Selecting Informative Genes from Microarray Data by using a Hybrid Algorithm for Cancer Classification

M.S. Mohamad[1] S. Omatu[1] S. Deris[2] and M.F. Misman[2]

[1]*Department of Computer Science and Intelligent Systems, Graduate School of Engineering, Osaka Prefecture University, Sakai, Osaka 599-8531, Japan*
(Tel : 81-72-254-9278; Fax : 81-72-257-1788)
(mohd.saberi@sig.cs.osakafu-u.ac.jp; omatu@cs.osakafu-u.ac.jp)

[2]*Department of Software Engineering, Faculty of Computer Science and Information Systems, Universiti Teknologi Malaysia, 81310 Skudai, Johore, Malaysia*
(Tel : 60-7-553-7784; Fax : 60-7-556-5044)
(safaai@utm.my; faizmisman@gmail.com)

Abstract: The development of microarray-based high-throughput gene expression has led to the hope that this technology could provide an efficient cancer diagnosis and classification platform. A major problem in these gene expression data is that the number of genes greatly exceeds the number of tissue samples. Moreover, these data have a noisy nature. It has been shown that selecting a small subset of informative genes can lead to improved classification accuracy. Thus, this paper aims to select a small subset of informative genes that are most relevant for the classification task. To achieve this aim, a hybrid algorithm that combines two hybrid methods has been developed. This algorithm is assessed on two well-known microarray data sets, showing competitive results.

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

I. 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.[1] Microarray is a device used to measure the expression level 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 such as cancer versus normal.[2]

Given N tissue samples and expression of M genes, the data is stored in a N x (M+1) matrix as shown in Figure 1. Cancer classification using gene expression data poses a major challenge because of the following characteristics:

- M >> N. For typical data sets, M is in the range of 2000-20000, while N is in the range of 30-200.
- Most genes are not relevant to the classification of different tissue types.
- These data have a noisy nature.

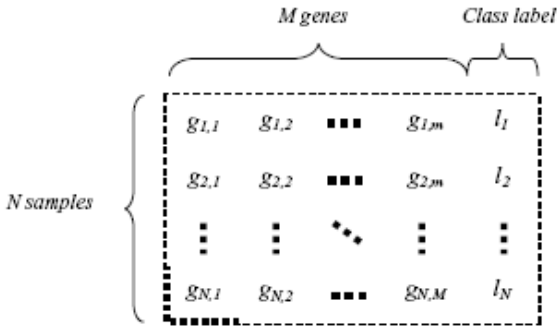

Fig.1. The data matrix. $g_{i,j}$ is a numeric value presenting the gene expression level of gene j in sample i. l_i in the last column is the class label for sample i

To overcome the problems, gene selection approach is used to select a small subset of informative genes that maximises the classifier's ability to classify samples accurately.[2] Gene selection has several advantages:

- It can improve classification accuracy.
- It can reduce the cost in a clinical setting. At the moment, expression arrays with thousands of genes are impractical to be employed in medical laboratories.
- It could enable biologists to gain significant insight into the genetic nature of the disease and the mechanisms responsible for it. This would assist in drug discovery and early diagnosis.

Gene selection method 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.[3] The application of hybrid approach using genetic algorithm (GA) with a classifier has grown in recent years. From the previous works, the GA performed well but only on the data that have number of features less than 1000.[3]

In this paper, gene selection problem is focused on and a hybrid algorithm is proposed to select a small subset of informative genes for cancer classification. This algorithm is developed to improve the performance of a hybrid of GA and support vector machine classifier (GASVM) for genes selection and classification.[3]

II. PROPOSED HYBRID ALGORITHM

Mohamad et al. reported that GASVM and NewGASVM have several advantages and disadvantages.[3] 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, NewGASVM performs well in the high dimensional data. Therefore, it can reduce the complexity of search space and maybe able to evaluate all possible subsets of genes. However, the drawback of NewGASVM is that it manually selects a number of genes to yield a gene subset.

As a result, this paper proposes a hybrid algorithm based on GASVM and NewGASVM for selecting informative genes for cancer classification. Figure 2 shows that this proposed hybrid algorithm sequentially combines NewGASVM and GASVM. In the first stage, NewGASVM is applied to manually select genes from the overall gene expression to produce a subset of genes data. It is used to reduce the dimensionality of the data, and therefore the complexity of the search or solution space can also be decreased. It also performs well in higher dimensional data.

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

Step 1: Select a number of genes and produce initial populations with each chromosome represented by integer string.
Step 2: Evaluate each individual (chromosome) in a population using a fitness function.
  Step 2.1: Selecting genes based on position of the integer values:
    Step 2.1.1: Save the integer values from chromosomes to array.
    Step 2.1.2: Sort the integer values.
    Step 2.1.3: Store the selected genes based on the integer values (e.g: if integer value=10, then select 10th gene).
  Step 2.2 Evaluate each individual (chromosome) using SVM classifier.
Step 3: GA operates on the population to evolve the best solution (a subset of selected genes) until the final generation.
  Step 3.1: Apply the 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 produces by Step 4 and produce new initial populations with each chromosome represented by bits (0 and 1) string.
Step 6: Evaluate each individual (chromosome) in a population using a fitness function.
  Step 6.1: Selecting genes based on bit values (bit 1=select; bit 0=unselect):
    Step 6.1.1: Save bits from chromosomes to array.
    Step 6.1.2: Store the selected genes based on the position of bit 1.
  Step 6.2: Evaluate each individual (chromosome) using SVM classifier.
Step 7: GA operates on the population to evolve the best solution (the best subset of genes) until the final generation.
  Step 7.1: Repeat **Step 3.1**.
  Step 7.2: Repeat **Step 6**.
Step 8: Return the best subset of genes.
Step 9: Classify the best subset using SVM classifier.

Fig.2. A Proposed Hybrid Algorithm

Ambroise and Mclachlan (2002) indicated that testing results could be overoptimistic, caused by the "selection bias" if the testing samples were not excluded from the classifier building process.[4] Therefore, the proposed hybrid algorithm totally

excluded the testing samples from the classifier building process in order to avoid the influence of bias.

The fitness of an individual is calculated as follows:

$$fitness(x) = w1 \times A(x) + w2((M - R(x))/M) \quad (1)$$

in which $A(x) \in [0,1]$ is the leave-one-out-cross-validation (LOOCV) accuracy on training data using only the expression values of the selected genes in subset x, $R(x)$ is the number of selected genes in x. M is the total number of genes. $w1$ and $w2$ are two weights corresponding to the importance of accuracy and the number of selected genes, $w1 \in [0,0.9]$ and $w2 = 1 - w1$. In this paper, accuracy is more important than number of selected genes. Hence, w1 and w2 are set to 0.8 and 0.2 respectively for Leukemia data set, and 0.7 and 0.3 respectively for Colon data set. These values are based on experimental results in Mohamad et al.'s paper.[5]

III. EXPERIMENTAL RESULTS

3.1. Data Sets

Two benchmark data sets are used to evaluate the proposed algorithm: Leukemia cancer and Colon cancer. Leukemia cancer data set contains examples of human acute leukemia. It can be obtained at http://www.broad.mit.edu/cgi-bin/cancer/datasets.cgi, while Colon cancer data set can be downloaded at http://microarray.princeton.edu/oncology/. For Leukemia cancer data set, LOOCV procedure is applied on training data, and accuracy test measurement on testing data to measure classification accuracy. While for Colon cancer data set, only LOOCV procedure is used because this data set only has training data.

3.2. Experimental Setup

Three criteria following its important are considered to evaluate the performances of the proposed hybrid algorithm: test accuracy, LOOCV accuracy and number of selected genes.

The experimental results presented in this section pursue two objectives. The first objective is to show that gene selection using the proposed hybrid algorithm is needed for better classification of microarray data. Furthermore, the second objective is to show that the proposed hybrid algorithm is better than the original version of GASVM. To achieve these objectives, several experiments are conducted five times for both data sets by using different number of selected genes (10, 20, 30,…, 600). Moreover, SVM and GASVM are also performed in this research for comparison with the proposed algorithm. GASVM has used the fitness function such as in the proposed algorithm.

3.3. Result Analysis and Discussion

Table 1 shows that the highest averages of LOOCV and test accuracies for classifying Leukemia cancer samples are 100% and 90.59% respectively, while 99.68% LOOCV accuracy for Colon data set. In this table, a value of the form $x \pm y$ represents average value x with standard deviation y.

Table 1. Classification accuracies for different gene subsets using the hybrid algorithm (5 runs on average)

Number of Pre-Selected Genes	Average for Leukemia Data Set			Average for Colon Data Set	
	LOOCV %	Test %	Number of Final Selected Genes	LOOCV %	Number of Final Selected Genes
600	100 ± 0	75.88 ± 7.32	35 ± 1.58	98.39 ± 0	59.6 ± 5.32
400	100 ± 0	74.71 ± 7.67	12 ± 2.35	99.68 ± 0.72	27.2 ± 4.44
200	100 ± 0	72.94 ± 6.71	3.2 ± 0.84	99.36 ± 0.88	10.6 ± 2.70
100	100 ± 0	69.41 ± 7.95	3.2 ± 0.84	97.42 ± 4.92	9.2 ± 0.84
90	100 ± 0	76.47 ± 6.24	4.2 ± 1.30	99.03 ± 0.88	11 ± 1.23
80	100 ± 0	71.77 ± 11.12	4.2 ± 1.10	99.03 ± 0.88	10.8 ± 1.48
70	100 ± 0	67.06 ± 12.20	4.6 ± 1.14	99.68 ± 0.72	11.2 ± 1.30
60	100 ± 0	77.06 ± 6.71	4.8 ± 1.30	99.03 ± 0.88	9.6 ± 1.67
50	100 ± 0	74.71 ± 6.44	5 ± 1.22	95.16 ± 5.59	9.4 ± 1.34
40	100 ± 0	60.59 ± 10.73	4 ± 1.58	96.77 ± 4.70	10 ± 0.71
30	100 ± 0	74.71 ± 6.78	5.2 ± 1.92	94.52 ± 6.10	7.6 ± 1.52
20	100 ± 0	74.71 ± 12.58	3.6 ± 1.14	89.36 ± 8.11	4.4 ± 1.14
10	100 ± 0	90.59 ± 8.92	2.4 ± 0.55	87.74 ± 6.40	3.2 ± 0.84

Note: Result of the best subsets shown in shaded cells.

Only 2.4 genes were finally selected to obtain the highest accuracies of Leukemia data set, whereas either 11.2 genes or 27.2 genes were of Colon dataset. Overall, LOOCV accuracy in Leukemia data set was much higher than test accuracy due to overfitting of these data sets. Overfitting is a major problem in classification of microarray data when LOOCV accuracy is higher than test accuracy. This problem happened because of the number of training samples is smaller than the number of test samples, and many expression values of test samples may be different from those of the training samples.

Table 2 shows the best performances (LOOCV and test accuracies) were 100% and 97.06% respectively for Leukemia data set using only two genes. For Colon data set, the highest LOOCV accuracy was 100% using 9, 11 or 12 genes. The best performances for Leukemia data set have been found in the second, fourth and fifth experiments, while for Colon data set, the best performances are found in the first, third, fourth and fifth experiments.

Table 2. Results of the best gene subset in 5 runs

Data set	Pre-Selected Genes	LOOCV %	Test %	Final Selected Genes
Leukemia	10	100	97.06	2
Colon	70	100	-	9,11,12

The benchmark of the proposed hybrid algorithm comparing with GASVM method and SVM classifier 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 and showcased the best results respectively. In the table, the algorithm has outperformed GASVM and SVM in terms of LOOCV accuracy, test accuracy and number of selected genes on average result and the best result. A smaller size gene subset that is produced by the proposed hybrid algorithm results in higher classification accuracy hence may provide more insights into the molecular classification and diagnosis of cancers. This suggests that gene selection is needed for cancer classification of microarray data.

IV. CONCLUSION

In this paper, a hybrid algorithm is designed, developed and analysed for gene selection and classification on two microrray data sets. This research found many combinations of gene subsets that are not equal number of genes as produced in the same classification accuracy. These findings suggest that there are many redundant or noisy genes in microarray data and some of them act negatively on the acquired accuracy by the relevant genes. From the experimental results, the performances of the proposed hybrid algorithm were superior to the GASVM and SVM. This is due to the fact that NewGASVM and GASVM are combined into the algorithm. NewGASVM is applied to reduce the dimensionality of the data, while GASVM is used to automatically select and optimise a small subset of genes from the subset that is produced by NewGASVM for removing irrelevant and noisy genes.

It can also be applied in other applications such as robotics, computer intrusion detection and computer graphics. Even though this proposed algorithm has classified tumours with higher accuracy, it is still can not avoid overfitting problem. Recursive algorithm in hybrid method is currently studied to better select a small subset of genes for cancer classification.

REFERENCES

[1] Wang L, Chu F, Xie W (2007) Accurate cancer classification using expressions of very few genes. IEEE/ACM Trans Comput Bio Bioinf 4(1):40–53
[2] Mohamad MS, Omatu S, Deris S, et al. (2007) A model for gene selection and classification of gene expression data. Artif Life Robotics 11(2):219–222
[3] Mohamad MS, Deris S, Illias RM (2005) A hybrid of genetic algorithm and support vector machine for features selection and classification of gene expression microarray. J Comput Intell Appl. 5: 1–17
[4] Ambroise C, McLachlan GJ (2002) Selection bias in gene extraction on the basis of microarray gene-expression data. Proc Nat Aca Sci USA 6562–6566
[5] Mohamad MS, Omatu S, Deris S, et al. (Accepted) A multi-objective strategy in genetic algorithm for gene selection of gene expression data. Symp Artif Life Robotics

Table 3. Benchmark of proposed hybrid algorithm with GASVM and SVM on each data set

Method	Leukemia Data Set (Average, The Best)			Colon Data Set (Average, The Best)	
	Number of Selected Genes	Accuracy (%)		Number of Selected Genes	LOOCV Accuracy (%)
		LOOCV	Test		
Proposed Hybrid Algorithm	(2.4, 2)	(100, 100)	(90.59, 97.06)	(11.2, 9)	(99.68, 100)
GASVM (multi-objective)	(2225.4, 2199)	(95.79, 97.37)	(85.88, 88.24)	(455.2, 438)	(93.55, 93.55)
SVM	(7129, 7129)	(94.74, 94.74)	(85.29, 85.29)	(2000, 2000)	(85.48, 85.48)

Note: Best result shown in shaded cells.