Multi-Objective Optimization Using Genetic Algorithm for Gene Selection from Microarray Data

Mohd Saberi Mohamad[1,2], Sigeru Omatu[1], Safaai Deris[2], Michifumi Yoshioka[1]
[1] *Department of Computer Science and Intelligent Systems, Graduate School of Engineering, Osaka Prefecture University, Sakai, Osaka 599-8531, Japan*
(Tel : 81-72-254-9278; Fax : 81-72-257-1788)
mohd.saberi@sig.cs.osakafu-u.ac.jp (Corresponding author)
[2] *Department of Software Engineering, Faculty of Computer Science and Information Systems, Universiti Teknologi Malaysia, 81310 Skudai, Johore, Malaysia*
(Tel : 60-7-553-7784; Fax : 60-7-556-5044)

Abstract

Microarray technology has been increasingly used in cancer research because of its potential for measuring expression levels of thousands of genes simultaneously in tissue samples. It is used to collect the information from tissue samples regarding gene expression differences that could be useful for cancer classification. However, this classification task faces many challenges due to availability of a smaller number of samples compared to the huge number of genes, and many of the genes are not relevant to the classification. It has been shown that selecting a small subset of genes can lead to an improved accuracy of the classification. Hence, this paper proposes a solution to the problem of gene selection by using a multi-objective approach in genetic algorithm. This approach is experimented on two microarray data sets such as Lung cancer and Mixed-Lineage Leukemia cancer. It obtains encouraging result on those data sets as compared with an approach that uses single-objective approach.

I. INTRODUCTION

Gene expression is the 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 expression levels of thousands of genes simultaneously and determine whether those genes are active, hyperactive or silent in normal or cancerous tissues. Furthermore, this technology finally produces gene expression data. This data is also known as microarray data.

Current studies on molecular level classification of tissue have produced remarkable results and indicated that microarray data could significantly aid in the development of an efficient cancer classification.[1] However, classification based on the data confronts with more challenges. One of the major challenges is the overwhelming number of genes relative to the number of samples in the data. Moreover, many of the genes are not relevant to the classification process. Hence, the genes selection is the key of molecular classification and very important.

The task of cancer classification using microarray data is to classify tissue samples into related classes of phenotypes such as cancer and normal.[2] The process of gene selection is to reduce the number of genes used in classification while maintaining acceptable classification accuracy. Gene selection method 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 wrapper (hybrid) approach. Most previous works 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] Application of a hybrid approach using genetic algorithm (GA) with a classifier has grown in recent years.

Multi-objective optimization (MOO) is an optimization problem that involves multiple objectives or goals. Generally, the objectives may estimate very different aspects of the solution. Being aware that gene selection is also a multi-objective optimization problem in the sense of classification accuracy maximization, and gene subset size minimization.

978-1-4244-1692-9/08/$25.00 ©2008 IEEE

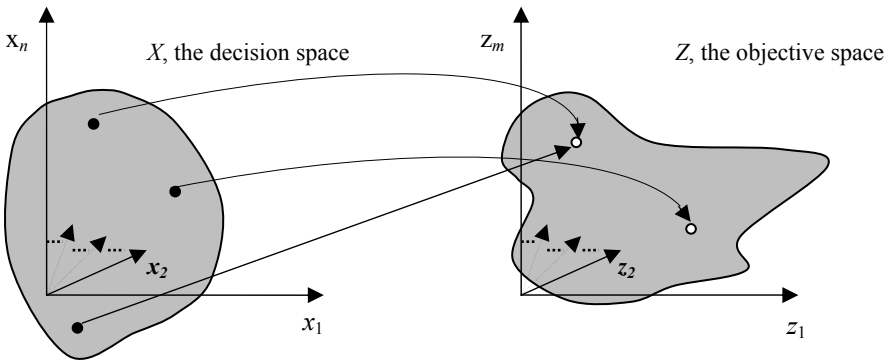

Figure 1. The *n*-dimensional decision space maps to the *m*-dimensional objective space

Therefore, this research proposes a multi-objective approach in a hybrid of GA and support vector machine classifier (GASVM) for gene selection and classification of microarray data. It is known as MOGASVM.

II. MULTI-OBJECTIVE OPTIMIZATION USING GA

MOGASVM is developed to improve the performance of GASVM in previous work[1] that uses single-objective. All information of GASVM such as flowchart, algorithm, chromosome representation, fitness function, 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 multi-objective optimization problem. Formally, each gene subset (a solution) x (*n*-dimensional decision vector) is associated with a vector objective function $f(x)$:

$$f(x) = (f_1(x), f_2(x), ..., f_m(x)) \quad (1)$$

with $x = (x_1, x_2, ..., x_n) \in X$,

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 i^{th} objective. The vector $z = f(x)$ is an objective vector. The image of X in objective space is the set of all attainable points, z (see Fig. 1). If all objective functions are for maximization, a subset x is said to dominate another subset x^* if and only if:

$$x > x^* \text{ iff}$$

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

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 objectives. The set of all feasible non-dominated 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]

Pareto front in this research is defined as the set of non-dominated gene subsets. MOGASVM is one of the promising approaches to find or approximate the Pareto front. The roles of this approach are guided with the search towards the Pareto front and preserving the non-dominated solutions as diverse as possible. Therefore, original GASVM is customized to accommodate multi-objective problem by using specialized fitness function. The ultimate goal of a MOGASVM is to identify a non-dominated 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 used in a fitness function. Therefore, the fitness of an individual is calculated such equation (4):

$$f_1 = w_1 \times A(x) \quad (2)$$

$$f_2 = w_2 \times ((M - R(x)) / M) \quad (3)$$

$$fitness(x) = f_1 + f_2 \quad (4)$$

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

In this paper, accuracy is more important than number of selected genes (gene subset size). Ambroise and Mclachlan (2002) indicated that testing results could be overoptimistic, caused by the "selection bias", if the testing samples were not excluded from the

classifier building process.[4] Therefore, the proposed MOGASVM is totally excluded the testing samples from the classifier building process in order to avoid the influence of bias.

III. EXPERIMENTAL RESULT

A. Data Sets

Two microarray data sets are used to evaluate the proposed approach: Lung cancer and Mixed-Lineage Leukemia (MLL) cancer. The Lung cancer data set has two classes: malignant pleural mesothelioma (MPM) and adenocarcinoma (ADCA). There are 181 samples (31 MPM and 150 ADCA). The training set contains 32 (16 MPM and 16 ADCA) of them. The rest 149 samples are used for testing set. Each sample is described by 12,533 genes. It can be obtained at http://chestsurg.org/publications/2002-microarray.aspx.

The MLL cancer data set is a multi-classes data set. It has three leukemia classes: acute lymphoblastic leukemia (ALL), acute myeloid leukemia (AML), and MLL. The training set contains 57 samples, while the testing set contains 19 samples. There are 12,582 genes in each sample. This data set can be downloaded at http://www.broad.mit.edu/cgi-bin/cancer/datasets.cgi

B. Experimental Setup

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

The experimental results presented in this section pursue two objectives. The first objective is to show that gene selection using MOGASVM is needed for reducing number of genes and in achieving better classification of microarray data. Furthermore, the second objective is to show that the MOGASVM is better than GASVM (single-objective) and SVM. To achieve these objectives, several experiments are conducted 10 times on both data sets using different values of w_1 and w_2. The subset that produces the highest LOOCV accuracy with the smallest number of selected genes is chosen as the best subset.

C. Result Analysis

Table 1 displays 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 average value x with standard deviation y. Overall, classification accuracy and number of selected genes for both data sets were more fluctuating because of the diversity of the solutions based on adjusted weights (w_1 and w_2). Moreover, multiple objectives simultaneously search 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 accuracy and test accuracy for classifying Lung samples were 73.31% and 85.84%, respectively, while 94.74% and 90%, respectively of MLL data set. The highest averages of the accuracies on both data sets were obtained by using $w_1 = 0.7$ and $w_2 = 0.3$. All LOOCV results of both data sets were much higher than the test results due to the problem of over-fitting. The data set properties, i.e., thousand of genes with less than hundred of samples in the training sets can possibly cause the over-fitting. In this problem, a learning of decision function performs well on the training data, but bad on the testing data.

TABLE 1: CLASSIFICATION ACCURACIES FOR DIFFERENT GENE SUBSETS USING MOGASVM (10 RUNS ON AVERAGE)

Weight		Average for Lung Data Set			Average for MLL Data Set		
		Accuracy (%)		Number of	Accuracy (%)		Number of
w_1	w_2	LOOCV	Test	Selected Genes	LOOCV	Test	Selected Genes
0.1	0.9	75 ± 0	84.43 ± 4.16	4,416.5 ± 17.90	94.74 ± 0	88.67 ± 5.49	4,472.1 ± 29.40
0.2	0.8	75 ± 0	85.24 ± 4.68	4,421.3 ± 21.53	94.74 ± 0	89.33 ± 4.66	4,470.6 ± 16.54
0.3	0.7	75 ± 0	84.16 ± 3.79	4,416.6 ± 13.59	94.74 ± 0	88.67 ± 7.06	4,466.9 ± 21.25
0.4	0.6	75 ± 0	81.75 ± 4.30	4,410.3 ± 26.30	94.74 ± 0	89.33 ± 4.66	4,471.4 ± 19.50
0.5	0.5	75 ± 0	84.10 ± 4.78	4,415.7 ± 25.40	94.74 ± 0	89.33 ± 5.62	4,465.3 ± 24.60
0.6	0.4	75 ± 0	84.90 ± 4.04	4,423.2 ± 19.62	94.74 ± 0	88.67 ± 3.22	4,479.2 ± 21.73
0.7	0.3	75.31 ± 0.99	85.84 ± 3.97	4,418.5 ± 50.19	94.74 ± 0	90.00 ± 3.51	4,465.2 ± 18.34
0.8	0.2	75 ± 0	83.22 ± 4.86	4,419 ± 15.25	94.74 ± 0	88.00 ± 6.13	4,479.3 ± 22.24
0.9	0.1	75 ± 0	83.83 ± 4.30	4,423.3 ± 19.66	94.74 ± 0	88.00 ± 6.13	4,468.4 ± 16.03

Note: Best result shown in shaded cells.

TABLE 2: RESULT OF THE BEST SUBSETS IN 10 RUNS USING $w_1 = 0.7$ AND $w_2 = 0.3$ ON MOGASVM

Data set	LOOCV (%)	Test (%)	Experiment No.	Number of Selected Genes
Lung	78.13	93.29	7	4,433
MLL	94.74	93.33	7	4,437

4418.5 genes (average) in a subset were finally selected to obtain the highest accuracies (LOOCV and test) of Lung data set, whereas 4465.2 genes (average) of MLL data set. Hence, these subsets were being chosen as the best subsets for Lung and MLL data sets, respectively. It is called best-known Pareto front because it is close to the true Pareto front. MOGASVM could obtain the best subsets since it successfully distributed diverse gene subsets over solution space.

Table 2 shows that the best performances (LOOCV and test accuracies) were 78.13% and 93.29%, respectively for Lung data set using 4433 genes, while 94.74% and 93.33%, respectively for MLL data set by using 4437 genes. The best performances, for both data sets have been found in the seventh experiment.

In Table 3 shows that the performance of MOGASVM was better than GASVM and SVM in terms of LOOCV accuracy, test accuracy, and number of selected genes on average result and the best result. In general, MOGASVM has reduced about a third of the total number of genes, whereas about a half of GASVM. This is due to the ability of the MOGASVM to simultaneously search different regions of a solution space and therefore it is possible to find a diverse set of solution in higher dimensional space. Moreover it may also exploit structures of good solutions with respect to different objectives to create new non-dominated solutions in unexplored parts of the Pareto optimal set. This suggests that gene selection using multi-objective approach in GASVM is needed for cancer classification of microarray data.

IV. CONCLUSION

This paper has investigated the important issues of selection a subset of genes from thousands of genes measured on microarray. A MOGASVM is designed, developed, and analyzed to solve the issues on two benchmark microarray data sets. This research found that classification accuracy and number of selected genes for both data sets were more fluctuating when using different values of w_1 and w_2. This result concludes that there are many irrelevant genes in gene expression data and some of them act negatively on the acquired accuracy by the relevant genes.

From the experimental results, generally, the MOGASVM achieved significant LOOCV accuracy, test accuracy, and number of selected genes, and were better than GASVM and SVM since the multi-objective approach in it can find a diverse solution in Pareto optimal set. However, the number of selected genes using MOGASVM was still higher. Thus, it needs further research to reduce the number. MOGASVM can also be extended to other applications such as computer vision and cognitive science.

REFERENCES

[1] 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", *International Journal of Computational Intelligence and Applications*, Imperial College Press, 5, 2005, pp. 91–107.

[2] M.S. Mohamad, S. Omatu, S. Deris, and S.Z.M. Hashim, "A Model for Gene Selection and Classification of Gene Expression Data", *International Journal of Artificial Life & Robotics*, Springer, 11(2), 2007, pp. 219–222.

[3] J. Handl, D.B. Kell, and J. Knowles, "Multi-objective Optimisation in Bioinformatics and Computational Biology", *IEEE/ACM Transaction on Computational Biology and Bioinformatics*, IEEE, 4(2), 2007, pp. 279–292.

[4] C. Ambroise, and G.J. McLachlan, "Selection Bias in Gene Extraction on the Basis of Microarray Gene-expression Data", *Proceedings of the National Academy of Science of the USA*, USA, 99 (10), May 14, 2002, pp. 6562–6566.

TABLE 3: BENCHMARK OF THE MOGASVM WITH GASVM AND SVM

Method	Lung Data Set (Average; The Best)			MLL Data Set (Average; The Best)		
	Number of Selected Genes	Accuracy (%)		Number of Selected Genes	Accuracy (%)	
		LOOCV	Test		LOOCV	Test
MOGASVM	(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)	(90.00 ± 3.51; 93.33)
GASVM (single-objective)	(6,267.8 ± 56.34; 6,342)	(75.00 ± 0; 75.00)	(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	(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: Best result shown in shaded cells.