An Iterative GASVM-Based Method: Gene Selection and Classification of Microarray Data

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

[1] Department of Computer Science and Intelligent Systems, Graduate School of Engineering, Osaka Prefecture University, Sakai, Osaka 599-8531, Japan
mohd.saberi@sig.cs.osakafu-u.ac.jp,
{omatu,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
safaai@utm.my

Abstract. Microarray technology has provided biologists with the ability to measure the expression levels of thousands of genes in a single experiment. One of the urgent issues in the use of microarray data 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 the small number of samples compared to the huge number of genes (higher-dimensional data). In this study, we propose an iterative method based on hybrid genetic algorithms to select a near-optimal (smaller) subset of informative genes in classification of the microarray data. The experimental results show that our proposed method is capable in selecting the near-optimal subset to obtain better classification accuracies than other related previous works as well as four methods experimented in this work. Additionally, a list of informative genes in the best gene subsets is also presented for biological usage.

Keywords: Gene selection, genetic algorithm, iterative method, hybrid approach, microarray data.

1 Introduction

The recent development of microarray technologies has enabled biologists to quantify the expression levels of thousands of genes in a single experiment. It 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.

S. Omatu et al. (Eds.): IWANN 2009, Part II, LNCS 5518, pp. 187–194, 2009.

To overcome the challenge, a gene selection method is normally used to select a subset of genes that increases the classifier's ability to classify samples more accurately. Efficient gene selection methods can yield a more compact gene subset without the loss of classification accuracy. These methods also can reduce the dimensionality of data, and remove irrelevant and noisy genes. In addition, a smaller number of selected genes can be more conveniently and economically used for diagnostic purposes in clinical settings.

There are two types of the 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],[4-6]. The drawbacks of the hybrid methods (GASVM-based methods) in the previous works are: 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) was also reported in a review paper in Saeys *et al*. [3].

In order to overcome the limitations of the previous works and solve the problems derived from microarray data, we propose an iterative GASVM-based method (I-GASVM). 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 four real microarray data sets.

2 The Proposed Iterative GASVM-Based Method (I-GASVM)

In this paper, we propose I-GASVM to overcome the problems derived from the previous works and microarray data [1],[2],[4-6]. I-GASVM is a hybrid approach based on MOGASVM. Details of MOGASVM can be found in Mohamad *et al*. [4]. I-GASVM in our work differs from the methods in the previous works in one major part [1],[2],[4-6]. The major difference is that our proposed method involves an iterative approach, whereas the previous works did not use the iterative approach for gene selection. The general procedure of I-GASVM is shown in Fig. 1. Basically, I-GASVM 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. The number of cycles is based on the satisfied condition of genes numbers. In each cycle of I-GASVM, 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.

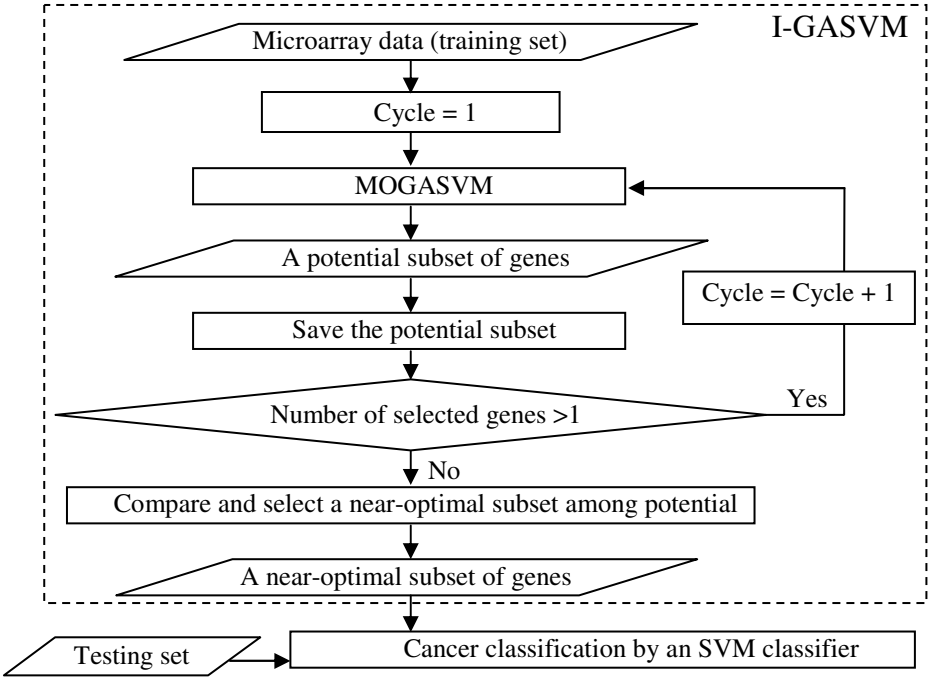

Fig. 1. The general flowchart of I-GASVM

3 Experiments

3.1 Data Sets and Experimental Setup

Four real microarray data sets that contain binary classes and multi-classes are used to evaluate I-GASVM: leukaemia cancer, colon cancer, lung cancer, and mixed-lineage leukaemia (MLL) cancer data sets. Table 1 summarises the data sets. For the colon data set, only the training set is available in the downloaded source.

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
Leukaemia	2 (ALL and AML)	38 (27 ALL and 11 AML)	34 (20 ALL and 14 AML)	7,129	http://www.broad.mit.edu/cgi-bin/cancer/datasets.cgi
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
Colon	2 (Normal and tumour)	62 (22 normal and 40 tumour)	Not available	2,000	http://microarray.princeton.edu/oncology/affydata/index.html

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

Three criteria following their importance are considered to evaluate the performances of I-GASVM 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-GASVM and other experimental methods such as GASVM (single-objective), 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.

3.2 Experimental Results

Tables 2 and 3 show that the classification accuracy for each run using I-GASVM on all data sets. Interestingly, almost all runs have achieved 100% LOOCV accuracy. This has proven that I-GASVM 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 on a cycle by cycle basis. Therefore, I-GASVM yields the near-optimal gene subset (a smaller subset of informative genes with higher classification accuracy) successfully.

Generally, near-optimal subsets that obtained from almost all run on the data sets contain less than 10 genes. This is inline with the diagnostic goal of developed medical procedures that needs the least number of possible informative genes to detect diseases. The conservativeness of the results in Tables 2 and 3 is controlled and maintained by the iterative approach and the fitness function of I-GASVM that maximises the classification accuracy and meanwhile, minimises the number of selected genes.

Practically, the best subset of a data set is firstly chosen and the genes in it are then listed for biological usage. The best subset is chosen based on the highest classification accuracy with the smallest number of selected genes. The highest accuracy gives confidence to us for the most accurate classification of cancer types. Moreover, the smallest number of selected genes for cancer classification can reduce the cost in clinical settings.

Table 2. Results for each run using I-GASVM on the leukaemia and lung data sets

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 represents a number of selected genes.

Table 3. Results for each run using I-GASVM on the MLL and colon data sets

Run#	MLL Data Set			Colon Data Set	
	LOOCV (%)	Test (%)	#Selected Genes	LOOCV (%)	#Selected Genes
1	100	86.67	8	100	13
2	100	100	6	100	13
3	100	80	9	100	14
4	100	73.33	9	95.16	5
5	100	86.67	8	96.77	6
6	100	80	6	100	7
7	100	86.67	7	100	10
8	100	93.33	8	98.39	9
9	100	93.33	7	100	10
10	100	80	6	100	10
Average ± S.D	100 ± 0	86 ± 7.98	7.4 ± 1.17	99.03 ± 1.73	9.70 ± 3.06

Note: Results of the best subsets shown in shaded cells. S.D. denotes the standard deviation, whereas #Selected Genes represents a number of selected genes. The colon data set only has LOOCV accuracy since it only has the training set.

Informative genes in the best gene subsets as produced by the proposed I-GASVM and reported in Tables 2 and 3, are listed in Table 4. 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 in the future.

For an objective comparison, we only compare our work with related previous works that used GASVM-based methods in their work [1],[2],[4-6]. Moreover, the previous works also produced the average of classification accuracy results since they used hybrid approaches. We make the comparison using the averages of LOOCV accuracy and the number of selected genes. This is due to the most previous works only evaluated the performance of their approaches using the LOOCV procedure or *k*-fold-cross-validation and the number of selected genes on averages.

Table 4. 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]
MLL	2	35083_at	Human DNA sequence from clone RP4-681N20 on chromosome 20p12.1-
		36436_at	Homo sapiens mRNA for LECT2 precursor, complete cds
		36873_at	Human gene for very low density lipoprotein receptor, 5'flanking.
		40518_at	Human mRNA for T200 leukocyte common antigen (CD45, LC-A)
		35794_at	Homo sapiens mRNA for KIAA0942 protein, partial cds
		41827_f_at	Homo sapiens cDNA, 3' end
Colon	6	H80240	INTER-ALPHA-TRYPSIN INHIBITOR COMPLEX COMPONENT II PRECURSOR (Homo sapiens)
		T62220	CALPACTIN I LIGHT CHAIN (HUMAN);.
		H22688	UBIQUITIN (HUMAN);.
		T88902	COT PROTO-ONCOGENE SERINE/THREONINE-PROTEIN KINASE (Homo sapiens)
		U00968	STEROL REGULATORY ELEMENT BINDING PROTEIN 1 (HUMAN);.
		T84082	ER LUMEN PROTEIN RETAINING RECEPTOR 1 (HUMAN);.
		T62947	60S RIBOSOMAL PROTEIN L24 (Arabidopsis thaliana)

Note: Run# represents a run number.

According to Tables 5 and 6, I-GASVM 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-GASVM was also lower. This small gap shows that the risk of the overfitting problem can be reduced. Therefore, I-GASVM 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.

Table 5. The benchmark of the proposed I-GASVM with the other experimental methods and related previous works on the leukaemia and lung cancer data sets

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-GASVM	(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
Huang and Chang [6]	(3.4 ± NA; NA)	(100 using 10-CV ± NA; NA)	NA	NA	NA	NA

Note: The best result shown in shaded cells. S.D. denotes the standard deviation, whereas #Selected Genes and 10-CV represent a number of selected genes and 10-fold-cross-validation, respectively. 'NA' means that a result is not reported in the related previous works. Methods in *italic* style are experimented in this work.

Table 6. The benchmark of the proposed I-GASVM with the other experimental methods and related previous works on the MLL and colon cancer data sets

Method	MLL Data Set (Average ± S.D; The Best)			Colon Data Set (Average ± S.D; The Best)	
	#Selected Genes	Accuracy (%)		#Selected Genes	LOOCV Accuracy (%)
		LOOCV	Test		
I-GASVM	(7.4 ± 1.17; 6)	(100 ± 0; 100)	(86 ± 7.98; 100)	(9.7 ± 3.06; 7)	(99.03 ± 1.73; 100)
GASVM-II [2]	(30 ± 0; 30)	(100 ± 0; 100)	(84.67 ± 6.33; 93.33)	(30 ± 0; 30)	(99.03 ± 0.83; 100)
MOGASVM [4]	(4,465.2 ± 18.34; 437)	(94.74 ± 0; 94.74)	(90 ± 3.51; 93.33)	(446.3 ± 8.90; 446)	(93.23 ± 1.02; 95.16)
GASVM [2]	(6,298.8 ± 51.51; 224)	(94.74 ± 0; 94.74)	(87.33 ± 2.11; 86.67)	(979.8 ± 5.80; 940)	(91.77 ± 0.51; 91.94)
SVM [2]	(12,582 ± 0; 12,582)	(92.98 ± 0; 92.98)	(86.67 ± 0; 86.67)	(2,000 ± 0; 2,000)	(85.48 ± 0; 85.48)
Li *et al.* [1]	NA	NA	NA	15 ± NA; NA	(93.55 ± NA; NA)
Peng *et al.* [5]	NA	NA	NA	(12 ± NA; NA)	(93.55 ± NA; NA)

4 Conclusions

In this paper, I-GASVM has been proposed and tested for gene selection on four real microarray data. Based on the experimental results, the performance of I-GASVM was superior to the other experimental methods and related previous works. This is due to the fact that I-GASVM can automatically reduce the dimensionality of the data on a cycle by cycle basis. When the dimensionality was reduced, the combination of genes 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-GASVM 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-GASVM 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.: Gene selection using genetic algorithm and support vectors machines. Soft Comput. 12, 693–698 (2008)
2. Mohamad, M.S., Deris, S., Illias, R.M.: 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 (2005)
3. Saeys, Y., Inza, I., Larranaga, P.: A review of feature selection techniques in bioinformatics. Bioinformatics 23(19), 2507–2517 (2007)
4. Mohamad, M.S., Omatu, S., Deris, S., Misman, M.F., Yoshioka, M.: A multi-objective strategy in genetic algorithm for gene selection of gene expression data. J. Artif. Life. & Rob. 13(2), 410–413 (2009)
5. Peng, S., Xu, Q., Ling, X.B., Peng, X., Du, W., Chen, L.: Molecular classification of cancer types from microarray data using the combination of genetic algorithms and support vector machines. FEBS Lett. 555, 358–362 (2003)
6. Huang, H.L., Chang, F.L.: ESVM: Evolutionary support vector machine for automatic feature selection and classification of microarray data. BioSystems 90, 516–528 (2007)