Artif Life Robotics (2007) 11:219–222
DOI 10.1007/s10015-007-0432-1

ORIGINAL ARTICLE

Mohd Saberi Mohamad · Sigeru Omatu · Safaai Deris
Siti Zaiton Mohd Hashim

A model for gene selection and classification of gene expression data

Received: January 24, 2007 / Accepted: April 23, 2007

Abstract Gene expression data are expected to be of significant help in the development of efficient cancer diagnosis and classification platforms. One problem arising from these data is how to select a small subset of genes from thousands of genes and a few samples that are inherently noisy. This research aims to select a small subset of informative genes from the gene expression data which will maximize the classification accuracy. A model for gene selection and classification has been developed by using a filter approach, and an improved hybrid of the genetic algorithm and a support vector machine classifier. We show that the classification accuracy of the proposed model is useful for the cancer classification of one widely used gene expression benchmark data set.

Key words Gene selection · Hybrid approach · Filter approach · Gene expression data

1 Introduction

Owing to recent advances in biotechnology, gene expression can now be quantitatively monitored on a large scale. Gene expression represents the activation level of each gene within an organism at a particular point of time.[1] Recent studies on the molecular-level classification of tissue have indicated that gene expression data could significantly aid in the development of efficient cancer diagnosis and classification platforms.[2] However, classification based on gene expression data confronts us with more challenges, one of the major ones being the overwhelming number of genes relative to the number of training samples in the data sets.[1] Also, most genes are not relevant to the distinction between different tissue classes, and introduce noise in the classification process.

Gene selection, or feature selection, is the task of selecting a subset of features that maximizes the classifier's ability to classify samples accurately.[3] Gene selection methods 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.[1] Most previous work has 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.[4]

The filter approach has been widely applied by many researchers to select features or genes in various applications. The threshold number of misclassification (TNoM) score,[2] information gain (IG),[5] the signal-to-noise ratio,[6] and the relief algorithm (RA)[7] are some of the widely known filter approaches, and have been successfully applied to gene selection problems. The application of the hybrid approach using the genetic algorithm (GA) and a classifier has grown in recent year.[3] For example, a hybrid of GA and a neural network classifier (GANN), incorporating GA and the support vector machine (SVM) classifier (GASVM), and combining GA via the weight voting classifier, are some of the widely known hybrid approaches, and have been successfully used in various applications.[4] While a large number of supervised and unsupervised methods from the pattern recognition literature have been proposed in bioinformatics research, a method based on the SVM classifier has proven to be the most popular and is reasonably accurate.[4]

A major goal of diagnostic research is to develop a diagnostic procedure based on the least number of possible genes needed to detect diseases.[1] By identifying a small subset of genes on which to base a diagnosis, it is possible to improve classification accuracy. This research deals with selecting a small subset of informative genes from gene

M.S. Mohamad (✉) · S. Deris · S.Z.M. Hashim
Department of Software Engineering, Faculty of Computer Science and Information Systems, Universiti Teknologi Malaysia, 81310 Skudai, Johore, Malaysia
Tel. +60-7-553-2438; Fax +60-7-556-5044
e-mail: {saberi, safaai, sitizaiton}@utm.my

S. Omatu
Department of Computer Science and Intelligent Systems, Graduate School of Engineering, Osaka Prefecture University, Sakai, Osaka, Japan

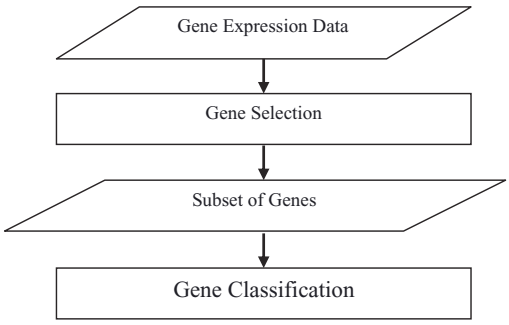

Fig. 1. Model of gene selection and classification

expression data which maximizes the classification accuracy. Hence, this article proposes a model of gene selection and classification using the filter approach, and an improved hybrid of the GA and the SVM classifier (NewGASVM).

2 Model for gene selection and classification

Generally, a model for gene selection and classification has two stages: gene selection and gene classification.[5] Figure 1 shows that this model exhibits a classification stage which includes training and testing phases.

The gene selection method needs to select some genes that are closely related to particular classes for classification; these are called informative genes.[6] This process is called gene selection. The general process of classification is to train a classifier by using training samples, and then classify test samples with the trained classifier.

2.1 Filter approach and hybrid approach for gene selection

Filter approaches, such as IG and the RA are used in this research. The genes with the highest scores are selected as top genes.

Suppose that a gene expression pattern is represented as g_i (for example, i = 1–7129 in leukemia cancer data). Each g_i is a vector of gene expression levels from N samples, $g_i = (e_1, e_2, \ldots, e_N)$, while c_j represents a class of sample j where $j = 1 - N$.[5] If the number of genes excited ($P(g_i)$) or not excited ($P(\bar{g}_i)$) in class ($P(c_j)$) is counted, the coefficient of the IG becomes

$$IG(g_i, c_j) = P(g_i, c_j)\log\frac{P(g_i, c_j)}{P(c_j)\cdot P(g_i)} + P(\bar{g}_i, c_j)\log\frac{P(\bar{g}_i, c_j)}{P(c_j)\cdot P(\bar{g}_i)} \quad (1)$$

The basic idea of the RA is to draw instances at random, compute their nearest neighbors, and adjust a gene weighting vector to give more weight to genes that distinguish this set from neighbors of different classes. Specifically, it tries to find a good estimate of the following probability in order to assign a weight for each gene:[7]

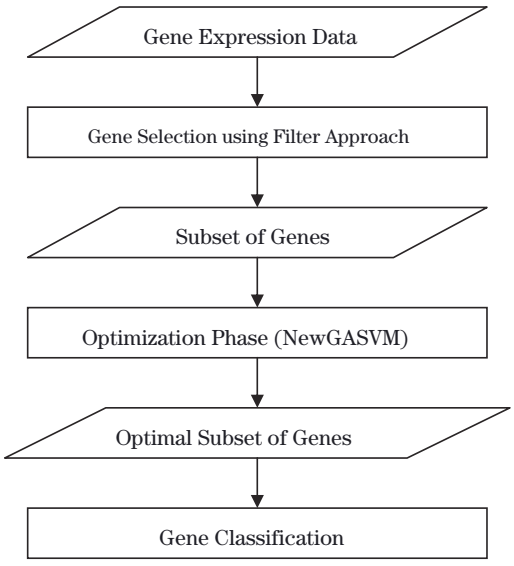

Fig. 2. Proposed model of gene selection and classification

$$W_{gi} = \mathrm{P}(\text{different value of } g_i \mid \text{different class}) - \mathrm{P}(\text{different value of } g_i \mid \text{same class}) \quad (2)$$

This research uses a hybrid approach such as GASVM or NewGASVM. Details of GASVM and NewGASVM can be found in Mohamad et al.[4]

3 Proposed model

This article proposes a model for gene selection and classification using a filter approach and the NewGASVM. Previous work has used a model of gene selection and a classification model, as shown in Fig. 1, which involves two stages. However, our proposed model has three stages: gene selection, gene optimization, and gene classification. Figure 2 shows this model.

The gene selection stage removes irrelevant genes using a filter approach such as IG or the RA. Selecting genes by a filter approach also presents an overall pattern of gene expression data. Therefore, it is a nice starting point for the data analysis. As a result, this stage produces a small subset of genes. The optimization stage selects and optimizes a subset of genes from the small subset by using NewGASVM. GASVM can also be used to replace the NewGASVM at this stage. If the subset is small, the combination of genes is not very complex, and then the NewGASVM can easily find the optimized subset. Moreover, the NewGASVM can also remove noise genes because the filter approach has reduced the size and complexity of the search space. Thus, the NewGASVM is more efficient by using a small subset to complete its task quickly. Lastly, the classification stage builds an SVM classifier using the optimal subset of training sets, and tests it using a test set.

4 Experimental results

4.1 Data set

The leukemia cancer benchmark data set is used to evaluate the proposed model. This data set contains examples of human acute leukemia, originally analyzed by Golub et al.[6] It has the expression levels of 7129 genes and can be obtained at http://www.broad.mit.edu/cgi-bin/cancer/datasets.cgi.

Two criteria are considered to evaluate the performances of the proposed model: the *leave one out cross validation* (LOOCV) accuracy and test accuracy. The LOOCV procedure is applied on training data and the accuracy test measurement on test data to measure the classification accuracy.[4]

4.2 Experimental environment

The experiments were conducted using six methods obtained from a combination of GASVM, NewGASVM, and the filter approaches (IG and RA). Firstly, the GASVM and NewGASVM methods were applied following the stages shown in Fig. 1. Furthermore, by following the stages in Fig. 2, four methods are obtained: IG+GASVM, RA+GASVM, IG+NewGASVM, and RA+NewGASVM. The filter approach was used to select 100 genes from the whole set of genes. Methods based on the NewGASVM were tried using 5, 10, 20, 30, 40, 50, 60, 70, 80, or 90 genes in order to choose the optimal subset of genes. However, methods based on the GASVM, such as GASVM, IG+GASVM, and RA+GASVM, were not tried using the different numbers of selected genes because they were unable to fix the selections.

Figures 3 and 4 show the highest LOOCV and test accuracies for classifying leukemia cancer samples, which are 99.47% and 94.71%, respectively. The IG+NewGASVM method used 20 genes to reach the highest accuracy.

In general, the IG+NewGASVM and RA+NewGASVM methods performed consistently and were much better than the NewGASVM method owing to the application of a filter and hybrid approach in the proposed model. Hence, applying IG+NewGASVM or RA+NewGASVM has improved the accuracy by removing irrelevant genes from whole genes and optimizing the remaining genes. These figures also indicate that the accuracy depends on the number of selected genes.

Table 1 shows the high accuracy of the six methods. In general, the GASVM and NewGASVM methods produced poorer results. In contrast, when a filter approach was applied prior to these methods, the results improved. Hence, the methods that applied the filter approach and NewGASVM out-performed the methods that applied the filter approach and GASVM.

The highest accuracy of the IG+NewGASVM method was 99.47% and 94.71% for LOOCV and test accuracies, respectively, using 20 selected genes in the leukemia cancer data set. On the other hand, the original work of Golub et al.,[6] required about 50 genes to achieve 94.74% for LOOCV accuracy and 85.29% for test accuracy.

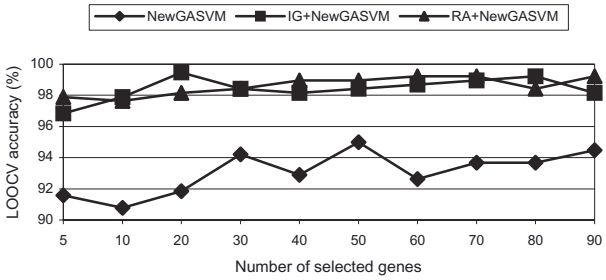

Fig. 3. Correlation between *LOOCV* accuracy and the number of selected genes. *IG*, information gain; *RA*, relief algorithm

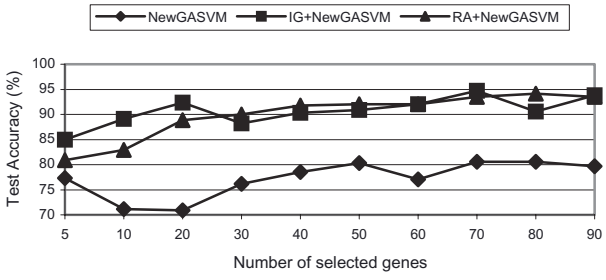

Fig. 4. Correlation between test accuracy and the number of selected genes. *IG*, information gain; *RA*, relief algorithm

Table 1. The benchmark of the highest classification accuracy on leukemia cancer data set

Method	Leukemia cancer data set	
	LOOCV (%)	Test (%)
GASVM	94.74	83.53
IG+GASVM	98.95	93.53
RA+GASVM	97.63	91.76
NewGASVM	95.00	80.59
IG+NewGASVM	99.47	94.71
RA+NewGASVM	99.21	94.12

In general, the IG+NewGASVM and RA+NewGASVM methods performed consistently and with a higher accuracy percentage than other methods because the filter approach was applied before the optimization phase. The filter approach selects and reduces the number of candidate genes from the total number of genes in order to remove irrelevant and noisy genes. Hence, the IG+NewGASVM and RA+NewGASVM methods are more efficient at producing the optimized subset of genes by using the small subset that is produced from the filter approach. However, their performance was much lower than the methods that used a filter approach and the NewGASVM method because it was unable to fix the number of genes selected.

5 Conclusion

In this article, we have designed and applied a new model for gene selection and the classification of gene expression data. Generally, the IG+NewGASVM and RA+NewGASVM methods achieved significant LOOCV and test accuracies, and performed better than other methods because a filter approach was applied before the optimization phase. The filter approach can produce a small subset of genes. Thus, the hybrid method can be more efficient at producing an optimized subset of genes using the small subset that is produced from the filter approach. Hence, applying a filter approach and the hybrid approach in our proposed model is useful because it produces significant classification accuracy. However, this model suffers from a drawback which motivates us to look for further improvements. This limitation is that it has produced inconsistent results when it runs independently in terms of accuracy and the number of selected genes. Even though the proposed model has been successfully applied in the bioinformatics area, it can also be applied and extended in other applications such as robotics, pattern recognition, and computer graphics. We are currently studying a multiobjective strategy with a hybrid method for better optimization of a small subset of genes from thousands of genes.

References

1. Inza I, Larranaga P, Blanco R, et al. (2004) Filter versus wrapper gene selection approaches in DNA microarray domains. J Art Intell Med 31:91–103
2. Ben-Dor A, Bruhn L, Friedman N, et al. (2000) Tissue classification with gene expression profiles. J Comput Biol 7:559–584
3. Zhang P, Verma B, Kumar K (2005) Neural vs statistical classifier in conjunction with genetic algorithm-based feature selection. Patt Recog Lett 26:909–919
4. 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
5. Ryu J, Cho SB (2002) Towards optimal feature and classifier for gene expression classification of cancer. Proceedings of the 2002 AFSS International Conference on Fuzzy Systems (AFSS2002), 2002, Calcutta, India, LNCS 2275 Springer-Verlag, London, UK, pp 310–317
6. Golub TR, Slonim DK, Tomayo P, et al. (1999) Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science 286:531–537
7. Wang Y, Makedon F (2004) Application of Relief-F feature filtering algorithm to selecting informative genes for cancer classification using microarray data. Proceedings of the IEEE Conference on Computational Systems Bioinformatics (CSB'04), 2004, IEEE Press, Stanford, CA, USA, pp 497–498