristics of mutual information, the termination of the searching process is not determined in an ad hoc basis. Third, using mutual information, the highly redundant selection results can be avoided in a systematic way. The experimental results can support the benefits of our study. At last, based on the generic characteristics of data distribution, the appropriate selected feature subsets can be reliably estimated. However it is noted that, for different classification model, the optimal size of the selected feature subset may vary. Thus, further work will be focused on the determination of the optimal feature subset for a given classifier.

Reference:

1. 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, pp. 531--537, October 1999.
2. G. J. Gordon, R. V. Jensen, Li-Li 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, vol. 62, pp. 4963-967, 2002.

3. J. Li, H. Liu, J. R. Downing, A. E. Yeoh, and L. Wong. Simple rules underlying gene expression profiles of more than six subtypes of acute lymphoblastic leukemia (ALL) patients. Bioinformatics, vol. 19, pp. 71-78, 2003.
4. E. F. Petricoin, A. M Ardekani, B. A Hitt, P. J. Levine, V. A Fusaro, S. M. Steinberg, G. B. Mills, C. Simone, D. A Fishman, E. C. Kohn and L. A Liotta. Use of proteomic patterns in serum to identify ovarian cancer. Lancet, vol. 359, pp. 572-577, 2002.
5. D. Singh, P. G. Febbo1, K. Ross, D. G. Jackson, J. Manola, C. Ladd, P. Tamayo, A. A. Renshaw, A. V. D'Amico, J. P. Richie, E. S. Lander, M. Loda, P. W. Kantoff, T. R. Golub and W. R. Sellers. Gene expression correlates of clinical prostate cancer behavior. Cancer Cell, vol. 1, pp. 203-209, March 2002.
6. E. Yeoh, M. E. Ross, S. A. Shurtleff, W. K Williams, D. Patel, R. Mahfouz, F. G. Behm, S. C. Raimondi, M. V. Relling, Anami Patel, Cheng Cheng, D. Campana, D. Wilkins, X. Zhou, J. Li, H. Liu, C. Pui, W. E. Evans, C. Naeve, L. Wong and J. R. Downing. Classification, subtype discovery, and prediction of outcome in pediatric acute lymphoblastic leukemia by gene expression profiling. Cancer Cell, vol. 1, pp. 133-143, March 2002.
7. I. Guyon, J. Weston and S. Barnhill. Gene selection for cancer classification using support vector machines. Machine learning, vol. 46, pp. 389-422, 2002.
8. E. P. Xing, M. I. Jordan and R.M. Karp. Feature selection for high-dimensional genomic microarry data. In Proc. 18th International Conf. On Machine Learning, 2001.
9. T. M. Cover and J. A. Thomas. Elements of Information Theory. New York: John Wiley, 1994.
10. A. M. Fraser and H. L. Swinney. Independent coordinates for strange attractors from mutual information. Phys. Rev. A, 33(2), pp. 1134-1140, 1986.
11. K. Torkkola. Feature extraction by non-parametric mutual information maximization. Journal of Machine Learning Research, pp. 1415-1438, March, 2003.
12. J. C., Principe, J. W. Fisher III and D. Xu. Information theoretic learning. In Unsupervised Adaptive Filtering, eds. S. Haykin, New York, NY: Wiley, 2000.
13. E. Parzen. On the estimation of a probability density function and mode. Ann. Math. Statist., vol. 33, pp. 1064-1076, 1962.
14. B. W. Silverman. Density Estimation for Statistics and Data Analysis. London: Chapman-Hall, 1986.
15. Battiti, R. Using mutual information for selecting features in supervised neural net learning. IEEE Trans. on Neural Network, vol. 5, no. 4, pp537-550, 1994.
16. N. Kwak and C-H. Choi. Input feature selection by mutual information based on Parzen window. IEEE. Trans. on Neural Networks, vol. 13, no. 1, pp143-159, 2002.

17. J.W. Han and M. Kamber. Data Mining: Concepts and Techniques. San Meteo, Calif: Morgan Kaufmann Publishers, 2001.
18. T. W. S. Chow and D. Huang, Estimating the optimal features subset using efficient estimate of high dimensional mutual information, IEEE Trans. on Neural Networks, vol. 16, no. 1, pp. 213-224, January 2005.
19. D. Huang, T. W. S. Chow, Effective feature selection scheme using mutual information. Neurocomputing, vol. 63, pp. 325-343, August 2004.
20. B. Bonnlander. Nonparametric Selection of Input Variables for Connectionist Learning. Ph.D. thesis, CU-CS-812-96, University of Colorado at Boulder, 1996.
21. G. D., Tourassi, E. D. Frederick, M. K. Markey and C. E. Jr. Floyd, Application of the mutual information criterion for feature selection in computer-aided diagnosis. Med, Phys. vol. 28(12), pp. 2394-402, 2001.
22. P. A. Devijver, J. Kittler. Pattern Recognition: A Statistical Approach. Englewood Cliffs: Prentice Hall, 1982.
23. U. Alon, N. Barkar, D. A. Notterman, K. Gish, S. Ybarra, D. Mack, and A. J. Levine. Broad pattern of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. Proc. Natl. Acad. Sci. U.S.A. vol. 96(12), pp. 6745-6750, 1999.
24. M. Young, B. Rajagopalan, U. Lall, Estimation of mutual information using kernel density estimators. Phys. Rev. E, vol. 52, no. 3 (B), pp. 2318-2321, Sept 1995.
25. M. Schena, D. Shalon, R. W. Davis, and P. O. Brown, Quantitive monitoring of gene expression with a complementary DNA microarray. Science, vol. 270. pp. 467-471, 1995.
26. G. C. Tseng, A Comparative review of gene clustering in expression profile. In 8th International Conference on Control, Automation, Robotics and Vision (ICARCV), pp. 1320-1324, 2004.
27. D. Huang, Searching the Salient Features and Samples for Pattern Recognition. Ph.D. thesis, City University of Hong Kong, 2005.

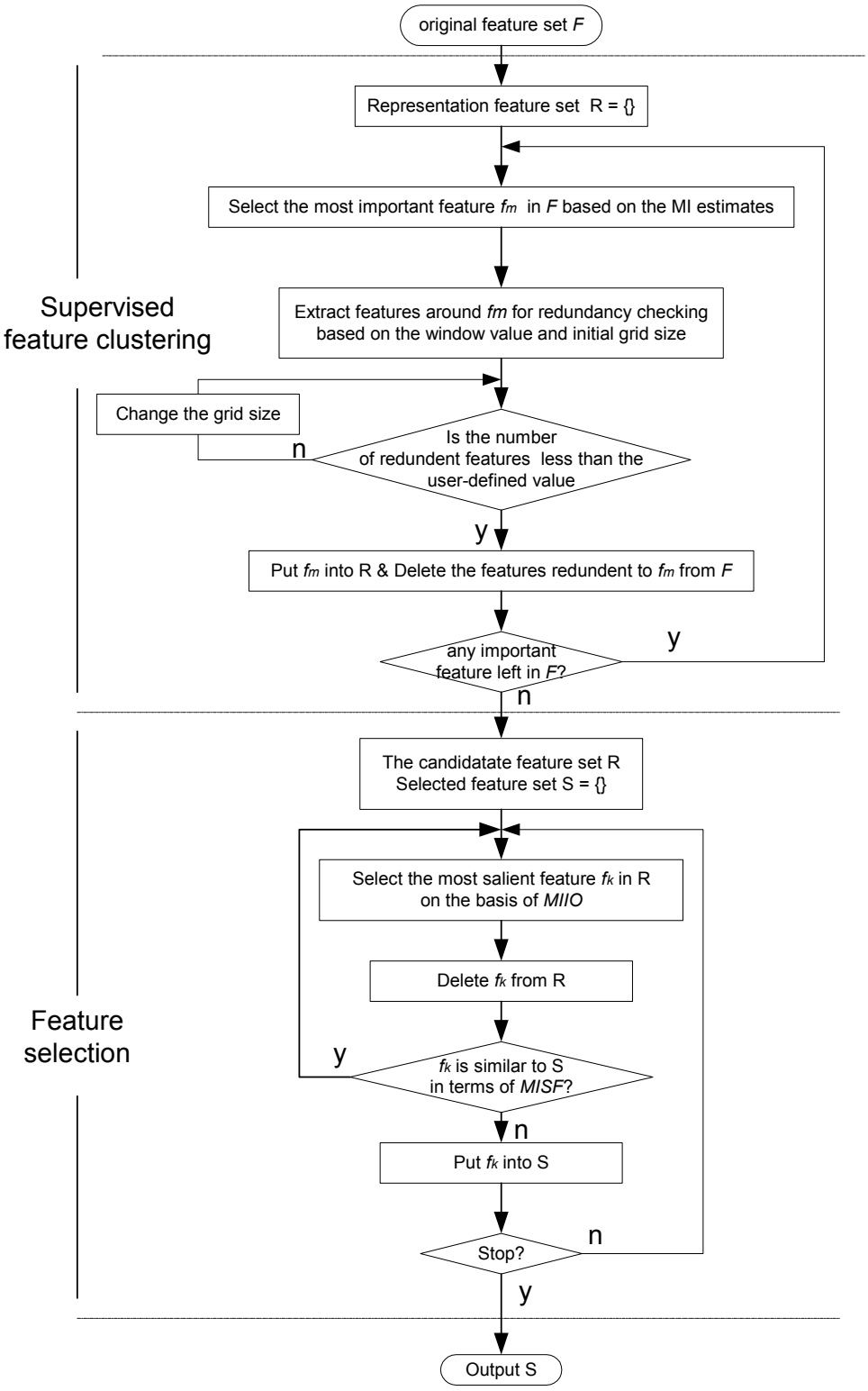

Fig. 1. The block diagram of the proposed QMIFS-GC.

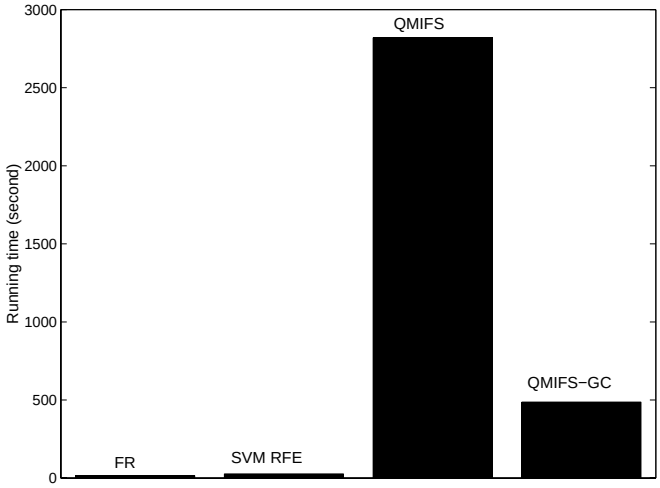

Fig. 2. Comparisons in terms of the running time on the prostate cancer classification data.

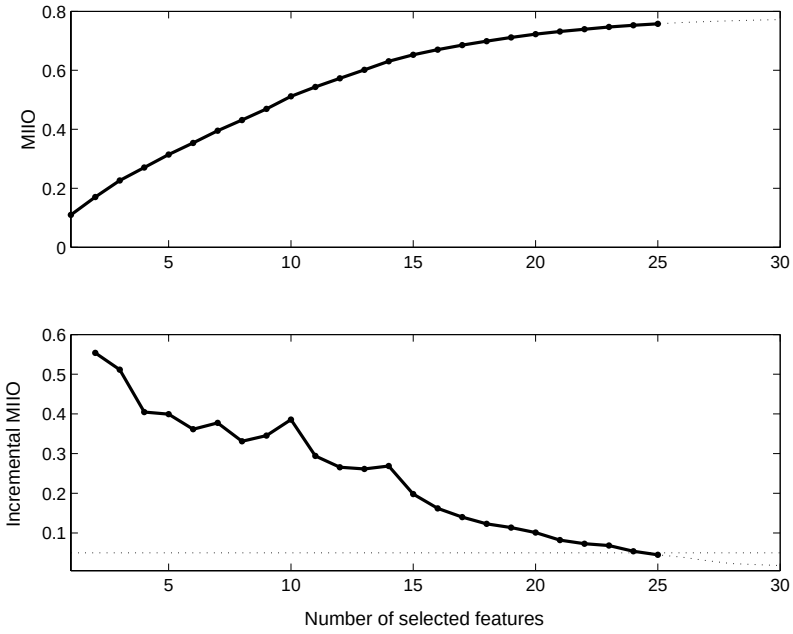

Fig. 3. The change of MIIO and the incremental MIIO with the number of the selected features on the prostate classification data.

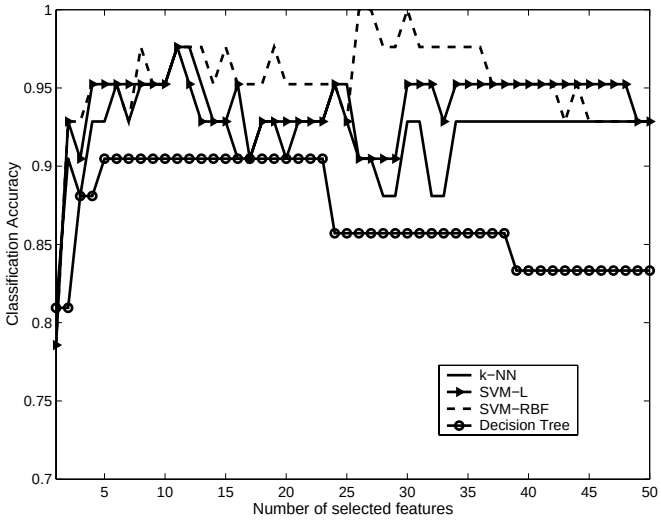

Fig. 4. Classification results of the features selected by QMIFS-GC on the prostate cancer classification data.

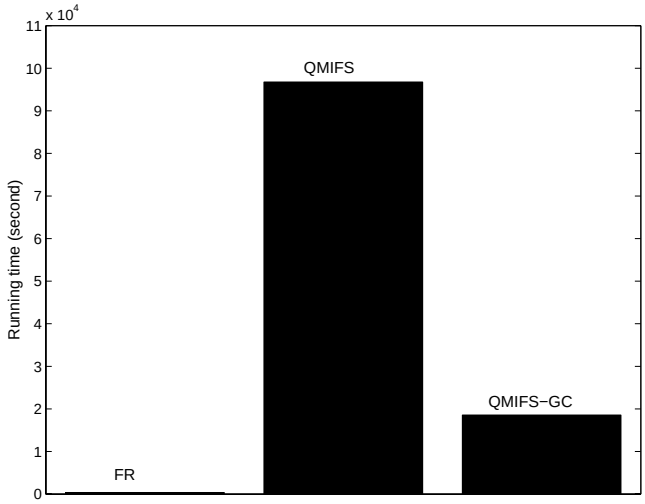

Fig. 5. Comparisons in terms of running time on the ALL subtype classification data.

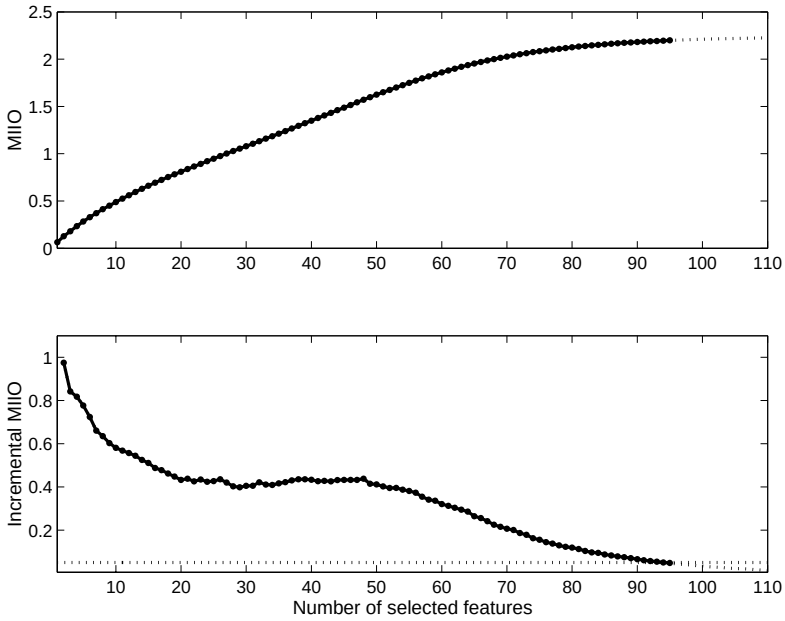

Fig. 6. The change of *MIIO* and the incremental *MIIO* with the number of the selected features on the ALL subtype classification data.

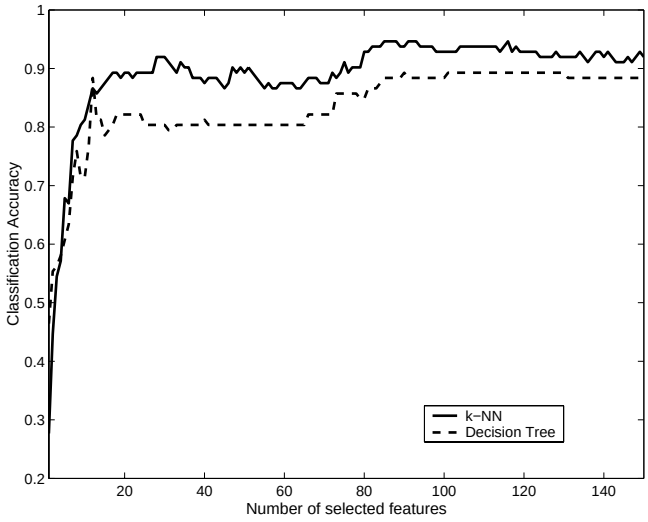

Fig. 7. Classification results of the features selected by QMIFS-GC on the ALL subtype classification data.

Table 1 Comparisons of classification accuracy on the prostate cancer classification dataset. The best result in each case is highlighted in bold face.

Number of selected features		FR	SVM-RFE	QMIFS	QMIFS-GC
k-NN	4	0.83	0.88	**0.93**	**0.93**
	8	0.83	0.93	**0.95**	**0.95**
	16	0.83	**0.90**	0.88	**0.90**
SVM-R	4	0.78	0.90	**0.95**	**0.95**
	8	0.81	0.93	**0.95**	**0.95**
	16	0.86	0.93	**0.93**	**0.95**
SVM-L	4	0.76	0.93	**0.95**	**0.95**
	8	0.83	0.93	**0.98**	**0.98**
	16	0.83	0.90	**0.98**	0.95
Decision Tree	4	0.76	**0.88**	**0.88**	**0.88**
	8	0.76	0.81	**0.90**	**0.90**
	16	0.71	0.81	**0.90**	**0.90**

Table 2 The gene selection result of QMIFS-GC on the prostate cancer classification dataset. GAN represents the gene accession number.

Selection Order	Reference Number	GeneBank	Description
1	37639_at	X07732	Human hepatoma mRNA for serine protease hepsin
2	37720_at	M22382	Human mitochondrial matrix protein P1 (nuclear encoded) mRNA.
3	38028_at	AL050152	Homo sapiens mRNA; cDNA DKFZp586K1220 (from clone DKFZp586K1220)
4	41504_s_at	AF055376	Homo sapiens short form transcription factor C-MAF (c-maf) mRNA.
5	32786_at	X51345	Human jun-B mRNA for JUN-B protein
6	36864_at	AJ001625	Homo sapiens mRNA for Pex3 protein
7	35644_at	AB014598	Homo sapiens mRNA for KIAA0698 protein.
8	38087_s_at	W72186	zd69b10.s1 Homo sapiens cDNA.

Table 3 Comparisons of classification accuracy on the ALL subtype classification dataset. The best result in each case is highlighted in bold face.

Number of selected features		FR	QMIFS	QMIFS-GC
k-NN	4	0.46	**0.57**	**0.57**
	8	0.74	**0.79**	**0.79**
	16	0.73	**0.88**	**0.88**
	32	0.72	0.89	**0.90**
Decision Tree	4	0.43	**0.58**	**0.58**
	8	0.71	**0.76**	**0.76**
	16	0.72	**0.79**	**0.79**
	32	0.76	**0.79**	**0.79**

Table 4. Comparisons between QMIFS-GC and the methods mentioned in [6]. The numbers listed in this table are classification accuracy. To evaluate a gene selection method, binary classification schemes for each subtype are constructed by using 50 genes determined by that gene selection method. The details about this classification scheme can be found in [6]. (a) The results of SVM. (b) The results of *k*-NN rule.

(a)

	QMIFS-GC	Chi sq	CFS	T-stats	SOM/DAV
T-ALL	1.00	1.00	1.00	1.00	1.00
E2A-PBX1	1.00	1.00	1.00	1.00	1.00
TEL-AML1	0.98	0.99	0.99	0.98	0.97
BCR-ABL	0.97	0.95	0.97	0.94	0.97
MLL	0.96	1.00	0.98	1.00	0.97
H>50	0.97	0.96	0.96	0.96	0.95

(b)

	QMIFS-GC	Chi sq	CFS	T-stats	SOM/DAV
T-ALL	1.00	1.00	1.00	1.00	1.00
E2A-PBX1	1.00	1.00	1.00	1.00	1.00
TEL-AML1	0.98	0.98	0.98	0.99	1.00
BCR-ABL	0.97	0.94	0.97	0.95	0.93
MLL	0.96	1.00	0.98	0.95	1.00
H>50	0.97	0.98	0.96	0.94	0.98

Table 5. The gene selection results of QMIFS-GC on the ALL subtype classification dataset. GAN represents the gene accession number.

Selection Order	Reference Number	GAN	Description
1	1077_at	M29474	Human recombination activating protein (RAG-1) gene.
2	36239_at	Z49194	H.sapiens mRNA for oct-binding factor.
3	41442_at	AB010419	Homo sapiens mRNA for MTG8-related protein MTG16a.
4	38319_at	AA919102	Homo sapiens cDNA.
5	36937_s_at	U90878	Homo sapiens carboxyl terminal LIM domain protein (CLIM1) mRNA.
6	35614_at	AB012124	Homo sapiens TCFL5 mRNA for transcription factor-like 5.
7	38968_at	AB005047	Homo sapiens mRNA for SH3 binding protein.
8	36985_at	X17025	Human homolog of yeast IPP isomerase.
9	38518_at	Y18004	Homo sapiens mRNA for SCML2 protein.
10	41097_at,	AF002999	Homo sapiens TTAGGG repeat binding factor 2.
11	33355_at	AL049381	Homo sapiens mRNA; cDNA DKFZp586J2118 (from clone DKFZp586J2118).
12	38596_i_at	D50402	Human mRNA for NRAMP1.
13	36620_at	X02317	Human mRNA for Cu/Zn superoxide dismutase
14	38242_at	AF068180	Homo sapiens B cell linker protein BLNK mRNA, alternatively spliced.
15	39728_at	J03909	Human gamma-interferon-inducible protein (IP-30) mRNA.
16	38652_at	AF070644	Homo sapiens clone 24742 mRNA sequence.
17	39878_at	AI524125	Homo sapiens cDNA.
18	2087_s_at	D21254	Human mRNA for OB-cadherin-1.
19	37344_at	X62744	Human RING6 mRNA for HLA class II alpha chain-like product.
20	35974_at	U10485	Human lymphoid-restricted membrane protein (Jaw1) mRNA.

Table 6. Comparisons on other cancer classification problems. In the columns for listing best classification accuracy, the left value is the best classification accuracy of top 50 feature subsets, and the right value is the smallest size of the feature subsets with the best classification accuracy. The running time of QMIFS and QMIFS-GC is the time for them selecting 50 features.

Feature selection methodology	Running time (second)	Best classification accuracy			
		k-NN	SVM-R	SVM-L	DT
ovarian cancer classification					
FR	46	0.99; 26	0.98; 9	1.00; 22	0.96;10
SVM RFE	351	1.00; 8	1.00; 4	1.00; 4	0.96; 4
QMIFS	2.0×10^4	1.00; 3	1.00; 3	1.00; 3	0.99; 3
QMIFS-GC	1.3×10^3	1.00; 3	1.00; 3	1.00; 3	0.99; 3
colon cancer classification					
FR	1.2	0.76; 3	0.86; 4	0.90; 12	0.81; 8
SVM RFE	3.5	0.86; 8	0.90; 8	0.81; 8	0.76; 8
QMIFS	211.6	0.86; 3	0.90; 9	0.95; 9	0.81; 3
QMIFS-GC	81.1	0.90; 11	0.95; 11	0.90; 3	0.81; 3

Response to the Editors and the Reviewers

We would like to thank the Editors and the Reviewers' comments, which are most useful in making this manuscript into a better quality. We believe that this revised paper is now in a much better format for publication.

We have gone through all the comments carefully and have made appropriate amendments accordingly. Below, these amendments are listed.

1) **Comment**: In my experience, WEKA GUI does have memory limitation. However, we can avoid this problem by entering the command line mode. Hence, if possible, I still hope the authors could make some comparisons with WEKA package in the feature selection process.

In WEKA, there are several typical feature selection methods, such as consistency-based method, and correlation-based one. In the study, we did compare the proposed MI based method with those in WEKA and other MI based ones in great detail by using the conventional data. However, the discussions of this type are beyond the scope of this paper in which the microarray type data are focused on. Thus, in the revised manuscript, we added the following explanation and a new reference 27.

"Detailed discussion on MI, especially the advantages of MI over other feature selection criteria, such as consistency and correlation-based feature selection, has been thoroughly discussed and can be referred to [11, 12, 15, 18, 21, 27]." (Page 4)

2) **Comment**: From references 18 and 19, we know the authors have two accepted papers related to the mutual information. Due to not yet being able to access for these two papers, please attach them to reviewers and point out this manuscript's novelty compared with the mentioned two papers.

We have revised the reference 18 and 19, because they have been recently published and can be accessed from Internet.

Also, the main novelty of this study is the grid-based algorithm for eliminating the redundancy (detailed in Section 3). With this algorithm, the computational difficulty of the typical MI based methods (including the ones developed in the reference 18 and 19)

is efficiently addressed. In the revised manuscript, the following lines are added in page 4.

"…it is worth noting that these MI based methods employ a forward feature searching process, in which the computational complexity is $O(M^2)$, where M is the number of the given features. Obviously, it is very difficult to apply these methods directly to explore the huge gene space in a microarray type dataset. " (Page 4)

3) **Comment:** In experimental 2, we know that the classifier is used to distinguish six ALL subtypes. Please list the individual accuracy for each subtype. In addition, from reference 3 content, we can see that the total accuracy for the same problem is 92.8%~94.6% (104/112, 106/112). It is better than this paper's best prediction result 90%. I know the present comparison is not fair (one is the hierarchical two-classes classification problem, another is multi-classes classification problem). Based on the same type classifier, please do more simulations for comparison.

To compare the proposed method with those mentioned in the reference 6, we added new experimental results in Table 4. In order to perform a fair comparison, we constructed classifiers in a way adopted in that reference.

4) **Comment:** a variety of methods have been developed that partition genes or samples into groups, or clusters, with maximum similarity, thus enabling the identification of gene signatures or informative gene subsets. It is better if the authors are able to give more explanations why their proposed method works better than the comparative methods on the test sets according to the biological meanings and/or discoveries of the clusters, selected genes, and features. Biological experts concern on biological meanings.

More discussions about the gene clustering methods are added in page 3 as following.

"Also, different clustering methods are developed for gene selection [26]. This type of methods is generally very computationally demanding because they require determining a large amount of similarity between genes. The determination of the number of clusters (i.e., the number of the selected genes) is also problem dependent and has always been computationally tough."

Also, we clarified in page 8 the reason that, for the purpose of redundancy elimination, we developed a new grid based algorithm rather than directly employed the existing gene clustering methods. The revised lines are

"*Filtering out those redundant genes efficiently before performing feature selection will greatly enhance the computational efficiency. For this task, The existing gene clustering methods are too computationally demanding to be employed.*"