SELECTING INFORMATIVE GENES OF LUNG CANCERS BY A COMBINATION OF HYBRID METHODS

Mohd Saberi Mohamad[1,2], Sigeru Omatu[1], Safaai Deris[2] and Michifuci Yoshioka[1]

[1]Department of Computer Science and Intelligent Systems,
Graduate School of Engineering, Osaka Prefecture University,
Sakai, Osaka 599-8531, Japan

[2]Department of Software Engineering,
Faculty of Computer Science and Information System,
Universiti Teknologi Malaysia, 81310 Skudai, Johor, Malaysia

Email: mohd.saberi@sig.cs.osakafu-u.ac.jp, {sigeru,yoshioka}@.cs.osakafu-u.ac.jp,
safaai@utm.my

Abstract: Gene expression technology namely microarray, offers the ability to measure the expression levels of thousands of genes simultaneously in biological organisms. Microarray data are expected to be of significant help in the development of efficient cancer diagnosis and classification platform. A major problem in these data is that the number of genes greatly exceeds the number of tissue samples. These data have also noisy genes. It has been shown from literature reviews that selecting a small subset of informative genes can lead to an improved classification accuracy. Thus, this paper aims to select a small subset of informative genes that are most relevant for the cancer classification. To achieve this aim, an approach that involved two hybrid methods has been proposed. This approach is assessed and evaluated on one well-known microarray data set, namely the lung cancer, showing competitive results.

Keywords: Cancer Classification, Genetic Algorithm, Gene Selection, Hybrid Method, Microarray Data.

1. INTRODUCTION

The traditional cancer diagnosis relies on a complex and inexact combination of clinical and histopathological data. This classic approach may fail when dealing with atypical tumours or morphologically indistinguishable tumour subtypes. Advances in the area of microarray-based expression analysis have led to the promise of cancer diagnosis using new molecular-based approaches [9]. A microarray machine is used to measure the expression levels of thousands of genes simultaneously in a cell mixture, and finally it produces microarray data.

SELECTING INFORMATIVE GENES OF LUNG CANCERS BY A COMBINATION OF HYBRID METHODS

Mohd Saberi Mohamad[1,2], Sigeru Omatu[1], Safaai Deris[2] and Michifuci Yoshioka[1]

[1]Department of Computer Science and Intelligent Systems,
Graduate School of Engineering, Osaka Prefecture University,
Sakai, Osaka 599-8531, Japan

[2]Department of Software Engineering,
Faculty of Computer Science and Information System,
Universiti Teknologi Malaysia, 81310 Skudai, Johor, Malaysia

Email: mohd.saberi@sig.cs.osakafu-u.ac.jp, {sigeru,yoshioka}@.cs.osakafu-u.ac.jp,
safaai@utm.my

Abstract: Gene expression technology namely microarray, offers the ability to measure the expression levels of thousands of genes simultaneously in biological organisms. Microarray data are expected to be of significant help in the development of efficient cancer diagnosis and classification platform. A major problem in these data is that the number of genes greatly exceeds the number of tissue samples. These data have also noisy genes. It has been shown from literature reviews that selecting a small subset of informative genes can lead to an improved classification accuracy. Thus, this paper aims to select a small subset of informative genes that are most relevant for the cancer classification. To achieve this aim, an approach that involved two hybrid methods has been proposed. This approach is assessed and evaluated on one well-known microarray data set, namely the lung cancer, showing competitive results.

Keywords: Cancer Classification, Genetic Algorithm, Gene Selection, Hybrid Method, Microarray Data.

1. INTRODUCTION

The traditional cancer diagnosis relies on a complex and inexact combination of clinical and histopathological data. This classic approach may fail when dealing with atypical tumours or morphologically indistinguishable tumour subtypes. Advances in the area of microarray-based expression analysis have led to the promise of cancer diagnosis using new molecular-based approaches [9]. A microarray machine is used to measure the expression levels of thousands of genes simultaneously in a cell mixture, and finally it produces microarray data.

The task of cancer classification using microarray data is to classify tissue samples into related classes of phenotypes, e.g., cancer versus normal [5].

Given N tissue samples and expression of M genes, microarray data are stored in a matrix as shown in Figure 1. Cancer classification using these data poses a major challenge because of the following characteristics:

- $M >> N$. M is in the range of 2,000-20,000, while N is in the range of 30-200.
- Most genes are not relevant for classifying different tissue types.
- These data have a noisy nature.

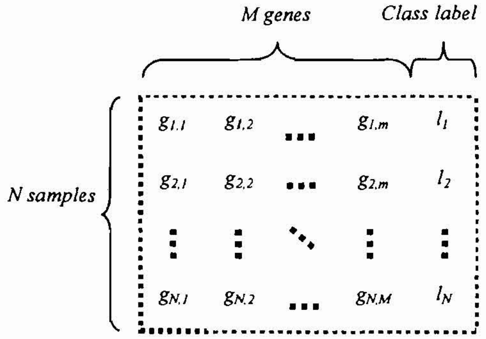
M genes
Class label
N samples
$g_{1,1}$ $g_{1,2}$ ... $g_{1,m}$ l_1
$g_{2,1}$ $g_{2,2}$ ... $g_{2,m}$ l_2
$g_{N,1}$ $g_{N,2}$ ... $g_{N,M}$ l_N

Figure 1. The matrix of microarray data $G_{N\times(M+1)}$. $g_{i,j}$ is a numeric value representing the gene expression level of the jth gene in the ith sample. l_i in the last column is the class label for the ith sample.

To overcome the challenge, a gene selection approach is usually used to select a small subset of informative genes that maximises the classifier's ability to classify samples accurately [5]. This approach has several advantages:

- It can maintain or improve classification accuracy.
- It can reduce the dimensionality of data.
- It can remove noisy genes.

Gene selection methods can be classified into two categories. If gene selection is carried out independently from the classification procedure, the method belongs to the filter approach. Otherwise, it is said to follow a hybrid approach. Most previous works have used the filter approach to select genes since it is computationally more efficient than the hybrid approach. However, the hybrid approach usually provides greater accuracy than the filter approach [4]. In this paper, an approach that involves two hybrid methods is proposed to select a small subset of informative genes for cancer classification.

```
Step 1: Select a number of genes and produce initial populations with each chromosome is represented by an integer string.
Step 2: Evaluate each individual (chromosome) in each population using a fitness function.
    Step 2.1: Sort integer values in a chromosome.
    Step 2.2: Select genes based on position of the integer values in a chromosome (e.g: if integer value=10, then select 10th gene).
    Step 2.3: Store the selected genes into a subset.
    Step 2.4: $fitness(x) = w_1 \times A(x) + (w_2(M - R(x)) / M)$
Step 3: GA operates on the populations to evolve the best solution (a subset of selected genes) until the final generation.
    Step 3.1: Apply a selection strategy and GA operators (crossover and mutation).
    Step 3.2: Repeat Step 2.
Step 4: Return a subset of genes (the highest fitness).
Step 5: Get the total number of genes from the subset of genes that is produced by Step 4, and produce new initial populations with each chromosome is represented by a bit (0 and 1) string.
Step 6: Evaluate each chromosome in each population using a fitness function.
    Step 6.1: Select genes based on bit values in a chromosome (bit 1=select; bit 0=unselect).
    Step 6.2: Store the selected genes into a subset.
    Step 6.3: $fitness(x) = w_1 \times A(x) + (w_2(M - R(x)) / M)$
Step 7: GA operates on the populations to evolve the best solution (the best subset of genes) until the final generation.
    Step 7.1: Apply a selection strategy and GA operators (crossover and mutation).
    Step 7.2: Repeat Step 6.
Step 8: Return the optimal subset of genes.
Step 9: Classify the optimal subset using an SVM classifier.
```

Figure 2. The algorithm of GASVM-II+GASVM.

2. THE PROPOSED APPROACH

Mohamad *et al.* have reported that a hybrid of genetic algorithms and support vector machines (GASVM), and an improved GASVM (NewGASVM) have several advantages and disadvantages [4]. In this paper, NewGASVM is called GASVM-II. All information of GASVM and GASVM-II are available in Mohamad *et al* [4]. The advantage of GASVM is that it can automatically select and optimise a number of genes to produce a gene subset.

However, it performs poorly in high-dimensional data. In contrast, GASVM-II performs well in the high-dimensional data. It can also reduce the complexity of search spaces and maybe able to evaluate all possible subsets of genes. Nevertheless, the drawback of GASVM-II is that it selects a number of genes manually to yield a gene subset.

As a result, this paper proposes an approach using two hybrid methods for selecting informative genes. This approach is called as GASVM-II+GASVM. It is developed to improve the performances of GASVM and GASVM-II. Figure 2 shows that the algorithm of GASVM-II+GASVM involving two stages. In the first stage, GASVM-II is applied to manually select genes from the overall microarray data to produce a subset of genes. It is used to reduce the dimensionality of the data, and therefore the complexity of the searches or solution spaces can also be decreased.

In the second stage, GASVM is used to select and optimise a small subset of informative genes from the subset that is produced by the first stage. If the size of the subset is small and the combination of genes is not very complex, GASVM can easily find and optimise the subset. GASVM is applied because it can automatically select a number of genes and finally produce an optimised gene subset. This second stage can also remove noisy genes because the first step has reduced the size and complexity of the search spaces.

Therefore, this proposed approach has totally excluded the test samples from the classifier building process in order to avoid the influence of selection bias [1]. The fitness of an individual is calculated as follows:

$$fitness(x) = w_1 \times A(x) + (w_2(M - R(x)) / M) \qquad (1)$$

in which $A(x) \in [0,1]$ is the leave-one-out-cross-validation (LOOCV) accuracy on the training data using the only expression values of the selected genes in a subset x, $R(x)$ is the number of selected genes in x. M is the total number of genes. w_1 and w_2 are two priority weights corresponding to the importance of accuracy and the number of selected genes, where $w_1 \in [0.1, 0.9]$ and $w_2 = 1 - w_1$. In this paper, the accuracy is more important than the number of selected genes. Hence, w_1 and w_2 are set to 0.7 and 0.3 respectively for the lung cancer data set. These values are based on experimental results in Mohamad *et al.*'s paper [6].

3. EXPERIMENTAL RESULTS

3.1 Data Sets

The lung cancer microarray data set is used to evaluate the proposed algorithm. This data set has two classes: malignant pleural mesothelioma (MPM) and adenocarcinoma (ADCA).

There are 181 samples (31 MPM and 150 ADCA), originally analysed by Gordon *et al.* [2]. The training set contains 32 of them (16 MPM and 16 ADCA). The rest 149 samples are used for the test set. Each sample is described by 12,533 genes. It can be obtained at http://chestsurg.org/publications/2002-microarray.aspx.

3.2 Experimental setup

Three criteria following its importance are considered to evaluate the performances of the proposed approach: the test accuracy, the LOOCV accuracy, and the number of selècted genes.

The experimental results presented in this section pursue two objectives. The first objective is to show that gene selection using GASVM-II+GASVM is needed for better classification of microarray data. The second objective is to show that it is better than GASVMs (single-objective and multi-objective) and GASVM-II. To achieve these objectives, several experiments are conducted on the proposed approach 10 times on each data set. In the first stage, it is experimented by using different number of pre-selected genes (10, 20, 30,..., 600). Furthermore, in the second stage, GASVM chooses a number of the final selected genes automatically. Lastly, it produces an optimised gene subset that contains the final selected genes. The subset that produces the highest LOOCV accuracy with the possible least number of selected genes is chosen as the best subset. SVM classifier, GASVMs, and GASVM-II are also experimented for comparison with GASVM-II+GASVM.

3.3 Result analysis and discussion

In this paper, a value of the form x ± y represents average value x with standard deviation y. Furthermore, #Pre-Selected Genes, #Final Selected Genes, and Run# represent the number of pre-selected genes, the number of the final selected genes, and a run number, respectively.

Figure 3 shows that the highest averages of LOOCV and test accuracies are 100% and 94.16.88%, respectively. Only 2.1 average genes were finally selected to obtain the highest average of the accuracies of the data set. Almost all the different numbers of pre-selected genes and the final selected genes have obtained 100% LOOCV accuracy, and this has proven that the proposed approach search and select the optimal solution (the best gene subset) in the solution space successfully. However, the test accuracy was much lower than the LOOCV accuracy due to over-fitting problem. This problem happened because of the number of training samples is smaller than the number of test samples, and many expression values of the test samples may be different from those of the training samples.

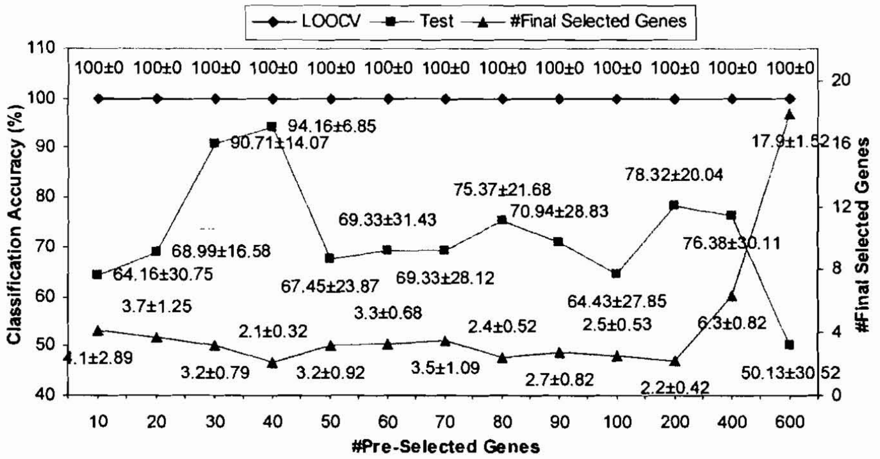

Figure 3. A relation between classification accuracies and the numbers of selected genes (#Pre-selected genes and #Final selected genes) on the lung data set (10 runs on average).

Table 1 shows that the best performances (LOOCV and test accuracies) of the proposed approach in the best subsets were 100% and 98.66%, respectively for the lung cancer data set using the only two genes. The best performances have been found in the second, sixth, seventh, eighth, ninth, and tenth runs.

Table 1. The result of the best gene subsets in 10 runs.

Data set	#Pre-Selected Genes	LOOCV (%)	Test (%)	#Final Selected Genes	Run#
Lung	40	100	98.66	2	2,6,7,8,9,10

The selected genes in the best gene subsets as founded by GASVM-II+GASVM in Table 1 are shown in Table 2. The probe-set name, gene description, and gene accession number of the selected informative genes are also given. Interestingly, all of the best gene subsets have similar type of genes. From this finding, the existence of some kinds of relations among the two selected genes of Lung data set is noted (gene description). Based on graph in Figure 3, different number of selected genes in a subset has produced dissimilar test accuracy. Thus, GASVM-II+GASVM preserves the interactions among the genes that result in the best classification accuracy by using only two genes of the data set. These genes among thousand of genes may be excellent candidate for medical investigations. Biologists can save much time since they can directly refer to the genes that have higher possibility to be useful for cancer diagnosis and drug target.

The benchmark of the proposed approach comparing with GASVM-II, GASVMs (single-objective and multi-objective), and SVM is summarised in Table 3. The LOOCV accuracy, test accuracy, and number of selected genes are written in the parenthesis; the first and second parts are average result and showcased the best result, respectively. In the table, GASVM-II+GASVM has outperformed GASVM-II, GASVMs, and SVM in terms of the LOOCV accuracy, test accuracy, and number of selected genes on average results and the best results. Generally, GASVM-II was better than GASVMs and SVM. A smaller size gene subset that is produced by the GASVM-II+GASVM results in higher classification accuracy. Hence, it may provide more insights into the molecular classification and diagnosis of cancers. This suggests that gene selection using the proposed approach is needed for cancer classification of microarray data.

Table 2. A list of informative genes in the best gene subsets.

Data Set	Run#	Probe-set Name	Gene Accession Number	Gene Description
Lung	2, 6, 7, 8, 9, 10	34320_at	AL050224	PTRF: polymerase I and transcript release factor
	2, 6, 7, 8, 9, 10	41286_at	X77753	TACSTD2: tumor-associated calcium signal transducer 2

Table 3. The benchmark of GASVM-II+GASVM with GASVMs, GASVM-II, and SVM.

Method	Lung Data Set (Average; The Best)		
	#Final Selected Genes	Accuracy (%)	
		LOOCV	Test
GASVM-II+GASVM	(2.1 ± 0.32; 2)	(100 ± 0; 100)	(94.16 ± 6.85; 98.66)
GASVM-II	(10 ± 0; 10)	(100 ± 0; 100)	(59.33 ± 29.32; 97.32)
GASVM (multi-objective)	(4,418.5 ± 50.19; 4,433)	(75.31 ± 0.99; 78.13)	(85.84 ± 3.97; 93.29)
GASVM (single-objective)	(6,267.8 ± 56.34; 6,342)	(75 ± 0; 75)	(84.77 ± 2.53; 87.92)
SVM classifier	(12,533 ± 0; 12,533)	(65.63 ± 0; 65.63)	(85.91 ± 0; 85.91)

Note: The best result shown in shaded cells.

Table 4. The benchmark of GASVM-II+GASVM with previous works.

Author [Reference]	Lung Data Set		
	#Final Selected Genes	Accuracy (%)	
		LOOCV	Test
Our work	2	100	98.66
Shah and Kusiak [7]	8	100	98.66
Gordon *et al.* [2]	4	-	97.32
Wang [8]	-	99.45	-
Li *et al.* [3]	-	-	97.99

Note: The best result shown in shaded cells. '–' means that result is not available.

Table 4 displays benchmark of the best results of this work and previous related works on the lung cancer data set. The best result of the proposed approach was obtained from the best subset in Table 1. Based on LOOCV and test accuracies, it was noted that the best results (100% LOOCV accuracy and 98.66% test accuracy) from this work were equal to the best result from current previous work [7]. However, this work only uses two genes to achieve the accuracies, whereas eight genes were used in the work of Shah and Kusiak [7]. Hence, GASVM-II+GASVM outperformed the previous works. The first original work [2] achieved only 97.32% test accuracy by using 4 genes, while 97.99% of the further previous work [3].

5. CONCLUSIONS

In this paper, an approach (GASVM-II+GASVM) that involved two hybrid methods has been proposed, developed, and analysed for gene selection and classification. This research found many combinations of gene subsets that were not equal number of genes have produced the different classification accuracy. This finding suggests that there are many irrelevant and noisy genes in microarray data. In addition, the performances of the GASVM-II+GASVM were superior to the GASVM-II, GASVMs, and SVM. Focusing attention on a smaller subset of genes is useful not only because it produces good classification accuracy, but also since informative genes in this subset may provide insights into the mechanisms responsible for the cancer itself.

It can also be applied in other applications such as robotics, computer intrusion detections, and computer graphics. Even though the proposed approach has classified tumours with higher accuracy, it is still can not avoid the over-fitting problem. A recursive genetic algorithm is currently studied to better select a small subset of genes for cancer classification.

ACKNOWLEDGEMENTS

This study was supported and approved by Universiti Teknologi Malaysia, Osaka Prefecture University, and Malaysian Ministry of Higher Education. The authors gratefully thank the referees for the helpful suggestions.

REFERENCES

[1] Ambroise, C. and McLachlan, G. J., "Selection bias in gene extraction on the basis of microarray gene-expression data", Proceedings of the National Academy of Science of the USA, Washington, USA, Volume 99, Issue 10, pp 6562–6566, 2002.

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

[3] Li, J., Liu, H., Ng, S. K. and Wong, L., "Discovery of significant rules for classifying cancer diagnosis data", Bioinformatics, Volume 19, pp.93–102, 2003.

[4] Mohamad, M. S., Deris, S. and Illias, R. M., "A hybrid of genetic algorithm and support vector machine for features selection and classification of gene expression microarray", International Journal of Computational Intelligence and Applications, Volume 5, pp.91–107, 2005.

[5] Mohamad, M. S., Omatu, S., Deris, S. and Hashim, S. Z. M., "A model for gene selection and classification of gene expression data", International Journal of Artificial Life & Robotics, Volume 11, Issue 2, pp.219–222, 2007.

[6] Mohamad, M. S., Omatu, S., Deris, S. and Misman, M. F., "A multi-objective strategy in genetic algorithm for gene selection of gene expression data", International Journal of Artificial Life & Robotics, Volume 13, Issue 2, 2008.

[7] Shah, S. and Kusiak, A., "Cancer gene search with data-mining and genetic algorithms", Computers in Biology and Medicine, Volume 37, Issue 2, pp.251–261, 2007.

[8] Wang, C. W., "New ensemble machine learning method for classification and prediction on gene expression data", Proceedings of the 28th IEEE Engineering in Medicine and Biology Society, IEEE Press, pp.3478–3481, 2006.

[9] Wang, L., Chu, F. and Xie, W., "Accurate cancer classification using expressions of very few genes", IEEE/ACM Transaction on Computational Biology & Bioinformatics Volume 4, Issue 1, pp.40–53, 2007.