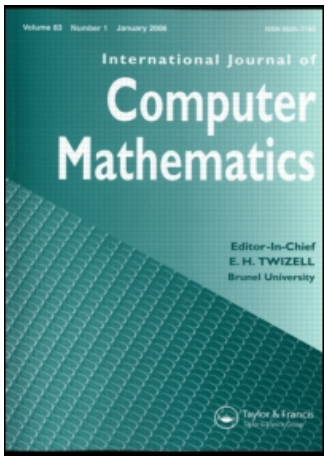

International Journal of Computer Mathematics
Publication details, including instructions for authors and subscription information:
http://www.informaworld.com/smpp/title~content=t713455451

Forward selection method with regression analysis for optimal gene selection in cancer classification

Online Publication Date: 01 May 2007
To cite this Article: Park, Han-Saem, Yoo, Si-Ho and Cho, Sung-Bae , (2007) 'Forward selection method with regression analysis for optimal gene selection in cancer classification', International Journal of Computer Mathematics, 84:5, 653 - 667
To link to this article: DOI: 10.1080/00207160701294384
URL: http://dx.doi.org/10.1080/00207160701294384

International Journal of Computer Mathematics
Vol. 84, No. 5, May 2007, 653–668

Taylor & Francis
Taylor & Francis Group

Forward selection method with regression analysis for optimal gene selection in cancer classification

HAN-SAEM PARK, SI-HO YOO and SUNG-BAE CHO*

Department of Computer Science, Yonsei University, Seoul 120–749, Korea

(Received 29 September 2006; revised version received 11 December 2006; second revision received 16 January 2007; accepted 25 January 2007)

The development of DNA microarray technology has facilitated in-depth research into cancer classification, and has made it possible to process thousands of genes simultaneously. Since not all genes are crucial for classifying cancer, it is necessary to select informative genes which are associated with cancer. Many gene selection methods have been investigated, but none is perfect. In this paper we investigate methods of finding optimal informative genes for classification of gene expression profiles. We propose a new gene selection method based on the forward selection method with regression analysis in order to find informative genes which predict cancer. The genes selected by this method tend to have information about the cancer that does not overlap with the other genes selected. We have measured the sensitivity, specificity, and recognition rate of the selected genes with the k-nearest-neighbour classifier for the colon cancer dataset and the lymphoma dataset. In most cases, the proposed method produces better results than gene selection based on other feature selection methods, yielding a high accuracy of 90.3% for the colon cancer dataset and 72% for the lymphoma dataset.

Keywords: Feature selection; Partial correlation; Cancer classification; Gene expression profiles

AMS Subject Classification: 68T10

1. Introduction

Microarray technology, which provides the expression patterns of thousands of genes simultaneously under particular experimental environments, has become an essential tool in cancer prediction and diagnosis, and computer-based analyses have been conducted to obtain useful information using this technique [1, 2]. Many researchers have have also studied cancer prediction using microarrays with gene expression data [3, 4]. However, only the genes associated with cancer are needed for prediction. Because some genes may have no function in cancer, it is very important to select informative genes before classification.

There have been many studies of feature selection methods for cancer prediction. Feature selection can be divided into the filtering approach and the wrapper approach based on evaluation criteria [5]. The filtering method evaluates feature subsets based on intrinsic characteristics of the data, whereas the wrapper method evaluates feature subsets based on the performance

*Corresponding author. Email: sbcho@cs.yonsei.ac.kr

International Journal of Computer Mathematics
ISSN 0020-7160 print/ISSN 1029-0265 online
http://www.tandf.co.uk/journals
DOI: 10.1080/00207160701294384

of the classifier. Usually, studies of gene selection for microarray data analysis focus on the filtering method because of the computational deficiency of the wrapper method, and several feature selection methods for selecting informative genes have been proposed [6]. However, filtering methods based on gene rank do not take into account the partial correlations among the selected genes because they only calculate the similarity between the target (cancer) and the gene itself on a one-to-one basis. If the partial correlations of the selected genes are not considered, the subset of the chosen genes can contain redundant information. Therefore, for efficient classification, it is important to consider partial correlations of the selected genes.

In this paper we propose a novel filtering method based on forward selection in regression analysis. This approach is different from previous methods because the genes are selected by their partial correlations rather than by their ranks. Correlations among selected genes are considered in order to minimize redundant information in the subset of selected genes [7]. Reducing the redundant information about the cancer in the selected genes helps to classify the cancer. The selected genes are input into a classifier which is trained with this input to adjust the result with the genes selected. Many classifiers have been used in cancer prediction, including the multilayer perceptron [8], the support vector machine (SVM) [9], and the k-nearest-neighbour [10]. We have used the k-nearest-neighbour classifier to verify the proposed method with a colon cancer dataset and a lymphoma dataset of gene expression profiles. The results are compared with six representative filter-based feature selection methods. We have used three measures (sensitivity, specificity, and recognition rate) to evaluate the performance of the proposed method.

2. Background

2.1 *DNA microarrays*

DNA arrays consist of a large number of DNA molecules spotted in a systematic order on a solid substrate. Depending on the diameter of each DNA spot on the array, arrays are categorized as microarrays (diameter $<250\,\mu$m) or macroarrays (diameter $>300\,\mu$m). Arrays on a small solid substrate are also referred to as DNA chips. This method is so powerful that gene information can be obtained very rapidly, because hundreds of genes can be analysed simultaneously on the DNA microarray.

There are two representative DNA microarray technologies: cDNA microarray technology and oligonucleotide microarray technology. cDNA microarrays are composed of thousands of individual DNA sequences printed in a high-density array on a glass microscope slide using a robotic arrayer. High-density oligonucleotide microarrays [2, 11, 12] are produced using spatially patterned light-directed combinatorial chemical synthesis, and contain hundreds of thousands of different oligonucleotides on a small glass surface.

mRNA samples obtained by cDNA microarray technology are labelled using different fluorescent dyes (red fluorescent dye Cy5 and green fluorescent dye Cy3). After the hybridization of these samples with the arrayed DNA probes, the slides are imaged using a scanner which makes fluorescence measurements for each dye. The log ratios of the intensities of the two dyes are used as the gene expression data:

$$gene\_expression = \log_2 \frac{\mathrm{Int(Cy5)}}{\mathrm{Int(Cy3)}} \tag{1}$$

where Int(Cy5) and Int(Cy3) are the intensities of the red and green dyes, respectively. Since a DNA microarray comprises hundreds of genes, we can obtain genome-wide information very rapidly.

2.2 *Related work*

Because the amount of DNA microarray data is usually very large, it is essential to analyse it efficiently. Not all of the thousands of genes whose expression levels are measured are needed for classification. Because microarray data consist of large number of genes in small samples, we need to select the informative genes for classification. This process is referred to as gene selection [10].

Gene selection approaches can be classified as either filtering or wrapper methods based on the evaluation criteria. There have been several studies of the wrapper method. Guyon *et al.* [13] proposed elimination of recursive features using the SVM (SVM-RFE), and several groups have used this method for gene selection [14–16]. Wrapper methods provide high accuracy and outperform filtering methods, but most studies on gene selection for microarray data analysis have focused on filtering methods because wrapper methods have a high computational cost when combined with sophisticated algorithms such as SVMs [17].

Filtering methods have been used as the preprocessing step for classification. Hall [18] and Golub *et al.* [19] used correlation-based feature selection methods, and Furey *et al.* [20] used the signal-to-noise ratio based on information theory. Euclidean distance and the cosine coefficient have also been used to calculate the similarity between genes, and information gain and mutual information have been calculated for dependencies among genes [6].

In addition to gene subset selection methods, methods which derive new features from original genes are available [21]. Nguyen and Rocke [22] compared partial least squares with principal components analysis (PCA) which is a representative technique for reducing dimensionality (number of features) [23]. A singular-value decomposition method has been proposed for the same purpose [24]. This approach seems to provide higher performance in some cases, but because features are extracted using transformation they cannot be analysed.

As mentioned earlier, most studies of gene selection adopt filtering methods. Statistical correlation-based methods such as Pearson's and Spearman's correlation coefficients are most commonly used to select informative genes by calculating the similarity between variables. These methods select a variable which is highly correlated with the target variable in order of rank. The correlation coefficient r varies from -1 to $+1$, so that the data distributed near the line biased in the $(+)$ direction will have positive coefficients, and the data near the line biased in the $(-)$ direction will have negative coefficients. Data near zero indicate that the relationship between two variables is very weak. The coefficients r_{pearson} and r_{spearman} for two vectors X and Y containing N elements are calculated as follows:

$$r_{pearson} = \frac{\sum XY - \left(\sum X \sum Y/N\right)}{\sqrt{\left(\left(\sum X^2 - \left(\sum X^2/N\right)\right)\left(\sum Y^2 - \left(\left(\sum Y\right)^2/N\right)\right)\right.}} \tag{2}$$

$$r_{spearman} = 1 - \frac{6\sum(Dx - Dy)^2}{N(N^2 - 1)} \tag{3}$$

where D_x and D_y are the rank matrices of X and Y, respectively.

The similarity between gene vectors X and Y can be thought of as a distance. Distance is a measure of how far apart the two vectors are located, and the distance tells us how likely it is that a certain gene belongs to a particular tumour class. If it is larger than a given threshold, the gene belongs to tumour class; otherwise, it belongs to the normal class. In this paper, we have adopted the cosine coefficient (r_{cosine}) represented by the following equation:

$$r_{cosine} = \frac{\sum XY}{\sqrt{\sum X^2 \sum Y^2}}. \tag{4}$$

We have also utilized the information gain and mutual information that are widely used in many fields such as data mining. Information gain (IG) is defined by

$$IG = \sum \left(\frac{l_i}{n} \log \frac{l_i}{n_l} + \frac{r_i}{n} \log \frac{r_i}{n_r} \right) - \sum_i \left(\frac{l_i + r_i}{n} \right) \log \left(\frac{l_i + r_i}{n} \right) \tag{5}$$

where n is the total number of genes, n_l (n_r) is the number of genes in the left (right) partition, l_i (r_i) is the number of genes belonging to class i in the left (right) partition, and c is the class of the ith gene [25]. Mutual information (MI) is defined by

$$MI = \sum_j \sum_i p(g_i, c_j) \log \frac{P(g_i, c_j)}{P(c_j) \cdot P(g_i)} \tag{6}$$

where $P(g_i)$ and $P(c_j)$ are the numbers of excited genes and unexcited genes, respectively.

Mutual information tells us the dependency relationship between two probabilistic variables of events. If two events are completely independent, the mutual information is zero. The more closely they are related, the higher the mutual information becomes [26]. Information gain is used when the sample features are extracted by inducing the relationship between gene and class by the presence frequency of the gene in the sample. Information gain measures the goodness of the gene using its presence and absence within the corresponding class.

Each gene g_i is from either a tumour sample or a normal sample. If we calculate the mean μ and standard deviation σ from the distribution of gene expressions within their classes, the signal-to-noise ratio $SN(g_i)$ of gene g_i can be determined as follows [19]:

$$SN(g_i) = \frac{\mu_{tumor}(g_i) - \mu_{normal}(g_i)}{\sigma_{tumor}(g_i) + \sigma_{normal}(g_i)}. \tag{7}$$

These filtering methods are widely used for gene selection, but have the disadvantage of ignoring the interrelations between the selected genes. The selected subset of genes may contain redundant information. Recently, some studies have considered this redundant information among selected features [27, 28]. Gilad-Bachrach *et al.* [27] proposed a non-redundant feature selection method based on a margin, i.e. a geometric measure which evaluates the confidence of a classifier with respect to its decision. In this paper, a non-redundant feature selection method based on regression analysis is proposed for gene selection.

3. Forward selection method

In regression analysis the partial correlations between the target variable and the variables that explain the target well are analysed. Unlike correlation analysis, regression analysis can predict or analyse the impact of one or more variables on another variable [29]. In this method, one variable is chosen as the target, and independent variables which affect this target variable are sought. If there is only one independent variable that explains the model, the model is known as a linear regression model, and if there is more than one variable that explains the model, it is known as a multiple regression model.

The basic concept of regression analysis is shown in figure 1. Although the amount of information represented by B is larger than that of C, the regression model selects C first because $(A + C)$ covers a larger area than $(A + B)$. The order of selecting genes is quite different. The correlation analysis selects (A, B, D, C) in order of the size of the area they cover. However, the regression analysis selects (A, C, D, B) in order of the size of exclusive

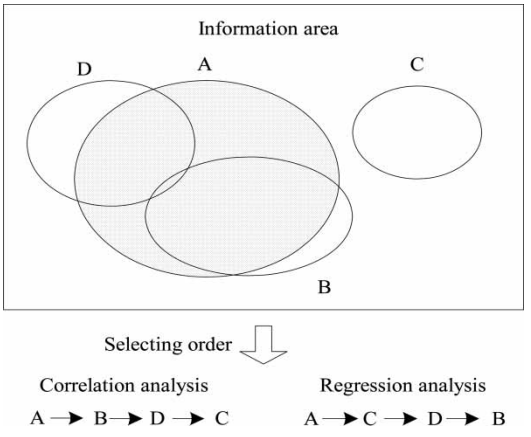

Figure 1. Basic concept of correlation and regression analyses.

area, considering the area covered by the previously selected genes. The regression analysis considers the relations between selected genes that minimize redundancy. When applying regression analysis to gene expression profile data, we use a multiple regression model because there are many genes which could affect the target variable, i.e. the presence of cancer in the sample.

A linear regression model with a target variable y and independent variable x is given by

$$y = \beta_0 + \beta_1 x_i + \epsilon_i, \quad i = 1, 2, \ldots, n \tag{8}$$

and a multiple regression model with the same target variable y and multiple variables x is given by

$$y = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \varepsilon_i, \quad i = 1, 2, \ldots, n \tag{9}$$

where β_0 and β_1 are constants estimated by observed values of x and the target variable y, and ε is estimated by normal distribution that has a mean of zero and a variance of σ^2.

The sum of the squared residuals is given by

$$SSE = \sum_{i=1}^{n} (y_i - \textit{predicted}\, y_i)^2. \tag{10}$$

A large value of *SSE* means that the regression line is predicted poorly. The total sum of squares is given by

$$SSTO = \sum_{i=1}^{n} (y_i, \bar{y})^2 \tag{11}$$

where $\bar{y}$ is the average of y_i, and the distance *SSR* between *SSTO* and *SSE* is a useful measure of the statistical performance of the prediction model:

$$SSR = SSTO - SSE. \tag{12}$$

In a regression model, selecting the variables which explain the target variable depends on R^2 value of the variables:

$$R^2 = \frac{SSR}{SSTO}. \tag{13}$$

This equation means that y is explained by $x$as a ratio of R^2. Therefore the variables that are good at explaining the target variable are selected in order of their R^2 values. The value of R^2

varies between 0 and 1. If the value of R^2 reaches 1 at a certain point when some variable x is added to the model, x is said to be a good variable that has a strong effect on the target value. Regression models are verified by the F-test, and each F-value of a regression model is calculated and evaluated as the fitness of that model. Selecting the model depends on fitness as assessed by the F-value:

$$F = \frac{SSR/1}{SSE/(n-2)}. \tag{14}$$

The label of each sample is set as a target variable of the model. The label is 1 if the sample is from a tumour and 0 if it is normal. The algorithm of the proposed feature selection approach using forward selection method is as follows.

G is a subset of selected genes and Max_R^2 is the maximum value of R^2 in the regression models developed. Initially, there are N genes and we develop a regression model for each gene; thus N regression models are obtained. Then we compute the R^2 value of each model and determine the Max_R^2 value. If $Max\_R^2 > 0$ (which means that the model could explain the target), the gene of that model is added to G, which is a subset of the selected genes. If $Max\_R^2 = 0$ (which means that the model cannot explain the target), the algorithm is terminated.

PROCEDURE

var N: total number of genes
G: set of selected genes
x_G: genes in G
function $Model(x)$: a function that applies gene x to a regression model as explained above
$Model(x_G, x)$: a function that applies gene x to a regression model as explained above and calculates R^2 value of x
$FindMax\_R^2(x)$: a function that finds a gene whose R^2 value is the largest
$FindMax\_R^2(x_G, x)$: a function that finds a gene whose R^2 value is the largest considering partial correlations of all x_G
$UpdateG(x)$: a function that updates set G by adding a new element x

```
begin
    for i = 1 to N
        Model(x_i)
    FindMax_R^2(x_i) and UpdateG(x_i)
    do
        for i = 1 to N
            Model(x_G,x_i), x_G ≠ x_i
        FindMax_R^2(x_G,x_i) and UpdateG(x_i)
    while Max_R^2(x_i) > 0
end
```

During the iterations, we make new gene sets ($gene_{selected} + gene_{new}$), where $gene_{selected}$ is a set of the genes selected up to the previous step and $gene_{new}$ is a gene selected in the current step. $Gene_{selected}$ starts with an empty set, and new regression models with the new gene sets are calculated to select the model with the largest R^2 value:

$$y = \beta_0 + \sum_{i=1}^{m} \beta_i gene_{selected,i} + \beta_{m+1} gene_{new} + \varepsilon \tag{15}$$

where m is the number of $genes_{selected}$ in the current step. Since this method selects the genes according to the relations with the selected genes, it decreases redundant information about the cancer and constructs an optimal gene set to predict cancer.

We have used the k-nearest-neighbour (KNN) classifier to classify the selected genes, because only a few samples are available. Since there are not as many samples in the microarray data as in other datasets, classifiers with parameter tuning, such as the neural network, experience difficulties in this domain. KNN is one of the most common methods of memory-based induction. Given an input vector, KNN extracts the k closest vectors in the reference set based on a similarity measure, and decides the label of input vector using the labels of the k nearest neighbours. Pearson's correlation can be used as the similarity measure. When we have an input X and a reference set $D = \{d_1, d_2, \ldots, d_N\}$, the probability $P(X, c_j)$ that X belongs to class c_j is defined as follows:

$$P(X, c_j) = \sum_{d_i \in kNN} \mathrm{Sim}(X, d_i) P(d_i, c_j) - b_j \tag{16}$$

where $\mathrm{Sim}(X, d_i)$ is the similarity between X and d_i, and b_j is a bias term.

4. Experiments

4.1 *Experimental environments*

The proposed method for identifying significant genes is applied to a colon cancer dataset containing 62 samples of colon epithelial cells taken from patients with colon cancer. Each sample contains 2000 gene expression levels. Forty of the 62 samples are colon cancer samples and the remainder are normal samples. Each sample was taken from either the tumour or the normal healthy part of the colon of the same patient and measured using high-density oligonucleotide arrays [3]. Half of the 62 samples were used as training data and the remaining 31 were used as test data (further details available at: http://www.sph.uth.tmc.edu/hgc/default.asp).

We have also applied the proposed method to a lymphoma dataset consisting of 24 GC B-like samples and 23 activated B-like samples. Each sample contains 4026 genes. Twenty-two of the 47 samples were used as training data and the remainder were used as test data (further details available at: http://genome-www.stanford.edu/lymphoma).

For evaluation, we have used sensitivity, specificity, and recognition rate. Sensitivity is the percentage of samples that are recognized as cancer which are really cancer. Specificity is the percentage of samples that are recognized as normal which are really normal.

4.2 *Results for the colon dataset*

Table 1 lists the genes selected by the forward selection method. We have selected 18 genes with $R^2 > 0$. The ID of the first gene selected is R8712 (MYOSIN HEAVY CHAIN, NON-MUSCLE) which shows higher specificity than sensitivity or recognition rate (figure 2). This gene provides useful information about normal samples, but not much information about tumour samples. The third gene selected U3662 (Human Y-chromosome RNA recognition motif protein gene) shows higher sensitivity than specificity (figure 2). This gene provides useful information about tumour samples.

Table 2 summarizes R^2 value, the F-value, and significant levels of the F-value of the genes selected by the forward selection method. Gene493 is the first gene selected from 2000 genes, with the greatest R^2 value in the regression models. This gene has an F-value of 50.33,

Table 1. Colon cancer dataset: genes selected by the forward selection method.

Rank	Gene ID	Gene annotation
1	R8712	MYOSIN HEAVY CHAIN, NONMUSCLE (*Gallus gallus*)
2	U0202	Human pre-B-cell enhancing factor (PBEF) mRNA, complete cds
3	U3662	Human Y-chromosome RNA recognition motif protein (YRRM) gene, exon 12, partial cds, subclone 7S2.
4	H6253	SPORE GERMINATION PROTEIN B2 (*Bacillus subtilis*)
5	T7102	Human (HUMAN)
6	H5607	GTP CYCLOHYDROLASE I (*Homo sapiens*)
7	T9947	GLUCOSE-6-PHOSPHATASE (*Homo sapiens*)
8	J0014	Human dihydrofolate reductase pseudo-gene (psi-hd1)
9	M2821	*Homo sapiens* low-density lipoprotein receptor (FH 10 mutant causing familial hypercholesterolemia) mRNA, 3' end
10	H2475	FRUCTOSE-BISPHOSPHATE ALDOLASE A (HUMAN)
11	R4985	COAGULATION FACTOR V PRECURSOR (*Homo sapiens*)
12	T9855	DNA-DIRECTED RNA POLYMERASES I AND III 16 KD POLYPEPTIDE (*Saccharomyces cerevisiae*)
13	T4964	MYRISTOYLATED ALANINE-RICH C-KINASE SUBSTRATE (*Homo sapiens*)
14	T6109	ENDOGLIN PRECURSOR (*Homo sapiens*)
15	M8473	Human autoantigen calreticulin mRNA, complete cds
16	H6439	CALCINEURIN B SUBUNIT ISOFORM 1 (*Homo sapiens*)
17	T7258	GLUTAMATE RECEPTOR 5 PRECURSOR (*Homo sapiens*)
18	H1506	PROTEIN KINASE CLK (*Mus musculus*)

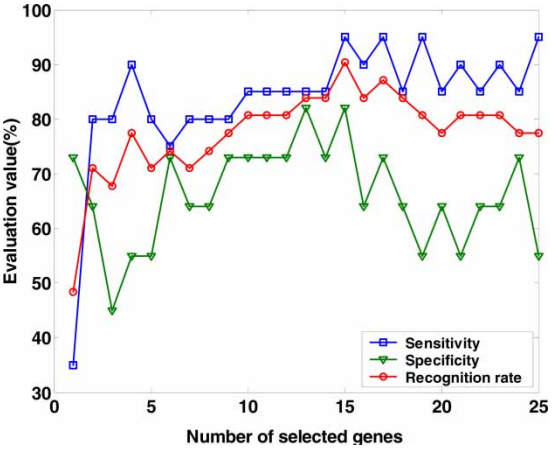

Figure 2. Traces of three evaluation criteria of the genes selected by the forward selection method.

which is very high, and the significance level of the F-value is less than 0.0001, which is very confident. The third gene selected gene has an F-value of 9.74, which is lower than the others, and low confidence. Except for a few genes with a low F-value, most of the genes selected by the forward selection method show quite high confidence level and provide important information about the cancer.

Figure 2 shows the results obtained with the forward selection method. When only one gene is selected, it performs very poorly (sensitivity 35.0%, specificity 73.0%, and recognition rate 48.4%). All three criteria are highest when the number of selected genes reaches about 15. Genes with low R^2 values, such as gene287, gene92, gene332, and gene858 in table 2, do not provide much information for classifying the cancer. This shows that genes with low R^2 values are less meaningful.

Table 2. Colon cancer dataset: R^2, F-value, and significant levels of the genes selected by the forward selection method.

Rank	Gene number	Partial R^2	F-value	Pr > F
1	gene493	0.6344	50.33	<0.0001
2	gene1147	0.1549	20.58	<0.0001
3	gene1927	0.0559	9.74	0.0043
4	gene1587	0.057	15.15	0.0006
5	gene66	0.0322	12.29	0.0017
6	gene1427	0.0218	11.99	0.002
7	gene597	0.0157	12.94	0.0015
8	gene1919	0.0133	19.93	0.0002
9	gene1584	0.0053	11.94	0.0024
10	gene55	0.0031	9.74	0.0054
11	gene459	0.0028	14.74	0.0011
12	gene1340	0.0019	19.84	0.0003
13	gene2000	0.0007	13.09	0.0021
14	gene955	0.0004	10.26	0.0055
15	gene287	0.0002	8.78	0.0097
16	gene92	0.0002	11.68	0.0042
17	gene332	0.0001	14.83	0.002
18	gene858	0.0001	20.94	0.0006

In figure 3, the pattern of the genes selected by the forward selection method is different and they have different expression levels on the same samples. They compensate each other and reduce the redundancy. However, the genes selected by Pearson's correlation coefficients have a similar pattern and similar expression levels. They are the top three ranked genes which have high correlation coefficients individually, but are not partially correlated with each other. The result of hierarchical clustering of the selected genes by the forward selection method does not show any distinguishable expression patterns (figure 4). This proves that the selected genes have different patterns of expression levels and low redundancy. We estimate the three measures for the selected genes and compare them to with those of the other six feature selection methods. Forward, Pearson, Spearman, Cosine, IG, MI, and S2N stand for forward

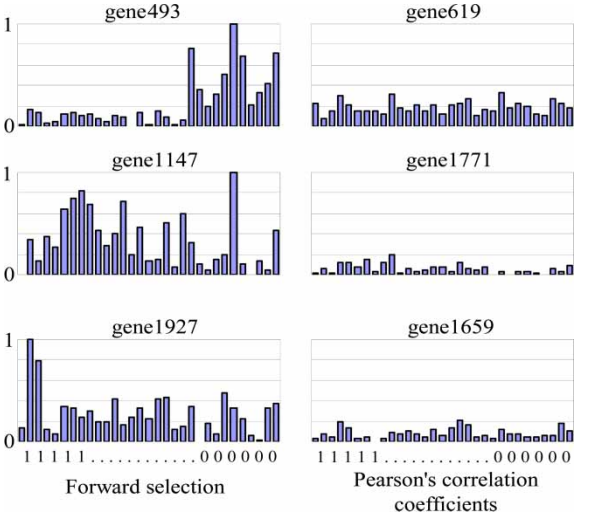

Figure 3. Expression levels of the genes selected. The expression levels of the top three ranked genes have been normalized as 0 to 1 and the figure shows the expression levels for all 31 training samples on these genes.

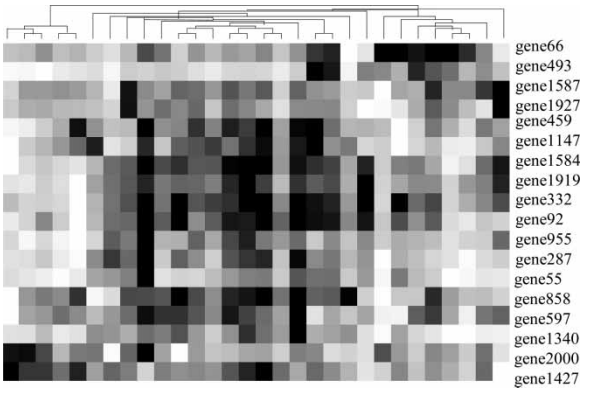

Figure 4. Hierarchical clustering of gene expression data for the colon data using the top 18 genes selected by the forward selection method.

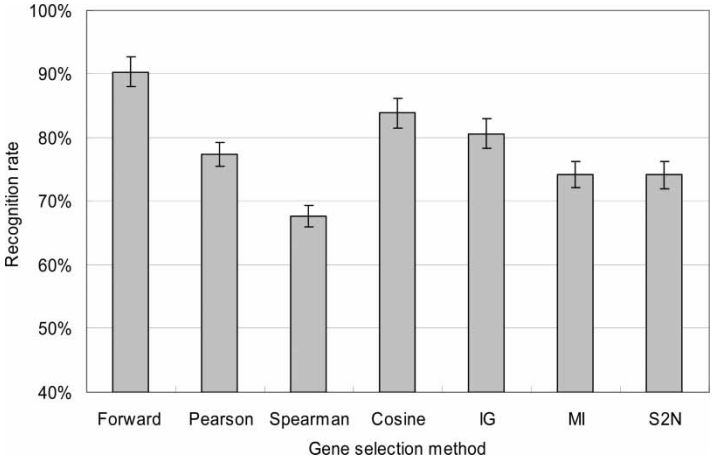

Figure 5. Colon cancer dataset: recognition rates of the gene selection methods.

selection method, Pearson's correlation coefficient, Spearman's correlation coefficient, cosine coefficient, information gain, mutual information, and signal-to-noise ratio, respectively.

We have conducted 15 runs by changing the value of k (1–15) in KNN and measured the recognition rates of the seven gene selection methods. Figure 5 shows the average values and the standard deviations for each case. The forward selection method has a higher average recognition rate than the other methods.

Table 3 shows the results of the best performance when the genes are selected by the forward selection method. We have determined the optimal number of genes from figure 2 (15 genes are selected in all the cases with the colon dataset). The forward selection method performs best performance in recognition rate and specificity, and shows a high performance in sensitivity.

Table 3. Colon cancer dataset: best results (%) of the evaluation measures by seven gene selection methods.

	Forward	Pearson	Spearman	Cosine	IG	MI	S2N
Recognition rate	90.3	77.4	67.7	83.9	80.7	74.2	74.2
Sensitivity	95.0	75.0	100.0	95.0	95.0	90.0	90.0
Specificity	82.0	82.0	9.0	63.6	54.5	54.5	54.5

Table 4. Colon cancer dataset: confusion matrix of selected genes by seven gene selection methods.

Forward				Pearson				Spearman				Cosine			
		P				P				P				P	
		0	1			0	1			0	1			0	1
A	0	9	2	A	0	9	2	A	0	1	10	A	0	7	4
	1	1	19		1	5	15		1	0	20		1	1	19

IG				MI				S2N			
		P				P				P	
		0	1			0	1			0	1
A	0	6	5	A	0	5	6	A	0	5	6
	1	1	19		1	2	18		1	2	18

P, number of predicted samples; A, number of actual samples.

In Spearman's correlation coefficient, the sensitivity is perfect (100%) and is better than the forward selection method (95%), but the recognition rate (77.4%) and specificity (9%) are low compared with the forward selection method. Spearman's correlation coefficient classifies the cancer samples very well, but its ability to classify normal samples is poor. The forward selection method classifies both tumour and normal samples well and shows a high performance in all the criteria.

Table 4 shows a confusion matrix for the selected genes. The test sample predicted to be tumour is denoted 1 and that predicted to be normal is denoted 0. The forward selection method predicts 19 (19/20) samples as tumours which are actually tumours, and predicts nine (9/11) samples as normal which are actually normal. The forward selection method performs well in terms of both sensitivity and specificity.

Figure 6 shows receiver operating characteristic (ROC) curves for the selected genes for all gene selection methods. FPF and TPF denote the false-positive fraction (the ratio of the number of false-positive decisions to the number of actual positive data) and the true positive

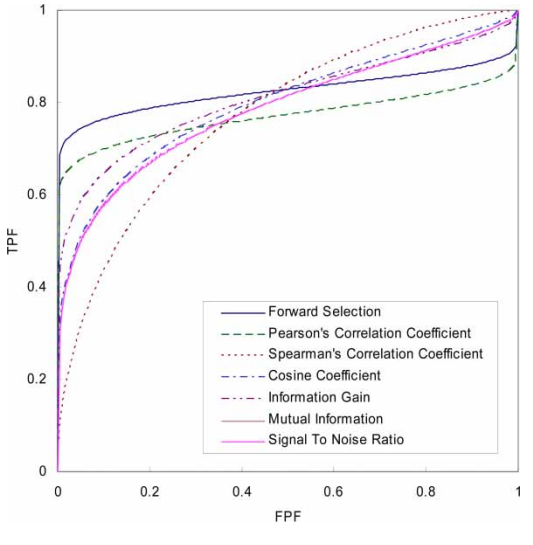

Figure 6. Colon cancer dataset: ROC curves for the seven gene selection methods.

Table 5. Lymphoma cancer dataset: best results (%) of the evaluation measures by seven gene selection methods.

	Forward	Pearson	Spearman	Cosine	IG	MI	S2N
Recognition rate	72.0	68.0	48.0	60.0	64.0	48.0	64.0
Sensitivity	90.9	63.6	54.5	72.7	100.0	63.6	100.0
Specificity	57.2	71.4	42.9	50.0	64.3	64.3	64.3

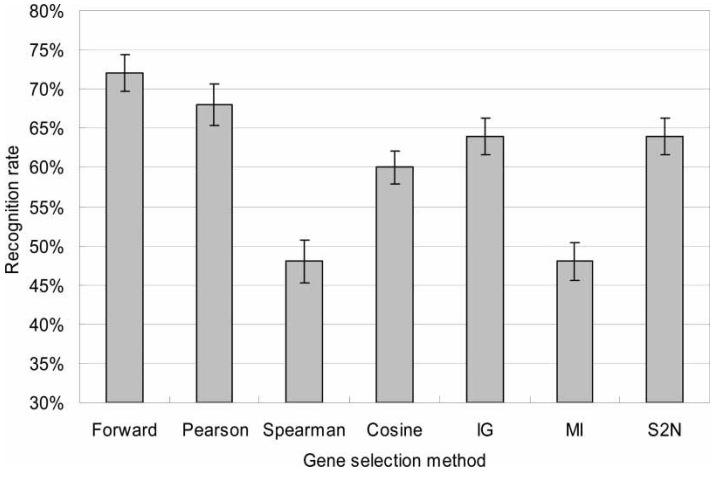

Figure 7. Lymphoma cancer dataset: recognition rates for the gene selection methods.

fraction (the ratio of the number of true positive decisions to the number of actual negative data), respectively. Here, positive means tumour and negative means normal. In the ROC curve, the area under the plotted graph indicates the accuracy. In figure 6, the area of the forward selection method is greatest.

Table 6. Lymphoma cancer dataset: confusion matrix of selected genes by seven gene selection methods.

Forward		P	
		0	1
A	0	8	6
	1	1	10

Pearson		P	
		0	1
A	0	10	4
	1	4	7

Spearman		P	
		0	1
A	0	6	8
	1	6	5

Cosine		P	
		0	1
A	0	7	7
	1	3	8

IG		P	
		0	1
A	0	5	9
	1	0	11

MI		P	
		0	1
A	0	5	9
	1	4	7

S2N		P	
		0	1
A	0	5	9
	1	0	11

P, number of predicted samples; A, number of actual samples.

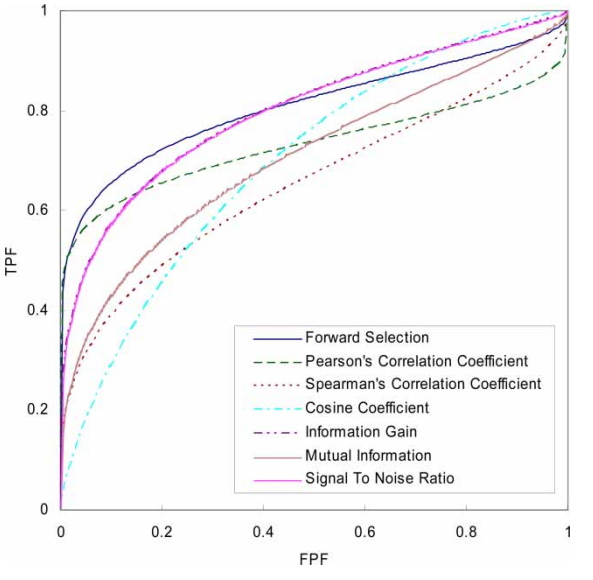

Figure 8. Lymphoma cancer dataset: ROC curves for the seven gene selection methods.

Table 7. Lymphoma cancer dataset: genes selected by the forward selection method.

Rank	Gene number	Gene annotation
1	gene1268	*CD10 = CALLA = Neprilysin = enkepalinase; Clone = 200814
2	gene544	*DRADA2a = dsRNA adenosine deaminase DRADA2a = RNA editing enzyme; Clone = 1326908
3	gene824	(Unknown; Clone = 1370669)
4	gene2313	(Unknown UG Hs.29205 alpha integrin binding protein 63; Clone = 1351211)
5	gene3125	(Unknown UG Hs.137428 ESTs, Highly similar to (defline not available 3249713) [*Homo sapiens*]; Clone = 1234298)
6	gene919	(Unknown UG Hs.117333 *Homo sapiens* mRNA for KIAA1093 protein, partial cds; Clone = 1337623)
7	gene667	(Unknown UG Hs.187585 ESTs; Clone = 825392)
8	gene2406	(Unknown UG Hs.100914 ESTs; Clone = 1335027)
9	gene233	*Unknown UG Hs.136819 ESTs; Clone = 1288950
10	gene3207	*Similar to DNA polymerase beta = DNA alkylation repair protein; Clone = 1358191

4.3 *Results for the lymphoma dataset*

Only 10 genes ($R^2 > 0$) were selected from the lymphoma cancer dataset. The sensitivity, specificity, and recognition rate are shown in table 5. The forward selection method shows the best performance in recognition rate (figure 7). It shows a better performance in sensitivity than all the other methods except information gain and signal to noise ratio, and an average performance in specificity. Considering all measures, the forward selection method, information gain, and signal-to-noise ratio perform better than the other methods. In general the proposed method provides the best performance overall, considering the results of both datasets.

Table 6 provides a confusion matrix of the 10 genes selected from the lymphoma dataset. The forward selection method predicts 10 (10/11) samples as tumours which are actually tumours, and predicts eight (8/14) samples as normal which are actually normal. Compared

Table 8. Lymphoma cancer dataset: R^2, F-value, and significant levels of the genes selected by the forward selection method.

Rank	Gene number	Partial R^2	F-value	Pr > F
1	gene1268	0.8169	89.25	<0.0001
2	gene544	0.1045	25.25	<0.001
3	gene824	0.0577	49.82	<0.0001
4	gene2313	0.0118	22.28	0.0002
5	gene3125	0.0042	14.07	0.0017
6	gene919	0.0026	17.23	0.0009
7	gene667	0.0013	21.23	0.0004
8	gene2406	0.0006	22.58	0.0004
9	gene233	0.0002	11.89	0.0048
10	gene3207	0.0001	61.90	<0.0001

with information gain and signal-to-noise ratio, the forward selection method is superior in terms of recognition rate, but not in terms of sensitivity and specificity.

Figure 8 shows the ROC curves of genes selected by all gene selection methods. As shown in tables 5 and 6, the area under the ROC curve for the forward selection method is greatest, although those of information gain and signal-to-noise ratio are almost as large.

Descriptions of the 10 genes selected from the lymphoma dataset are shown in table 7, and R^2, the F-value, and significant levels of the selected genes are shown in table 8. All the selected genes have a high F-value and a high confidence level. The first selected gene (gene1268) has a very high F-value (0.8169) which means that it gene explains the target (cancer) very well. The second selected gene (gene544) is related to adenosine deaminase and the RNA editing enzyme.

5. Concluding remarks

We have proposed a forward selection method for gene selection for the classification of cancer. The genes selected by the proposed method are able minimize redundant information about cancer. Most other methods are based on one-to-one correlation and do not consider correlations among the selected genes. However, the forward selection method selects genes that are partially correlated among the selected genes, reducing the redundancy in the subset of selected genes.

In the experiments, we used two different microarray datasets to show the utility of the proposed method. The genes selected by the forward selection method have shown its effectiveness in predicting cancer, with a high performance in both datasets. In the colon dataset, 15 genes were identified as crucial in cancer classification. We measured the sensitivity, specificity, and recognition rate of the selected genes using KNN and demonstrated the performance of the proposed method. The genes selected by the forward selection method performed better than genes selected by other selection methods. Only 10 genes were selected in the lymphoma dataset, and the forward selection method performed better than the other methods.

In future work we intend to apply the proposed method to larger datasets and to compare our method with other non-redundant feature selection methods. In addition, studies of the genes selected by the proposed method are required. The biological interpretation of the selected genes is essential for better understanding of the results. Deciding the optimal number of genes is also an interesting topic.

Acknowledgements

This work was supported by the Korea Science and Engineering Foundation (KOSEF) through the Biometric Engineering Research Center (BERC) at Yonsei University, Korea.

References

[1] Eisen, M.B. and Brown, P.O., 1999, DNA arrays for analysis of gene expression. *Methods in Enzymology*, **303**, 179–205.
[2] Harrington, C.A., Rosenow, C. and Retief, J., 2000, Monitoring gene expression using DNA microarrays. *Current Opinion in Microbiology*, **3**, 285–291.
[3] Ben-Dor, A., Bruhn, L., Friedman, N., Nachman, I., Schummer, M. and Yakhini, N., 2000, Tissue classification with gene expression profiles. *Journal of Computational Biology*, **7**, 559–584.
[4] Dudoit, S., Fridlyand, J. and Speed, T.P., 2000, Comparison of discrimination methods for the classification of tumors using gene expression data. Technical Report 576, Department of Statistics, University of California, Berkeley, CA.
[5] Devijver, P. and Kittler, J., 1982, *Pattern Recognition: A Statistical Approach* (Englewood Cliffs, NJ: Prentice-Hall).
[6] Cho, S.-B. and Ryu, J., 2002, Classifying gene expression data of cancer using classifier ensemble with mutually exclusive features. *Proceedings of the IEEE*, **90**, 1744–1753.
[7] Shannon, W.D., Watson, M.A., Perry, A. and Rich, K., 2002, Mantel statistics to correlate gene expression levels from microarrays with clinical covariates. *Genetic Epidemiology*, **23**, 97–96.
[8] Khan, J., Wei, J.S., Ringner, M., Saal, L.H., Ladanyi, M., Westermann, F., Berthold, F., Schwab, M., Antonescu, C.R., Peterson, C. and Meltzer, P.S., 2001, Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. *Nature Medicine*, **7**, 673–679.
[9] Furey, T.S., Cristianini, N., Duffy, N. Bednarski, D.W., Schummer, M. and Haussler, D., 2000, Support vector machine classification and validation of cancer tissue samples using microarray expression data. *Bioinformatics*, **16**, 906–914.
[10] Li, L., Weinberg, C.R., Darden, T.A. and Pedersen, L.G., 2001, Gene selection for sample classification based on gene expression data: study of sensitivity to choice of parameters of the GA/KNN method. *Bioinformatics*, **17**, 1131–1142.
[11] Shamir, R. and Sharan, R., 2001, Algorithmic approaches to clustering gene expression data. In: T. Jiang, T. Smith, Y. Xu and M.Q. Zhang (Eds) *Current Topics in Computational Biology* (Cambridge, MA: MIT Press).
[12] Lipshutz, R.J., Fodor, S.P.A., Gingeras, T.R. and Lockhart, D.J., 1999, High density synthetic oligonucleotide arrays. *Nature Genetics*, **21**, 20–24.
[13] Guyon, I., Weston, J., Barnhill, S. and Vapnik, V., 2002, Gene selection for cancer classification using support vector machines. *Machine Learning*, **46**, 389–422.
[14] Huang, T.M. and Kecman, V., 2005, Gene extraction for cancer diagnosis by support vector machines: an improvement. *Artificial Intelligence in Medicine*, **35**, 617–624.
[15] Fu, L.M. and Youn, E.S., 2003, Improving reliability of gene selection from microarray functional genomics data. *IEEE Transactions on Information Technology in Biomedicine*, **7**, 191–196.
[16] Duan, K.-B., Rajapakse, J.C., Wang, H. and Azuaje, F., 2005, Multiple SVM-RFE for gene selection in cancer classification with expression data. *IEEE Transactions on Nanobioscience*, **4**, 228–234.
[17] Wang, Y., Tetko, I.V., Hall, M.A., Frank, E., Facius, A. Mayer, K.F. and Mewes, H.W., 2005, Gene selection from microarray data for cancer classification: a machine learning approach. *Computational Biology and Chemistry*, **29**, 37–46.
[18] Hall, M.A., 1999, Correlation-based feature selection for machine learning. PhD thesis, Department of Computer Science, University of Waikato, New Zealand.
[19] Golub, T.R., Slonim, D.K. and Tamayo, P., 1999, Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, **186**, 531–537.
[20] Furey, T.S., Cristianini, N., Duffy, N., Bednarski, D.W., Schummer, M. and Haussler, D., 2000, Support vector machine classification and validation of cancer tissue samples using microarray expression data. *Bioinformatics*, **16**, 906–914.
[21] Guyon, I. and Elisseeff, A., 2003, An introduction to variable and feature selection. *Journal of Machine Learning Research*, **3**, 1157–1182.
[22] Nguyen, D.V. and Rocke, D.M., 2002, Tumor classification by partial least squares using microarray gene expression data. *Bioinformatics*, **18**, 39–50.
[23] Xing, E.P., Jordan, M.I. and Karp, R., 2001, Feature selection for high-dimensional genomic microarray data. In: *Proceedings of the 18th International Conference on Machine Learning*, pp. 601–608 (San Francisco, CA: Morgan Kaufman).
[24] Liu, J. and Iba, H., 2001, Selecting informative genes with parallel genetic algorithms in tissue classification. *Genome Informatics*, **12**, 14–23.
[25] Su, Y., Murali, T.M., Pavlovic, V., Schaffer, M. and Kasif, S., 2002, Rankgene: a program to rank genes from expression data. Available online at: http://genomics10.bu.edu/yangsu/rankgene/ (accessed 21 March 2007).

[26] Peng, H., Long, F. and Ding, C., 2005, Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **27**, 1226–1238.
[27] Gilad-Bachrach, R., Navot, A. and Tishby, N., 2004, Margin based feature selection: theory and algorithms. In: *Proceedings of the 21st International Conference on Machine Learning*, pp. 43–50.
[28] Vasconcelos, N., 2003, Feature selection by maximum marginal diversity: optimality and implications for visual recognition. In: *Proceedings of the 2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 762–769.
[29] West, M., Nevins, J.R., Marks, J.R., Spang, C., Blanchette, C. and Zuzan, H., 2000, DNA microarray data analysis and regression modeling for genetic expression profiling. *ISDS Discussion*, 00–15.