The Fourteenth International Symposium on Artificial Life and Robotics 2009 (AROB 14th '09),
B-Con Plaza, Beppu, Oita, Japan, February 5 - 7, 2009

Gene Subset Selection Using an Iterative Approach Based on Genetic Algorithms

M.S. Mohamad[1,2] S. Omatu[1] S. Deris[2] and M. Yoshioka[1]

[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; yoshioka@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)

Abstract: Microarray data are expected to be useful for cancer classification. The main problem that needs to be addressed is the selection of a smaller subset of genes from the thousands of genes in the data that contributes to a disease. This selection process is difficult due to many irrelevant genes, noisy genes, and the availability of a small number of samples compared to a huge number of genes (higher-dimensional data). Hence, this paper aims to select a near-optimal (smaller) subset of informative genes that is most relevant for the cancer classification. To achieve the aim, an iterative approach based on genetic algorithms has been proposed. Experimental results show that the performance of the proposed approach is superior to other related previous works as well as four methods experimented in this work. In addition a list of informative genes in the best gene subsets is also presented for biological usage.

Keywords: Gene selection, Genetic algorithm, Iterative approach, Microarray data**.**

I. INTRODUCTION

Advances in the area of microarray-based gene expression analyses have led to a promising future of cancer diagnosis using new molecular-based approaches. This microarray technology is used to measure the expression levels of thousands of genes simultaneously, and finally produce microarray data. A comparison between the gene expression levels of cancerous and normal tissues can also be done. This comparison is useful to select those genes that might anticipate the clinical behaviour of cancers. Thus, there is a need to select informative genes that contribute to a cancerous state. However, the gene selection process poses a major challenge because of the characteristics of microarray data: the huge number of genes compared to the small number of samples (higher-dimensional data), irrelevant genes, and noisy data.

To overcome the challenge, a gene selection method is used to select a subset of genes that increases the classifier's ability to classify samples more accurately. The gene selection method has several advantages such as improving classification accuracy, reducing the dimensionality of data, and removing irrelevant and noisy genes.

There are two types of gene selection methods: [1,2] if a gene selection method is carried out independently from a classifier, it belongs to the filter approach; otherwise, it is said to follow a hybrid (wrapper) approach. In the early era of microarray analysis, most previous works have used the filter approach to select genes because it is computationally more efficient than the hybrid approach. However, the hybrid approach usually provides greater accuracy than the filter approach since the genes are selected by considering and optimising relations among genes.[3] Until now, several hybrid methods, especially a combination between a genetic algorithm (GA) and a support vector machine (SVM) classifier (GASVM), have been implemented to select informative genes.[1,2,4,5] The drawbacks of the hybrid methods (GASVM-based methods) in the previous works are:[1,2,4,5] 1) intractable to efficiently produce a near-optimal subset of informative genes when the total number of genes is too large (higher-dimensional data) due to the drawback of binary chromosome representation; 2) the high risk of over-fitting problems. The over-fitting problem that occurred on hybrid methods (e.g., GASVM-based methods) is also reported in a review paper in Saeys *et al.*[3]

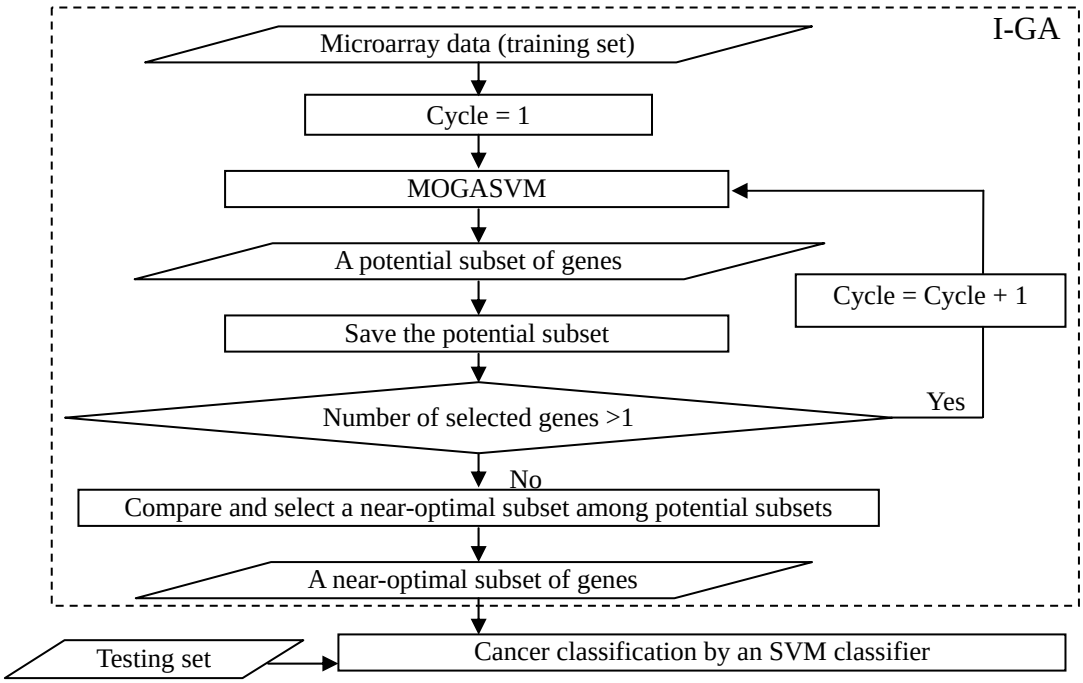

Fig.1. The general procedure of I-GA

In order to overcome the limitations of the previous works and solve the problems derived from microarray data, we propose an iterative approach based on multi-objective GASVM (MOGASVM). The ultimate goal of this paper is to automatically select a near-optimal (smaller) subset of informative genes that is most relevant for the cancer classification. To achieve the goal, we adopt the proposed method. It is evaluated on two real microarray data sets.

II. THE PROPOSED ITERATIVE APPROACH BASED ON MOGASVM (I-GA)

In this paper, we propose I-GA to overcome the problems derived from the previous works and microarray data.[1,2,4,5] I-GA is a hybrid approach based on MOGASVM. Details of MOGASVM can be found in Mohamad *et al.*[4] I-GA in our work differs from the methods in the previous works in one major part.[1,2,4,5] The major difference is that our proposed method involves an iterative approach, whereas the previous works did not use any iterative process for gene selection. The general procedure of I-GA is shown in Fig. 1.

Basically, I-GA repeats the process of MOGASVM to reduce the dimensionality of data iteratively. The description of each step is explained as follows:

Step 1: Starting an iterative process. It is repeated until the number of selected genes in the potential subset of the current cycle c is equal or less than 1. Every cycle is started here. In each cycle of I-GA, a number of selected genes are automatically selected by MOGASVM and the dimensionality is iteratively reduced.

Step 2: Starting MOGASVM to find and produce a potential subset of genes.

Step 3: Producing and saving the potential subset of selected genes. This potential subset is used for the next cycle (cycle c+1) as an input set. The selection of genes in the next cycle (cycle c+1) only uses genes in the potential subset that is resulted by the previous cycle (cycle c). Therefore, the dimensionality and complexity of solution spaces can be decreased on a cycle by cycle basis.

Step 4: A near-optimal subset is selected among the potential subsets based on the highest fitness value (the highest LOOCV accuracy with the smallest number of selected genes).

Step 5: An iterative process (Steps 1-4) results a near-optimal subset of genes. This subset is

possible to be found due to the dimensionality of data has been iteratively reduced. The near-optimal subset is then used to construct an SVM classifier, and the constructed SVM is tested by using the test set.

III. EXPERIMENTS

3.1. Data Sets

Two real microarray data sets are used to evaluate I-GA: Leukaemia cancer and lung cancer. The leukaemia data set contains the expression levels of 7,129 genes and can be obtained at http://www.broad.mit.edu/cgi-bin/cancer/datasets.cgi. it has two cancer classes: acute lymphoblastic leukaemia, acute myeloid leukaemia. In this data set, bone marrow and blood samples were taken from 72 patients (the training set contains 38 samples; the test set consists 34 samples). There are 181 samples in the lung cancer data set (the training set contains 32; the test set consists 149 samples). It has two tumour classes (malignant pleural mesothelioma and adenocarcinoma) and can be obtained at http://chestsurg.org/publications/2002-microarray.aspx.

3.2. Experimental Setup

Three criteria following their importance are considered to evaluate the performances of I-GA and other experimental methods: test accuracy, leave-one-out-cross-validation (LOOCV) accuracy, and the number of selected genes. Several experiments are conducted 10 times on each data set using I-GA and other experimental methods such as GASVM MOGASVM, GASVM version 2 (GASVM-II), and SVM. Next, an average result of the 10 independent runs is obtained. A near-optimal subset that produces the highest classification accuracies with the possible least number of genes is selected as the best subset.

3.3. Experimental Results

Table 1 shows the classification accuracy for each run using I-GA on both data sets. Interestingly, all runs have achieved 100% LOOCV accuracy on the data sets. This has proven that I-GA has efficiently selected and produced a near-optimal solution in a solution space. This is due to the fact of its ability to automatically reduce the dimensionality and complexity of the solution space on a cycle by cycle basis. Therefore, I-GA yields the near-optimal gene subset (a smaller subset of informative genes with higher classification accuracy) successfully.

Table 1. Classification accuracies for each run using I-GA

Run#	Leukaemia Data Set			Lung Data Set		
	LOOCV (%)	Test (%)	#Selected Genes	LOOCV (%)	Test (%)	#Selected Genes
1	100	85.35	5	100	90.60	2
2	100	91.18	5	100	95.30	2
3	100	91.18	3	100	93.29	3
4	100	85.29	5	100	95.30	4
5	100	85.29	5	100	85.24	2
6	100	82.35	5	100	83.22	3
7	100	82.35	4	100	92.62	2
8	100	100	5	100	97.32	2
9	100	88.24	5	100	96.64	2
10	100	85.29	4	100	95.30	3
Average ± S.D	100 ± 0	87.65 ± 5.33	4.60 ± 0.70	100 ± 0	92.48 ± 4.80	2.5 ± 0.71

Note: Results of the best subsets shown in shaded cells. S.D. denotes the standard deviation, whereas #Selected Genes represent a number of selected genes.

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

Data Set	Run#	Probe-set Name	Gene Description
Leukaemia	8	L15388_at	G PROTEIN-COUPLED RECEPTOR KINASE GRK5
		M95678_at	PLCB2 Phospholipase C, beta 2
		X15357_at	GB DEF = Natriuretic peptide receptor (ANP-A receptor)
		X55668_at	PRTN3 Proteinase 3
		S76473_s_at	TrkB [human, brain, mRNA, 3194 nt]
Lung	8	33328_at	ESTs
		609_f_at	Highly similar to SMHU1B metallothionein 1B [H.sapiens]

Note: Run# denotes a run number.

Table 3. The benchmark of the proposed I-GA with the other experimental methods and related previous works

Method	Leukaemia Data Set (Average ± S.D; The Best)			Lung Data Set (Average ± S.D; The Best)		
	#Selected Genes	Accuracy (%)		#Selected Genes	Accuracy (%)	
		LOOCV	Test		LOOCV	Test
I-GA	(4.60 ± 0.70; 5)	(100 ± 0; 100)	(87.65 ± 5.33; 100)	(2.5 ± 0.71; 2)	(100 ± 0; 100)	(92.48 ± 4.80; 97.32)
GASVM-II [2]	(10 ± 0; 10)	(100 ± 0; 100)	(81.18 ± 10.21; 94.12)	(10 ± 0; 10)	(100 ± 0; 100)	(59.33 ± 29.32; 97.32)
MOGASVM [4]	(2,212.6 ± 26.63; 2,189)	(95.53 ± 1.27; 97.37)	(84.41 ± 2.42; 88.24)	(4,418.5 ± 50.19; 4,433)	(75.31 ± 0.99; 78.13)	(85.84 ± 3.97; 93.29)
GASVM [2]	(3,574.9 ± 40.05; 3,531)	(94.74 ± 0; 94.74)	(83.53 ± 2.48; 88.24)	(6,267.8 ± 56.34; 6,342)	(75 ± 0; 75)	(84.77 ± 2.53; 87.92)
SVM [2]	(7,129 ± 0; 7,129)	(94.74 ± 0; 94.74)	(85.29 ± 0; 85.29)	(12,533 ± 0; 12,533)	(65.63 ± 0; 65.63)	(85.91 ± 0; 85.91)
Li *et al.*[1]	(4 ± NA; NA)	(100 ± NA; NA)	NA	NA	NA	NA
Peng *et al.*[5]	(6 ± NA; NA)	(100 ± NA; NA)	NA	NA	NA	NA

Note: The best result shown in shaded cells. S.D. denotes the standard deviation, whereas #Selected Genes represent a number of selected genes. 'NA' means that a result is not reported in the related previous works. Methods in *italic* style are experimented in this work.

Informative genes in the best gene subsets as produced by the proposed I-GA and reported in Table 1 are listed in Table 2. These informative genes among the thousand of genes may be the excellent candidates for clinical and 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 in the future.

According to Table 3, I-GA has outperformed the other experimental methods and previous works in terms of LOOCV accuracy, test accuracy, and the number of selected genes. The gap between LOOCV accuracy and test accuracy that resulted by I-GA was also lower. This small gap shows that the risk of the over-fitting problem can be reduced. Therefore, I-GA is more efficient than other experimental methods since it has produced the higher classification accuracies, smaller number of selected genes, smaller standard deviations, and smaller gap between LOOCV accuracy and test accuracy. However, due to the iterative process, I-GA is computationally more extensive than other methods.

IV. CONCLUSION

In this paper, I-GA has been proposed and tested for gene selection on two real microarray data. Based on the experimental results, the performance of I-GA was superior to the other experimental methods and related previous works. This is due to the fact that I-GA can automatically reduce the dimensionality of the data on a cycle by cycle basis. When the dimensionality was reduced, the combination of genes and the complexity of solution spaces can also be automatically decreased iteratively. This iterative process is done to generate potential gene subsets in higher-dimensional data (microarray data), and finally produce a near-optimal subset of informative genes. Hence, the gene selection using I-GA is needed to produce a near-optimal (smaller) subset of informative genes for better cancer classification. Moreover, focusing the attention on the informative genes in the best subset may provide insights into the mechanisms responsible for the cancer itself. Even though I-GA has classified tumours with higher accuracy, it is still not able to completely avoid the over-fitting problem. Therefore, a combination between a constraint approach and a hybrid approach will be developed to solve the problem.

REFERENCES

[1] Li S, Wu X, Hu X (2008) Gene selection using genetic algorithm and support vectors machines. Soft Comput 12:693–698

[2] 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

[3] Saeys Y, Inza I, Larranaga P (2007) A review of feature selection techniques in bioinformatics. Bioinformatics 23(19):2507–2517

[4] Mohamad MS, Omatu S, Deris S, *et al*. (Appear in press) A multi-objective strategy in genetic algorithm for gene selection of gene expression data. Int J Artif Life & Rob 13(2)

[5] Peng S, Xu Q, Ling XB, *et al*. (2003) Molecular classification of cancer types from microarray data using the combination of genetic algorithms and support vector machines. FEBS Lett 555:358–362