Artif Life Robotics (2009) 13:410–413
DOI 10.1007/s10015-008-0533-5

ORIGINAL ARTICLE

Mohd Saberi Mohamad · Sigeru Omatu · Safaai Deris
Muhammad Faiz Misman · Michifumi Yoshioka

A multi-objective strategy in genetic algorithms for gene selection of gene expression data

Received and accepted: June 11, 2008

Abstract A microarray machine offers the capacity to measure the expression levels of thousands of genes simultaneously. It is used to collect information from tissue and cell samples regarding gene expression differences that could be useful for cancer classification. However, the urgent problems in the use of gene expression data are the availability of a huge number of genes relative to the small number of available samples, and the fact that many of the genes are not relevant to the classification. It has been shown that selecting a small subset of genes can lead to improved accuracy in the classification. Hence, this paper proposes a solution to the problems by using a multi-objective strategy in a genetic algorithm. This approach was tried on two benchmark gene expression data sets. It obtained encouraging results on those data sets as compared with an approach that used a single-objective strategy in a genetic algorithm.

Key words Cancer classification · Genetic algorithm · Gene expression data · Gene selection · Multi-objective

1 Introduction

Gene expression is a process by which mRNA, and eventually protein, are synthesized from the DNA template of each gene. Recent advances in microarray technology allow scientists to measure the expression levels of thousands of genes simultaneously, and to determine whether those genes are active, hyperactive, or silent in normal or cancerous tissues. This technology finally produces gene expression data. Current studies on the molecular level classification of tissues have produced remarkable results and have indicated that gene expression data could significantly aid in the development of an efficient cancer classification.[1] However, classification based on such data confronts the researcher with more challenges. One of the major challenges is the overwhelming number of genes relative to the small number of samples in a data set. Also, many of the genes are not relevant to the classification process. Hence, the selection of genes is the key to molecular classification, and should be given more attention.

The task of cancer classification using gene expression data is to classify tissue samples into related classes of phenotypes, e.g., cancer versus normal.[2] A gene selection process is used to reduce the number of genes used in the classification while maintaining an acceptable classification accuracy. Gene selection methods can be classified into two categories. If gene selection is carried out independently from the classification procedure, the methods belong to the filter approach. Otherwise, they are said to follow a wrapper (hybrid) approach. Most previous workers have 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.[1] The application of hybrid approaches using a genetic algorithm (GA) with a classifier has grown in recent years. In previous work the GA has performed well, but only on data that have fewer than 1000 features.

Multi-objective optimization (MOO) is an optimization problem that involves multiple objectives or goals. Generally, the objectives estimate different aspects of the solutions. It is necessary to be aware that gene selection is a MOO problem in the sense of classification accuracy maximization and gene subset size minimization. Therefore, this research proposes a multi-objective strategy in a hybrid GA and support vector machine classifier (GASVM) for gene selection and the classification of gene expression data. This is known as MOGASVM.

M.S. Mohamad (✉) · S. Omatu (✉) · M. 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; omatu@cs.osakafu-u.ac.jp

S. Deris · M.F. Misman
Department of Software Engineering, Faculty of Computer Science and Information Systems, Universiti Teknologi Malaysia, Johore, Malaysia

This work was presented in part at the 13th International Symposium on Artificial Life and Robotics, Oita, Japan, January 31–February 2, 2008

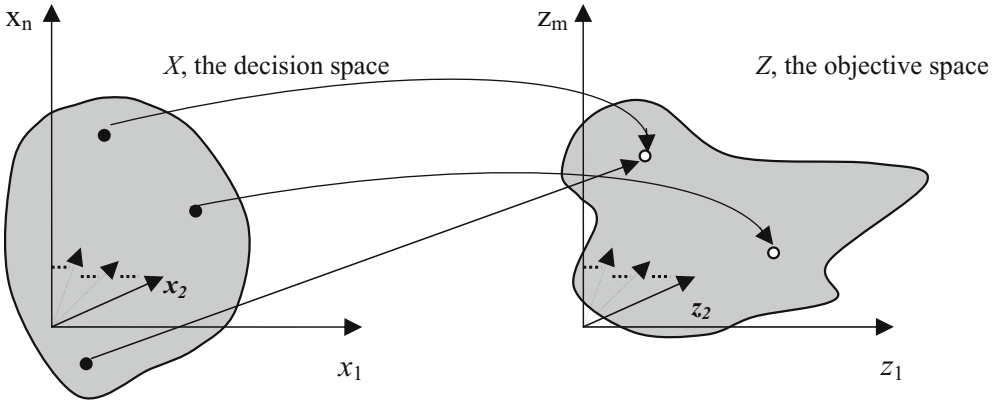

Fig. 1. The *n*-dimensional decision space maps to the *m*-dimensional objective space

2 Multi-objective strategy in GA

MOGASVM was developed to improve the performance of the GASVM that uses a single-objective.[1] All information about GASVM, such as flowcharts, algorithms, chromosome representations, fitness functions, and parameter values, are available in Mohamad et al.[1]

In the sense of classification accuracy maximization and gene subset size minimization, gene selection can be viewed as a MOO problem. Formally, each gene subset (a solution) is represented by x (n-dimensional decision vector). It is associated with a vector objective function $f(x)$:

$$f(x) = (f_1(x), f_2(x), \ldots, f_m(x)) \tag{1}$$

with $x = (x_1, x_2, \ldots, x_n) \in X$ where X is the decision space, i.e., the set of all expressible solutions. The vector objective function $f(x)$ maps X into $\Re^m$, where $\Re$ is the objective space and $m \geq 2$ is a number of objectives. f_i is the ith objective. The vector $z = f(x)$ is an objective vector. The image of X in the objective space is the set of all attainable points z (Fig. 1). If all objective functions are for maximization, a subset x is said to dominate another x (x^*) if and only if:

$x > x^*$ if
$\forall i \in 1..m, f_i(x) \geq f_i(x^*) \wedge \exists j \in 1..m, f_j(x) > f_j(x^*)$

A solution (gene subset) is said to be Pareto optimal if it is not dominated by any other solutions in the decision space. A Pareto optimal solution cannot be improved with respect to any objective without worsening at least one other objective. The set of all feasible nondominated solutions in X is referred to as the Pareto optimal set, and for a given Pareto optimal set, the corresponding objective function values in the objective space are called the Pareto front.[3]

The Pareto front in this research is defined as the set of nondominated gene subsets. MOGASVM is one promising approach to find or approximate the Pareto front. The roles of this approach are guided by the search towards the Pareto front while keeping the nondominated solutions as diverse as possible. Therefore, the original GASVM is customized to accommodate multiobjective problems by using specialized fitness functions. The ultimate goal of MOGASVM is to identify a nondominated gene subset Pareto front. This subset (individual) is evaluated by its accuracy on the training data and the number of genes selected in it. These criteria are denoted as f_1 and f_2 separately, and are used in a fitness function. Therefore, the fitness of an individual is calculated by Eq. 4 as follows:

$$f_1 = w_1 \times A(x) \tag{2}$$

$$f_2 = w_2 \times ((M - R(x))/M) \tag{3}$$

$$fitness(x) = f_1 + f_2 \tag{4}$$

where $A(x) \in [0,1]$ is the leave-one-out-cross-validation (LOOCV) accuracy of the training data using only the expression values of the selected genes in a subset x, where $R(x)$ is the number of selected genes in x. M is the total number of genes, w_1 and w_2 are two priority weights corresponding to the importance of the accuracy and the number of selected genes, respectively, where $w_1 \in [0.1,0.9]$ and $w_2 = 1 - w_1$, and f_2 is calculated as above in order to support the maximization function of the minimization of gene subset size. In this article, accuracy is more important than the number of selected genes (gene subset size).

Ambroise and McLachlan[4] have indicated that because of "selection bias," the test results could be over-optimistic if the test samples were not excluded from the classifier building process in a hybrid approach. Therefore, the proposed MOGASVM totally excludes the test samples from the classifier building process in order to avoid the influence of the bias.

3 Experimental results

3.1 Data sets

Two benchmark data sets are used to evaluate the proposed approach: leukemia cancer and colon cancer. The leukemia cancer data set contains examples of human acute leukemia. It can be obtained at http://www.broad.mit.edu/cgi-bin/cancer/datasets.cgi, while the colon cancer data set can be downloaded at http://microarray.princeton.edu/oncology/. For the leukemia cancer data set, LOOCV is applied on the training set, and an accuracy test measurement is

Table 1. Classification accuracies for different gene subsets using MOGASVM (10 runs on average)

Weight		Average for leukemia data set			Average for colon data set	
w_1	w_2	LOOCV (%)	Test (%)	Number of selected genes	LOOCV (%)	Number of selected genes
0.1	0.9	94.74 ± 0.00	84.12 ± 1.52	2196.5 ± 10.88	90.65 ± 1.27	398.8 ± 6.36
0.2	0.8	95.26 ± 1.11	83.24 ± 2.79	2205.1 ± 15.19	91.45 ± 1.08	419.5 ± 7.95
0.3	0.7	95.00 ± 0.83	83.24 ± 3.12	2199.1 ± 25.83	92.58 ± 0.83	429.2 ± 12.22
0.4	0.6	95.53 ± 1.27	83.53 ± 2.48	2220.8 ± 31.60	92.74 ± 0.85	430.1 ± 10.50
0.5	0.5	95.26 ± 1.11	82.65 ± 3.24	2231.2 ± 26.84	92.90 ± 0.83	443.0 ± 9.19
0.6	0.4	95.26 ± 1.11	82.65 ± 2.93	2210.9 ± 25.09	92.26 ± 0.68	429.0 ± 10.37
0.7	0.3	95.00 ± 0.83	83.24 ± 2.79	2201.4 ± 15.87	**93.23 ± 1.02**	**446.3 ± 18.90**
0.8	0.2	**95.53 ± 1.27**	**84.41 ± 2.42**	**2212.6 ± 26.63**	92.90 ± 1.13	445.9 ± 27.92
0.9	0.1	95.53 ± 1.27	83.82 ± 2.50	2218.3 ± 28.29	92.26 ± 0.68	435.3 ± 12.89

Note: The best results are shown in bold. The colon data set only has LOOCV accuracy since it only has the training set

applied on the testing set to measure classification accuracy. However, for the colon cancer data set, only the LOOCV procedure is used because this data set only has the training set.

3.2 Experimental setup

Three important criteria are used to evaluate the MOGASVM performances: test accuracy, LOOCV accuracy, and the number of selected genes.

The experimental results presented in this section pursue two objectives. The first objective is to show that gene selection using MOGASVM is needed in order to reduce the number of genes and achieve better classification of the gene expression data. The second objective is to show that MOGASVM is better than the original version of GASVM[1] that used a single-objective approach. To achieve these objectives, several experiments were conducted, 10 times each, for both data sets using different values of w_1 and w_2 ($w_1 \in [0.1,0.9]$ and $w_2 = 1 - w_1$). The subset that produces the highest LOOCV accuracy with the lowest number of selected genes is chosen as the best subset. SVM, GASVM (single-objective), and GASVM-II[1] were also used in this research as a comparison with MOGASVM.

3.3 Result analysis and discussion

Table 1 shows the results of the experiments for both data sets using different values of w_1 and w_2. A value of the form $x \pm y$ represents an average value x with a standard deviation y. Overall, the classification accuracy and the number of selected genes for both data sets fluctuated because of the diversity of the solutions based on adjusted weights (w_1 and w_2). Moreover, multiple objectives search simultaneously in a run, and consequently populations tend to converge to the solutions which are superior in one objective, but poor at others. The highest averages of LOOCV and test accuracies for classifying leukemia samples were 95.53% and 84.41%, respectively, using $w_1 = 0.8$ and $w_2 = 0.2$, while 93.23% LOOCV accuracy was obtained for the colon data set using $w_1 = 0.7$ and $w_2 = 0.3$.

A total of 2212.6 average genes in a subset were finally selected to obtain the highest accuracies (LOOCV and test) of the leukemia data set, whereas 446.3 average genes were selected of the colon data set. Hence, these subsets were chosen as the best subsets. This is called the best-known Pareto front because it is close to the true Pareto front. MOGASVM has found the best subsets since it distributed successfully diverse gene subsets over a solution space.

Table 2. The results of the best subset in 10 runs ($w_1 = 0.8$ and $w_2 = 0.2$ of the leukemia data set, $w_1 = 0.7$ and $w_2 = 0.3$ of the colon data set)

Data set	LOOCV (%)	Test (%)	Experiment no.	Number of selected genes
Leukemia	97.37	88.24	4	2252
Colon	95.16	–	7	446

All LOOCV results of the leukemia data set were much higher than the test results due to the problem of over-fitting. The data set properties, i.e., thousands of genes with less than a hundred samples in the training sets, probably cause the over-fitting, where a decision surface of the classifier performs well on the training set, but poorly on the test set.

Table 2 shows that the best performances (LOOCV and test accuracies) were 97.37% and 88.24%, respectively, for the leukemia data set using 2252 genes. For the colon data set, the highest LOOCV accuracy was 93.55% using 446 genes. The best performance for the leukemia data set was found in the fourth experiment, while the best performance for the colon data set was found in the seventh experiment.

In Table 3, the LOOCV accuracy, the test accuracy, and the number of selected genes are given in parentheses. The average results are given and the best result is highlighted. This table shows that the performance of MOGASVM was better than that of GASVM and SVM in terms of the LOOCV accuracy, the test accuracy, and the number of selected genes on average and for the best results. In general, MOGASVM reduces the number of genes to about a quarter of the total, whereas GASVM reduces the number to about a half of the total. This is due to the ability of MOGASVM to search different regions of a solution space simultaneously, and therefore it is possible to find a diverse set of solutions in a high-dimensional space. Moreover, it may also exploit the structures of good solutions with

Table 3. The benchmark of MOGASVM with GASVM (single-objective) and SVM on each data set

Method	Leukemia data set (average; the best)			Colon data set (average; the best)	
	Number of selected genes	Accuracy (%)		Number of selected Genes	LOOCV accuracy (%)
		LOOCV	Test		
MOGASVM	**(2212.6 ± 26.63; 2252)**	**(95.53 ± 1.27; 97.37)**	**(84.41 ± 2.42; 88.24)**	**(446.3 ± 18.90; 446)**	**(93.23 ± 1.02; 95.16)**
GASVM (single-objective)	(3574.9 ± 40.05; 3531)	(94.74 ± 0; 94.74)	(83.53 ± 2.48; 88.24)	(979.8 ± 35.80; 940)	(91.77 ± 0.51; 91.94)
SVM	(7129 ± 0; 7129)	(94.74 ± 0; 94.74)	(85.29 ± 0; 85.29)	(2000 ± 0; 2000)	(85.48 ± 0; 85.48)

Note: The best results are shown in bold

respect to different objectives to create new nondominated solutions in unexplored parts of the Pareto optimal set. This suggests that gene selection using the multi-objective approach is needed for disease classification of gene expression data.

4 Conclusion

MOGASVM has been designed, developed, and analysed to solve gene selection problems. By performing experiments, we found that classification accuracy and the number of selected genes for both data sets fluctuated more and were not equal when using different values of w_1 and w_2 This result shows that there are many irrelevant genes in gene expression data, and some of them act negatively on the accuracy acquired by the relevant genes.

Generally, MOGASVM achieved significant LOOCV accuracy, test accuracy, and the number of selected genes, and was better than GASVM (single-objective) and SVM because its multi-objective strategy could find a diverse solution in a Pareto optimal set. However, MOGASVM did not achieve the greatest accuracy, and the number of selected genes was still high. MOGASVM can also be extended to other applications such as pattern recognitions, computer visions, and cognitive sciences.

References

1. 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 Intel Appl 5:91–107
2. Mohamad MS, Omatu S, Deris S, Hashim SZM (2007) A model for gene selection and classification of gene expression data. Artif Life Robotics 11:219–222
3. Handl J, Kell DB, Knowles J (2007) Multi-objective optimisation in bioinformatics and computational biology. IEEE/ACM Trans Comput Biol Bioinf 4:279–292
4. Ambroise C, McLachlan GJ (2002) Selection bias in gene extraction on the basis of microarray gene-expression data. Proc Natl Acad Sci USA 99(10):6562–6566