A TWO-STAGE METHOD FOR IDENTIFYING A SMALLER SUBSET OF GENES IN MICROARRAY DATA

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: Microarray data measured by microarray are useful for cancer classification. However, it faces with several problems in selecting genes for the classification due to many irrelevant genes, noisy data, and the availability of a small number of samples compared to a huge number of genes (high-dimensional data). Hence, this paper proposes a two-stage gene selection method to select a smaller (near-optimal) subset of informative genes that is most relevant for the cancer classification. It has two stages: 1) pre-selecting genes using a filter method to produce a subset of genes; 2) optimising the gene subset using a multi-objective hybrid method to automatically yield a smaller subset of informative genes. Two microarray data sets are used to test the effectiveness of the proposed method. Experimental results show that the performance of the proposed method is superior to other experimental methods and related previous works.

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

1. INTRODUCTION

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 poses a major challenge because of the following characteristics of microarray data:

- High-dimensional data, for example, a huge number of genes and a small number of samples are in the ranges of 7,000-15,000 and 30-200, respectively.
- Most genes are not relevant for classifying different tissue types.
- These data have noisy genes.

To overcome the problems, a gene selection method is used to select a subset of genes that maximises 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.

In the context of cancer classification, gene selection methods can be classified into two categories. 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 since it is computationally more efficient than the hybrid method [9]. However, the hybrid approach usually provides greater accuracy than the filter approach. Until now, several hybrid methods [2-7], especially a combination between a genetic algorithm (GA) [1] and a support vector machine (SVM) [8] classifier (GASVM), have been implemented to select informative genes. Generally, our previous hybrid methods, i.e., GASVM-based methods performed well in high-dimensional data since we proposed a modified chromosome representation, a cyclic approach, and a multi-objective strategy [3-6]. However, the methods yielded inconsistent results when they were run independently.

The previous work of [2] that proposed GASVM-based methods can simultaneously optimise genes and SVM parameter settings. The work of [7] introduced a recursive feature elimination post-processing step after the step of a GASVM-based method in order to reduce the number of selected genes again. Nevertheless, the hybrid methods (GASVM-based methods) of the previous works are intractable to efficiently produce a smaller subset of genes in high-dimensional data due to their binary chromosome representation drawback [2],[7]. The total number of gene subsets produced by the GASVM-based methods in the previous works are calculated by $M_c = 2^M - 1$ where M_c is the total number of subsets, whereas M is the total number of genes. Based on this equation, the GASVM-based methods are almost impossible to evaluate all possible subsets of selected genes if M is too many (high-dimensional data). Although the work of [7] implemented a pre-processing step to decrease the dimensionality of data, but it can only reduce a small number of genes, and many genes are still available in the data. The GASVM-based methods [2],[7] also face with the high risk of over-fitting problems. An over-fitting problem is happened because the number of genes

greatly exceeds the number of samples. The over-fitting problem that occurred on hybrid methods (e.g., GASVM-based methods) is also reported in a review paper in [9].

In order to solve the problems derived from microarray data and overcome the limitation of the GASVM-based methods in the previous works [2-7], we propose a two-stage gene selection method (Filter+MOGASVM). This proposed method is able to perform well in high-dimensional data and reduce a risk of over-fitting problems since it has two stages as follows: stage 1 to decrease the dimensionality of data; stage 2 to produce a smaller (near-optimal) genes subset. The diagnostic goal is to develop a medical procedure based on the least number of possible genes that needed to detect diseases. Thus, the ultimate goal of this paper is to select a smaller subset of informative genes (minimise the number of selected genes) for yielding higher cancer classification accuracy (maximise the classification accuracy). To achieve the goal, we adopt Filter+MOGASVM. The proposed method is evaluated on two real microarray data sets of tumour samples.

The outline of this paper is as follows: Section 2 discusses the detail of the proposed Filter+MOGASVM. In Section 3, microarray data sets used, experimental setup, and experimental results are described. The conclusion of this paper is provided in Section 4.

2. THE PROPOSED TWO-STAGE GENE SELECTION METHOD (FILTER+MOGASVM)

In this paper, we propose Filter+MOGASVM to overcome the drawbacks of GASVM-based methods in the related previous works [2-7]. Filter+MOGASVM in our work differs from the methods in the previous works in one major part. The major difference is that our proposed method involves two stages (using a filter method and a hybrid method), whereas the previous works usually used only one stage (using a hybrid method) for gene selection. The difference is necessary in order to produce a smaller (near-optimal) gene subset from high-dimensional data and reduce the high risk of over-fitting problems. For more understanding, the general flowcharts of our work and the previous works are shown in Fig. 1 (a) and Fig. 1 (b), respectively. The detailed stages of Filter+MOGASVM are described as follows.

2.1 Stage 1: Pre-Selecting Genes Using a Filter Method

In the first stage, we apply a filter method such as gain ratio (GR) or information gain (IG) on the training set to pre-select genes and finally produce a subset of genes. After the pre-select process, the dimensionality of data is also decreased. The filter method calculates and ranks a score for each gene. Genes with the highest scores are selected and put into the gene subset. This subset is used as an input to the second stage.

Since GASVM-based methods in previous works performs poorly in high-dimensional data, and meanwhile, we use a GASVM-based method (MOGASVM) in the second stage of Filter+MOGASVM, a filter method (GR or IG) in this first stage is used to reduce the high-dimensional in order to overcome the drawback of GASVM-based methods. If the subset that produced by the filter method is small-dimension, the combination of genes is not complex, and then MOGASVM in the next stage can possible to produce a smaller (near-optimal) subset of informative genes.

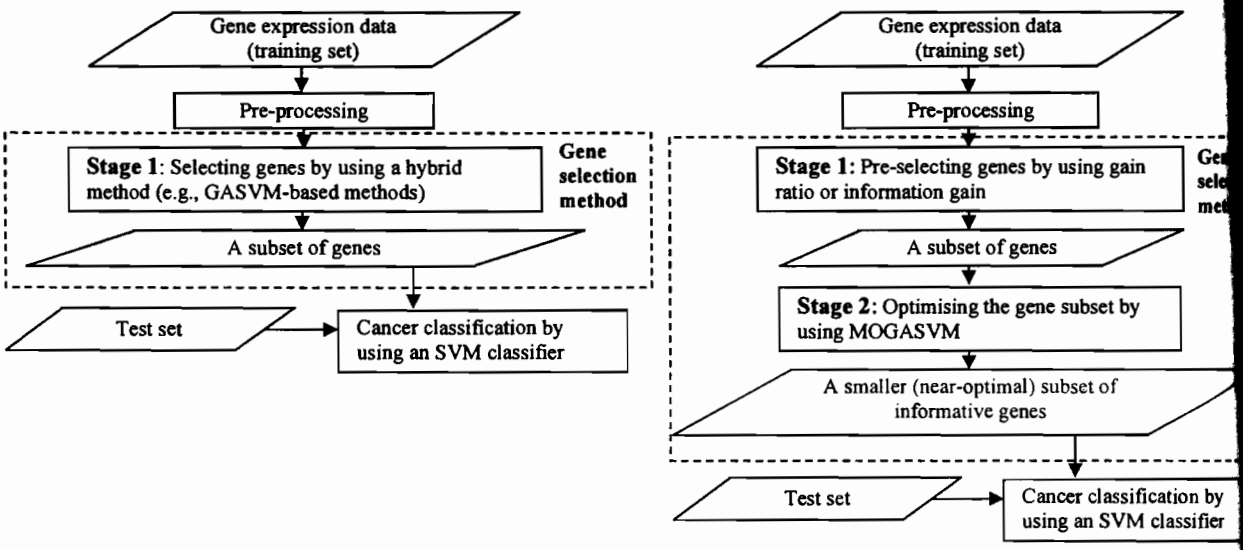

Figure 1. General flowcharts of (a) previous works (GASVM-based methods); (b) our work (Filter+MOGASVM).

2.2 Optimizing a Gene Subset Using MOGASVM

In this stage, we develop and use MOGASVM to automatically optimise the gene subset that is produced by the first stage, and finally yield a smaller (near-optimal) subset of informative genes. This smaller subset is identified by an evaluation function in MOGASVM that uses two criteria: maximisation of the leave-one-out-cross-validation (LOOCV) accuracy and minimisation of the number of selected genes. MOGASVM selects and optimises genes by considering relations among them in order to remove irrelevant and noisy genes. The smaller subset is possible to be found due to the dimensionality and complexity of data has been firstly reduced by the first stage. The high risk of over-fitting problems can be also decreased because of the reduction. The detail of MOGASVM can be found in [4].

Finally, the smaller subset of the training set is used to construct an SVM classifier for cancer classification, and the constructed SVM is then tested by using the test set

(independent set). This paper has produced two methods of Filter+MOGASVM obtained from combinations of two different filter methods (GR and IG) and MOGASVM. These methods are GR+MOGASVM and IG+MOGASVM.

3. EXPERIMENTS

3.1 Data Sets

Two benchmark microarray data sets that contain binary classes and multi-classes are used to evaluate Filter+MOGASVM. It is summarised in Table 1.

Table 1. The summary of microarray data sets.

Data set	Number of classes	Number of samples in the training set	Number of samples in the test set	Number of genes	Source
Lung	2 (MPM and ADCA)	32 (16 MPM and 16 ADCA)	149 (15 MPM and 134 ADCA	12,533	http://chestsurg.org/publications/2002-microarray.aspx
MLL	3 (ALL, MLL, and AML)	57 (20 ALL, 17 MLL, and 20 AML)	15 (4 ALL, 3 MLL, and 8 AML)	12,582	http://www.broad.mit.edu/cgi-bin/cancer/datasets.cgi

Note:
MPM = malignant pleural mesothelioma. MLL = mixed-lineage leukaemia. ADCA = adenocarcinoma.
ALL = acute lymphoblastic leukaemia. AML = acute myeloid leukaemia.

3.2 Experimental Setup

Since the number of training samples in microarray data is small, the cross-validation (CV) accuracy on the training set is calculated through an LOOCV procedure [3]. For the test accuracy, a classifier is built using all the training samples, and the classes of test samples from the test set are predicted one by one using the built classifier. The test accuracy is estimated by the number of the correctly classified, divided by the number of samples in the test set.

Table 2 contains parameter values for Filter+MOGASVM. These values are chosen based on the results of preliminary runs. Three criteria following their importance are considered to evaluate and compare the performance of Filter+MOGASVM with existing methods [2-7]: test accuracy, CV accuracy, and the number of selected genes. Higher accuracies and a smaller number of selected genes are needed to obtain an excellent performance. The top 200 genes are pre-selected by using GR and IG in the first stage of the proposed method, and are then used for the second stage. Several experiments are conducted 10 times on each data set using Filter+MOGASVM and other experimental methods such as GASVM (single-objective), MOGASVM, GASVM version 2 (GASVM-II), and SVM.

Filter+GASVM methods (IG+GASVM and GR+GASVM) are also experimented for the comparison. Next, an average result of the 10 independent runs is obtained.

Table 2. Parameter Settings for Filter+MOGASVM.

Parameter	Lung Data Set	MLL Data Set
Size of population	100	100
Number of generation	300	300
Crossover rate	0.7	0.7
Mutation rate	0.01	0.01
Weight 1, w_1	0.7	0.7
Weight 2, w_2	0.3	0.3
Cost for SVM	0.7	100

3.3 Experimental Results

3.3.1 LOOCV and test accuracies of selected genes with Filter+MOGASVM

Tables 3 and 4 show the results for each run on the lung and MLL data sets, respectively. The results of the best subsets are shown in shaded cells, whereas the results in boldface display the best result of averages. S.D. denotes the standard deviation. Run# and #Genes represent a run number and a number of selected genes, respectively. Almost all runs have achieved 100% LOOCV accuracy on all the data sets. This has proved that Filter+MOGASVM has efficiently selected and produced a near-optimal gene subset from a solution space.

Table 3. Classification accuracies using Filter+MOGASVM on the lung data set.

Run#	GR+MOGASVM (Filter+MOGASVM)			IG+MOGASVM (Filter+MOGASVM)		
	LOOCV (%)	Test (%)	#Genes	LOOCV (%)	Test (%)	#Genes
1	100	98.66	2	100	97.99	2
2	100	94.63	2	100	96.64	2
3	100	95.30	2	100	97.32	2
4	100	97.32	2	100	97.32	2
5	100	95.97	2	100	94.63	2
6	100	97.99	2	100	95.30	2
7	100	95.97	2	100	95.30	2
8	100	95.97	2	100	95.97	2
9	100	95.97	2	100	**99.33**	2
10	100	93.96	2	100	93.29	2
Average	100	96.18	2	**100**	**96.31**	**2**
± S.D.	± 0	± 1.45	± 0	**± 0**	**± 1.77**	± 0

Table 4. Classification accuracies using Filter+MOGASVM on the MLL data set.

Run#	GR+MOGASVM (Filter+MOGASVM)			IG+MOGASVM (Filter+MOGASVM)		
	LOOCV (%)	Test (%)	#Genes	LOOCV (%)	Test (%)	#Genes
1	100	93.33	6	100	93.33	7
2	100	93.33	6	100	93.33	6
3	100	100	5	100	100	7
4	100	93.33	7	98.25	100	6
5	100	100	5	100	93.33	7
6	100	93.33	6	100	93.33	5
7	100	100	5	100	100	7
8	100	100	7	100	100	6
9	100	100	5	100	100	5
10	100	93.33	4	100	86.67	7
Average ± S.D.	**100 ± 0**	**96.67 ± 3.51**	**5.60 ± 0.97**	99.83 ± 0.56	96.00 ± 4.66	6.30 ± 0.82

3.3.2 Filter+MOGASVM versus other experimental methods

The benchmark of Filter+MOGASVM in comparison with other experimental methods that have been experimented in this work is summarized in Table 5. Overall, the LOOCV and test accuracies of Filter+MOGASVM for all the data sets were higher than Filter+GASVM, MOGASVM, GASVM-II, GASVM, and SVM. Moreover, the number of selected genes by using Filter+MOGASVM was also lower.

Based on the standard deviations of LOOCV accuracy, test accuracy, and the number of selected genes, Filter+MOGASVM was also more consistent than the other experimental methods except the SVM classifier. This SVM classifier achieved 0 for the standard deviations in all experiments since it did not implement any gene selection approach. The gap between LOOCV accuracy and test accuracy that resulted by Filter+MOGASVM was also lower. This small gap shows that the risk of the over-fitting problem can be reduced. On the other hand, the results of LOOCV accuracy of the others were much higher than their test accuracy because they were unable to avoid or reduce the risk of over-fitting problems. Over-fitting is a major problem of hybrid methods in gene selection and classification of microarray data when the classification accuracy on training samples, e.g., LOOCV accuracy is much higher than the test accuracy.

GASVM and MOGASVM cannot produce a near-optimal subset of informative genes because they perform poorly in high-dimensional data due to their chromosome representation drawback. GASVM-II method is impractical to be used in real applications because a variety number of selected genes should be tested in order to obtain the near-optimal one. On the contrary, the proposed Filter+MOGASVM that pre-selects a number of genes in the first stage can automatically optimise the selected genes by the second stage in

order to remove irrelevant genes and produce a smaller (near-optimal) subset of informative genes.

Table 5. The benchmark of Filter+MOGASVM with Filter+GASVM and the previous methods on the lung and MLL data sets.

Method	Lung Data Set (Average ± S.D.; The Best)			MLL Data Set (Average ± S.D.; The Best)		
	#Genes	Accuracy (%)		#Genes	Accuracy (%)	
		LOOCV	Test		LOOCV	Test
GR+MOGASVM (Filter+MOGASVM)	2 ± 0; 2	100 ± 0; 100	96.18 ± 1.45; 98.66	5.60 ± 0.97; 5	100 ± 0; 100	96.67 ± 3.51; 100
IG+MOGASVM (Filter+MOGASVM)	2 ± 0; 2	100 ± 0; 100	96.31 ± 1.77; 99.33	6.30 ± 0.82; 5	99.83 ± 0.56; 100	96.00 ± 4.66; 100
GR+GASVM (Filter+GASVM)	101 ± 8.50; 105	100 ± 0; 100	86.04 ± 3.66; 90.60	100.40 ± 6.42; 98	100 ± 0; 100	90.67 ± 5.62; 100
IG+GASVM (Filter+GASVM)	100.3 ± 8.02; 87	100 ± 0; 100	84.30 ± 7.86; 88.59	100.20 ± 7.63; 99	100 ± 0; 100	88.67 ± 3.22; 93.33
A recursive GASVM [6]	2.80 ± 1.32; 4	100 ± 0; 100	93.69 ± 2.52; 98.66	12.0 ± 5.58; 20	100 ± 0; 100	91.33 ± 5.49; 100
GASVM-II+GASVM [5]	2.1 ± 0.32; 2	100 ± 0; 100	94.16 ± 6.85; 98.66	6.5 ± 0.71; 6	100 ± 0; 100	92 ± 8.20; 100
GASVM-II [3]	10 ± 0; 10	100 ± 0; 100	59.33 ± 29.32; 97.32	30 ± 0; 30	100 ± 0; 100	84.67 ± 6.33; 93.33
MOGASVM [4]	4,418.5 ± 50.19; 4,433	75.31 ± 0.99; 78.13	85.84 ± 3.97; 93.29	4,465.2 ± 18.34; 4,437	94.74 ± 0; 94.74	90 ± 3.51; 93.33
GASVM [3]	6,267.8 ± 56.34; 6,342	75 ± 0; 75	84.77 ± 2.53; 87.92	6,298.8 ± 51.51; 6,224	94.74 ± 0; 94.74	87.33 ± 2.11; 86.67
SVM classifier [4]	12,533 ± 0; 12,533	65.63 ± 0; 65.63	85.91 ± 0; 85.91	12,582 ± 0; 12,582	92.98 ± 0; 92.98	86.67 ± 0; 86.67

Note: The best result of each data set shown in shaded cells. S.D. denotes the standard deviation, whereas #Genes represents a number of selected genes.

Overall, this work has outperformed the related previous works on both the data sets in terms of classification accuracy and the number of selected genes. Filter+MOGASVM in our work has produced a near-optimal (smaller) gene subset from high-dimensional data and reduced the high risk of over-fitting problems. This is due to the fact that a filter method in the first stage of Filter+MOGASVM reduces the dimensionality of the solution space in order to produce a gene subset. Next, MOGASVM in the second stage of Filter+MOGASVM optimises the subset automatically to yield a smaller subset of informative genes with higher classification accuracy. This smaller subset is obtained since Filter+MOGASVM considers and optimises a relation among genes.

4. CONCLUSIONS

In this paper, Filter+MOGASVM has been proposed and tested for gene selection on two real microarray data sets that contain binary classes and multi-classes of tumour samples. Based on the experimental results, the performance of Filter+MOGASVM was superior to the other

experimental methods and related previous works. This is due to the fact that the filter method in the first stage of the proposed method can pre-select genes and reduce dimensionality of data in order to produce a subset of genes. When the dimensionality was reduced, the combination of genes and complexity of solution spaces were automatically decreased. The second stage of Filter+MOGASVM can automatically optimise the subset that is yielded by the first stage. This optimisation process is done to remove irrelevant and noisy genes, and finally produce a smaller (near-optimal) subset of informative genes. Hence, the gene selection using Filter+MOGASVM is needed to produce a smaller subset of informative genes for better cancer classification of microarray data. However, due to the application of a filter method in the first stage of Filter+MOGASVM, pre-selecting genes is difficult since it is manually done. Even though Filter+MOGASVM has classified tumours with higher accuracy, it is still not able to completely avoid the over-fitting problem. Therefore, a combination between constraint based reasoning methods and particle swarm optimisation techniques is recently developed to solve the over-fitting problem.

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] Elmahi, I., Grunder, O. and Elmoudni, A., "A modelling-optimization approach for discrete event systems using the (MAX,+) algebra and genetic algorithms", International Journal of Innovative Computing., Volume 2, pp.771–788, 2006.

[2] Huang, H. L. and Chang, F. L., "ESVM: Evolutionary support vector machine for automatic feature selection and classification of microarray data", BioSystems, Volume 90, pp.516–528, 2007.

[3] 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.

[4] Mohamad, M. S., Omatu, S., Deris, S., Misman, M. F. and Yoshioka, M., "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.

[5] Mohamad, M. S., Omatu, S., Deris, S., Misman, M. F. and Yoshioka, M., "Selecting informative genes from microarray data by using hybrid methods for cancer classification", International Journal of Artificial Life & Robotics, Volume 13, Issue 2, 2008.

[6] Mohamad, M. S., Omatu, S., Deris, S. and Yoshioka, M., "A recursive genetic algorithm to automatically select genes for cancer classification", Proceedings of the 2nd International Workshop on Practical Application of Computational Biology & Bioinformatics, Corchado, J. M., Paz, J. F. D., Rocha, M. P. and Juan, F. F. R, (eds.), Berlin/Heidelberg, Springer-Verlag, Advances in Soft Computing, Volume 49, pp.166–174, 2009.

[7] Peng, S., Xu, Q., Ling, X. B., Peng, X., Du, W. and Chen, L., "Molecular classification of cancer types from microarray data using the combination of genetic algorithms and support vector machines", FEBS Letters, Volume 555, pp.358–362, 2003.

[8] Q. She, H. Su, L. Dong and J. Chu, "Support vector machine with adaptive parameters in image coding", International Journal of Innovative Computing, Volume 4, pp.359–367, 2008.

[9] Saeys, Y., Inza, I. and Larranaga, P., "A review of feature selection techniques in bioinformatics", Bioinformatics, Volume 23, Issue 19, pp.2507–2517, 2007.