A Three-Stage Method to Select Informative Genes for Cancer Classification

Mohd Saberi Mohamad, Sigeru Omatu, Michifumi Yoshioka

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

E-mail: mohd.saberi@sig.cs.osakafu-u.ac.jp; {sigeru; yoshioka}@cs.osakafu-u.ac.jp

Safaai Deris

Department of Software Engineering, Faculty of Computer Science and Information
Systems, Universiti Teknologi Malaysia, 81310 Skudai, Johore, Malaysia

E-mail: 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 a three-stage gene selection method to select a smaller subset of informative genes that is most relevant for the cancer classification. It has three 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 yield near-optimal gene subsets; 3) analysing the frequency of appearance of each gene in the different near-optimal gene subsets to produce a smaller subset of informative genes. The experimental results show that our proposed method is capable in selecting the smaller 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.

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. An informative gene is useful for cancer classification. However, the gene selection process poses a major challenge because of the following charac-

teristics 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 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 gene selection methods [1]: 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 [2–5]. However, the filter approach results in inclusion of irrelevant and noisy genes in a gene subset for the cancer classification. The hybrid approach usually provides greater accuracy than the filter approach. 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],[6–11]. The drawbacks of the hybrid methods (GASVM-based methods) in the previous works are [1],[6–11]: 1) intractable to efficiently produce a smaller subset of informative genes when the total number of genes is too large (higher-dimensional data); 2) the high risk of over-fitting problems.

In order to solve the problems derived from microarray data and overcome the limitation of the hybrid methods in the previous works [1],[6–11], we propose a three-stage gene selection method (3-SGS). This proposed method is able to perform well in the higher-dimensional data and reduce the high risk of over-fitting problems since it has three stages as follows: stage 1 for producing a subset of

Fig. 1. The proposed method (3-SGS).

genes; stage 2 for resulting near-optimal subsets of genes; stage 3 for yielding a smaller (final) subset of informative genes based on the frequency of appearance for each gene in the near-optimal subsets. 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 3-SGS. 3-SGS is evaluated on two real microarray data sets of tumour samples.

The outline of this paper is as follows: Sections 2 and 3 discuss previous works and the detail of the proposed three-stage method, respectively. In Section 4, microarray data sets used, experimental setup, and experimental results are described. The conclusion of this paper is provided in Section 5.

2 Previous Works

Several hybrid methods, i.e., GASVM-based methods have been proposed for genes selection of microarray data [1],[6-11]. The hybrid methods usually provide greater accuracy than filter methods since genes are selected by considering relations among genes. Generally, our previous GASVM-based methods performed well in higher-dimensional data, e.g., microarray data since we proposed a modified chromosome representation, a cyclic approach, and a multi-objective approach [6-9]. However, the methods yielded inconsistent results when they were run independently. The work of Huang and Chang can simultaneously optimise genes and SVM parameter settings by using a GASVM-based method [10]. Next, integrated algorithms based on GASVM have been proposed by the work of Shah and Kusiak [1] to produce a small subset of genes. Peng *et al.* 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 [11].

Nevertheless, the GASVM-based methods of the previous works are still intractable to efficiently produce a smaller subset of informative genes from higher-dimensional data due to their binary chromosome representation drawback [1],[6-11]. The total number of

gene subsets produced by GASVM-based methods is calculated by $M_c = 2^M - 1$ where M_c is the total number of gene 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 (higher-dimensional data). Although the work of Peng *et al.* have 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 [11]. The GASVM-based methods also face with 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.* [12].

3 The Proposed Three-stage Gene Selection Method (3-SGS)

In order to overcome the drawbacks of GASVM-based methods in the related previous works [1],[6-11], we propose a three-stage gene selection method (3-SGS). 3-SGS in our work differs from the methods in the previous works in one major part. The major difference is that our proposed method involves three stages, whereas the previous works usually used only one stage (using a hybrid method) [1],[6-10] or two stages (using a filter method and a hybrid method) [11] for gene selection. In the third stage, our method implements frequency analysis to identify the most frequently selected genes in near-optimal gene subsets, whereas the previous works [1],[6-11] rely solely on a filter method or a hybrid method in the first stage of their methods. The difference is necessary in order to produce near-optimal gene subsets from higher-dimensional data, reduce the high risk of over-fitting problems, and finally yield a smaller subset of informative genes. 3-SGS is shown in Fig. 1 and described as follows:

3.1 Stage 1: Pre-selecting Genes Using a Filter Method

A filter method such as gain ratio (GR) or information gain (IG) is used 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

Table 1. The summary of microarray data sets

Data set [Reference]	Number of classes	Number of samples in the training set	Number of samples in the test set	Number of genes	Source
Leukaemia [3]	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 [4]	2 (MPM and ADCA)	32 (16 MPM and 16 ADCA)	149 (15 MPM and 134 ADCA)	12,533	http://chest Surg.org/publications/2002-microarray.aspx

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

calculates and ranks a score for each gene. Genes with the highest scores are selected and put into a gene subset. This subset is then used as an input to the second stage.

Since GASVM-based methods in previous works performs poorly in higher-dimensional data, and meanwhile, we also use a GASVM-based method, i.e., a multi-objective GASVM (MOGASVM) in the second stage of 3-SGS, a filter method (GR or IG) in this first stage is used to reduce the higher-dimensional in order to overcome the drawback of GASVM-based methods. If the subset that produced by the filter method is in small-dimension, the combination of genes is not complex, and then MOGASVM can possible to produce near-optimal genes subsets.

3.2 Stage 2: Optimising a Gene Subset Using MOGASVM

In this stage, we develop MOGASVM to automatically optimise a gene subset that is produced by the first stage, and finally yield near-optimal subsets of genes. This stage is cycled until the maximum number of cycles is satisfied. The near-optimal subsets are identified by an evaluation function in MOGASVM that uses two criteria: maximisation of 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 near-optimal subsets are 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 our previous work [8].

3.3 Stage 3: Analysing the Frequency of Each Gene in Near-optimal Subsets

The frequency of appearance for each gene in each near-optimal gene subset is examined and analysed to assess the relative importance of genes for cancer classification. The most frequently selected genes in near-optimal gene subsets are presumed to be the most relevant for the classification. Finally, a smaller (final) subset of informative genes (K genes, K is a number of genes) is produced and used to construct an SVM classifier. This subset contains a smaller number of informative genes with higher classification accuracy. This paper has produced two methods of 3-SGS obtained from combinations of two different filter methods (GR and IG) and MOGASVM. These methods are 3-SGS-GR and 3-SGS-IG.

4 Experiments

4.1 Data Sets and Experimental Setup

Two benchmark microarray data sets that contain binary classes and multi-classes of cancer samples are used to evaluate 3-SGS. It is summarised in Table 1. Table 2 contains parameter values for 3-SGS. These values are chosen based on the results of preliminary runs. Three criteria following their importance are considered to evaluate the performance of 3-SGS: test accuracy on the test set, LOOCV accuracy on the training set, 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 3-SGS, and are then used for the second stage.

Table 2. Parameter settings for 3-SGS

Parameter	Leukaemia data set	Lung data set
Size of population	100	100
Number of generation	300	300
Crossover rate	0.7	0.7
Mutation rate	0.01	0.01
Maximum number of cycles	10	10
Cost for an SVM classifier	100	0.7

4.2 Experimental Results

4.2.1 Classification accuracies of final informative genes

As shown in Fig. 2, the best results of the leukaemia (100% LOOCV and 97.06% test accuracies) and the lung (100% LOOCV and 99.33% test accuracies) data sets are obtained by using the only three (using 3-SGS-IG) and nine (using 3-SGS-GR) selected informative genes (K genes), respectively.

Many runs have achieved 100% LOOCV accuracy on all the data sets. This has proved that 3-SGS has efficiently selected and produced a smaller subset of informative genes from a solution space. This is due to the fact that a filter method in the first stage of 3-SGS reduces the dimensionality of the solution space in order to produce a gene subset. Next, MOGASVM in the second stage of 3-SGS optimise the subset automatically to yield near-optimal subsets of genes. These subsets are obtained since MOGASVM in 3-SGS considers and optimises a relation among genes. Finally, the first K genes appearing most frequently are selected as the final selected informative genes for cancer classification.

Fig. 2. A relation between classification accuracies and the number of final selected informative genes (K genes).

Table 3. The list of final selected informative genes in the final gene subsets

Data Set	Score	Gene ID	Gene Description
Leukaemia	5	M23197	CD33 antigen
	2	X95735	Zyxin
	2	X85116	Epb72 gene exon 1
	5	W28612	ESTs
	3	AL050224	polymerase I and transcript release factor
Lung	2	AB020647	f-box and leucine-rich repeat protein 7
	1	X05323	antigen identified by monoclonal antibody
	1	AI201108	related RAS viral (r-ras) oncogene homolog
	1	AL049381	Homo sapiens mRNA; cDNA DKFZp586J2118
	1	S71043	immunoglobulin heavy constant mu
	1	AJ012737	filamin C, gamma (actin-binding protein-280)
	1	Y00318	I factor (complement)

4.2.2 A list of informative genes for biological usage

The informative genes and their rank scores (frequency) of the final subsets as produced by the proposed 3-SGS

and reported in Fig. 2 are listed in Table 3. 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.

4.2.3 3-SGS versus other previous methods

Table 4 displays the benchmark of this work and previous related works that used filter and hybrid approaches. Overall, 3-SGS in this work has outperformed the previous works on all the data sets in terms of test accuracy, LOOCV accuracy, and the number of selected genes.

Generally, filter methods in previous works achieved poor performances since they may result in inclusion of irrelevant and noisy genes in a gene subset for the cancer classification [2-5]. This situation is happen because the methods evaluate a gene based on its discriminative power for the target classes without considering its relations with other genes.

Table 4. The benchmark of 3-SGS with previous methods.

Gene Selection Method (Category) [Reference]	Leukaemia Data Set			Lung Data Set		
	#Selected Genes	Accuracy (%)		#Selected Genes	Accuracy (%)	
		CV	Test		CV	Test
3-SGS (Filter, hybrid, and frequency analysis)	3	100	97.07	9	100	99.33
GASVM (Hybrid) [10]	(3.40)	(100)	-	-	-	-
GASVM (Filter and hybrid) [11]	6	100	-	-	-	-
GASVM (Hybrid) [1]	-	-	-	8	100	98.66
F-test and Cho's method (Filter) [2]	4	98.60	-	49	95.10	-
Signal-to-noise-ratio (Filter) [3]	50	94.74	85.29	-	-	-
t-test (Filter) [4]	-	-	-	4	-	97.32
Information gain (Filter) [5]	-	-	-	-	-	97.99
A Cyclic Approach (Hybrid) [9]	(2.90)	(100)	(88.82)	(2.80)	(100)	(93.69)
GASVM-II+GASVM (Hybrid) [7]	(4.50)	(100)	(85.88)	(2.10)	(100)	(94.16)
GASVM-II (Hybrid) [6]	(10)	(100)	(81.18)	(10)	(100)	(59.33)
MOGASVM (Hybrid) [8]	(2,212.60)	(95.53)	(84.41)	(4,418.5)	(75.31)	(85.84)
GASVM (Hybrid) [6]	(3,574.90)	(94.74)	(83.53)	(6,267.8)	(75)	(84.77)

Note: The results of the best subsets shown in shaded cells. '-' means that a result is not reported in the related previous work. A result in '()' denotes an average result. CV and #Selected Genes represent cross-validation and a number of selected genes, respectively. Methods in *italics* style are experimented in this work.

GASVM-based methods may be unable to produce a smaller subset of informative genes because they perform poorly in higher-dimensional data due to their chromosome representation drawback [1],[6-11]. 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 [6]. On the contrary, the proposed 3-SGS that pre-selects a number of genes at the first stage can reduce the data dimensionality and produce a gene subset. This subset is then optimised by MOGASVM in the second stage of 3-SGS to yield near-optimal subsets. Finally, the first K genes appearing most frequently are selected as the final selected informative genes (a smaller subset) for cancer classification.

The gap between LOOCV accuracy and test accuracy that resulted by 3-SGS 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 related previous works were much higher than their test accuracy because they were unable to avoid or reduce the risk of over-fitting problems [1],[6-9]. Other previous works that used GASVM-based methods did not provide any test accuracy results and thus, the over-fitting problem could not be investigated in their works [10-11]. Over-fitting is a major problem on 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. This is also supported by a review paper in Saeys et al. which reported that hybrid methods (e.g., GASVM-based methods) confront with the high risk of over-fitting problems because of the higher-dimensional data [12].

5 Conclusions

In this paper, a three-stage gene selection method (3-SGS) has been proposed and tested for gene selection on two microarray data sets that contain binary classes and multi-classes of tumour samples. Based on the experimental results, the performance of 3-SGS was superior to other methods in related previous works. This is due to the fact that the filter method in the first stage of the 3-SGS 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 3-SGS can automatically optimise the subset that is yielded by the first stage in order produce near-optimal gene subsets. Finally, the first K genes appearing most frequently are selected as the final selected informative genes (a smaller subset) for cancer classification. Hence, the gene selection using 3-SGS is needed to produce a smaller subset of informative genes for better cancer classification of microarray data. 3-SGS in this paper also obtains short running time because of the large number of genes are removed by a filter technique in the first step. However, due to the application of a filter method in the first stage of 3-SGS, the number of pre-selected genes is difficult since it is manually done. Even though 3-SGS has classified tumours with higher accuracy, it is still not able

to completely avoid the over-fitting problem. Therefore, a combination between a statistical approach and a hybrid method will be proposed to solve the problem.

References

- [1] S. Shah and A. Kusiak: Cancer gene search with data-mining and genetic algorithms, *Computers Biol. Medicine*, 37, pp.251-261, 2007.
- [2] K. Yang, Z. Cai, J. Li, and G. Lin: A stable gene selection in microarray data analysis, *BMC Bioinformatics*, 7, pp.228-246, 2006.
- [3] T.R. Golub, D.K. Slonim, P. Tomayo, C. Huard, M. Gaasenbeek, J.P. Mesirov, H. Coller, M.L. Loh, J.R. Downing, M.A. Caligiuri, C.D. Bloomfield, and E.S. Lander: Molecular classification of cancer: class discovery and class prediction by gene expression monitoring, *Sci.*, 286, pp.531-537, 1999.
- [4] G.J. Gordon, R.V. Jensen, L.L. Hsiao, S.R. Gullans, J.E. Blumenstock, S. Ramaswamy, W.G. Richards, D.J. Sugarbaker, and R. Bueno: Translation of microarray data into clinically relevant cancer diagnostic tests using gene expression ratios in lung cancer and mesothelioma, *Cancer Research*, 62, pp.4963-4967, 2002.
- [5] J. Li, H. Liu, S.K. Ng, and L. Wong: Discovery of significant rules for classifying cancer diagnosis data, *Bioinformatics*, 19, pp.93-102, 2003.
- [6] M.S. Mohamad, S. Deris, and R.M. Illias: A hybrid of genetic algorithm and support vector machine for features selection and classification of gene expression microarray, *Int. J. Comput. Intell. Appl.*, 5, pp.91-107, 2005.
- [7] M.S. Mohamad, S. Omatu, S. Deris, M.F. Misman, and M. Yoshioka: Selecting informative genes from microarray data by using hybrid methods for cancer classification, *Int. J. Artif. Life Rob.*, (Appear in Press)
- [8] M.S. Mohamad, S. Omatu, S. Deris, M.F. Misman, and M. Yoshioka: A multi-objective strategy in genetic algorithm for gene selection of microarray data, *Int. J. Artif. Life Rob.*, (Appear in Press)
- [9] M.S. Mohamad, S. Omatu, M. Yoshioka, and S. Deris: A cyclic hybrid method to select a smaller subset of informative genes for cancer classification, *Int. J. Innovative Comput. Inf. Control*, (Appear in Press)
- [9] H.L. Huang, F.L. Chang: ESVM: Evolutionary support vector machine for automatic feature selection and classification of microarray data, *Biosystems*, 90, pp.516-528, 2007.
- [10] S. Peng, Q. Xu, X.B. Ling, X. Peng, W. Du, and L. Chen: Molecular classification of cancer types from microarray data using the combination of genetic algorithms and support vector machines, *FEBS Lett.*, 555, pp.358-362, 2003.
- [11] Y. Saeys, I. Inza, and P. Larranaga: A review of feature selection techniques in bioinformatics, *Bioinformatics*, 23, pp.2507-2517, 2007.