A Multi-objective Strategy in Genetic Algorithm for Gene Selection of Gene Expression Data

M.S. Mohamad[1] S. Omatu[1] S. Deris[2] and M.F. Misman[2]

[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; omatu@cs.osakafu-u.ac.jp)

[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)
(safaai@utm.my; faizmisman@gmail.com)

Abstract: Microarray device offers the ability 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 issues in the use of gene expression data are the availability of huge number of genes relative to the small number of available samples, 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 strategy in genetic algorithm. This approach is experimented on two benchmark gene expression data sets and it presented the experimental results. It obtains encouraging result on those data sets as compared with an approach that uses single-objective strategy in genetic algorithm.

Keywords: Multi-objective, Genetic Algorithm, Gene Selection, Cancer Classification, Gene Expression Data.

I. INTRODUCTION

Gene expression is the process by which mRNA and eventually protein are synthesised 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. Current studies on molecular level classification of tissue have produced remarkable results and indicated that gene expression 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 sets. Moreover, many of the genes are not relevant to the classification process. Hence, the selection of genes is the key of molecular classification, and should be taken with more attention.

The task of cancer classification using gene expression data is to classify tissue samples into related classes of phenotypes such as cancer versus 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 hybrid approach using genetic algorithm (GA) with a classifier has grown in recent years. From the previous works, the GA performed well but only on the data that have number of features that are less than 1000.

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

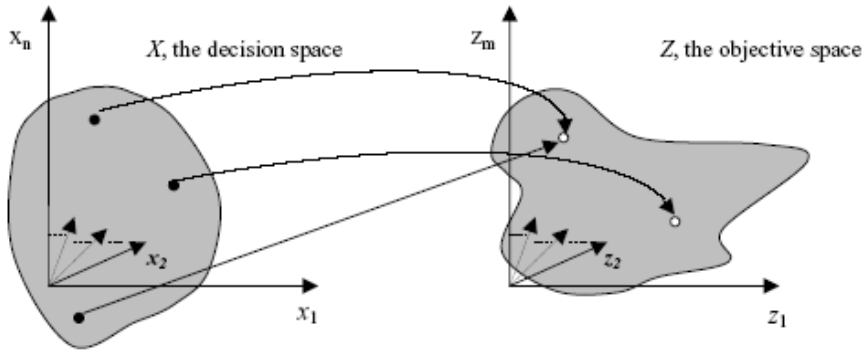

Fig.1. The n-dimensional parameter space maps to the m-dimensional objective space

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

II. MULTI OBJECTIVE STRATEGY IN GA

MOGASVM is developed to improve the performance of GASVM in previous work that uses single-objective.[1] 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 maximisation and gene subset size minimisation, gene selection can be viewed as a multi-objective optimisation 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 maximisation, a subset x is said to dominate than another subset x^* if and only if:

$x > x^*$ 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 solution 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 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 customised to accommodate multi-objective problems by using specialised fitness functions. 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):

$$f_1 = w1 \times A(x) \quad (2)$$

$$f_2 = w2 \times ((M - R(x))/M) \quad (3)$$

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

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 subset x, $R(x)$ is the number of selected genes in x. M is the total number of genes. $w1$ and $w2$ are two weights corresponding to the importance of accuracy and the number of selected genes respectively, $w1 \in [0,0.9]$ and $w2 = 1 - w1$. Formula of f_2 is calculated such above in order to support the maximisation function of minimisation of gene subset size. 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.[7] 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 RESULTS

3.1. Data Sets

Two benchmark data sets are used to evaluate the proposed algorithm: Leukemia cancer and Colon cancer. 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 Colon cancer data set can be downloaded at http://microarray.princeton.edu/oncology/. For Leukemia cancer data set, LOOCV procedure is applied on training data, and accuracy test measurement is applied on testing data to measure classification accuracy. While for Colon cancer data set, only LOOCV procedure is used because this data set only has training data.

3.2. Experimental Setup

Three criteria following its important are used to evaluate the MOGASVM performances: test accuracy, LOOCV accuracy and 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 for reducing number of genes and in achieving better classification of gene expression data. Furthermore, the second objective is to show that the MOGASVM is better than the original version of GASVM that uses the single-objective approach. To achieve these objectives, several experiments are conducted five times for both data sets using different value of *w1* and *w2* ($w1 \in [0,0.9]$ and $w2 = 1 - w1$). Moreover, SVM and GASVM are also performed in this research for comparison with the MOGASVM.

3.3. Result Analysis and Discussion

Table 1 displays results of the experiments for both data sets using different values of *w1* and *w2*. 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 (*w1* and *w2*). 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 and test accuracies for classifying Leukemia samples were 95.79% and 85.88% respectively, using *w1*=0.8 and *w2*=0.2, while 93.55% LOOCV accuracy using *w1*=0.7 and *w2*=0.3 for Colon data set.

2225.4 genes in a subset were finally selected to obtain the highest accuracies of Leukemia data set, whereas 455.2 genes were of Colon data set. Hence, these subsets were being chosen as the best subsets for Leukemia and Colon 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 1. Classification accuracies for different gene subsets using MOGASVM (five runs on average)

Weight		Average for Leukemia Data Set			Average for Colon Data Set	
w1	w2	LOOCV (%)	Test (%)	Number of Selected Genes	LOOCV (%)	Number of Selected Genes
0.1	0.9	94.74 ± 0	84.12 ± 1.61	2197.4 ± 11.46	90.97 ± 1.44	398.6 ± 398.60
0.2	0.8	95.26 ± 1.18	83.53 ± 3.35	2203.6 ± 18.82	90.97 ± 1.44	414.8 ± 414.80
0.3	0.7	95.26 ± 1.18	83.53 ± 3.35	2209.0 ± 29.40	92.90 ± 0.88	433 ± 433.00
0.4	0.6	95.26 ± 1.18	84.12 ± 2.63	2218.8 ± 37.71	93.22 ± 0.72	434.2 ± 434.20
0.5	0.5	96.32 ± 1.44	81.76 ± 1.32	2230.6 ± 28.46	93.22 ± 0.72	443.2 ± 443.20
0.6	0.4	94.74 ± 0	82.35 ± 3.60	2200.4 ± 22.50	92.58 ± 0.88	432.8 ± 432.80
0.7	0.3	95.26 ± 1.18	83.53 ± 1.61	2195.8 ± 10.09	93.55 ± 0	455.2 ± 455.20
0.8	0.2	95.79 ± 1.44	85.88 ± 2.46	2225.4 ± 22.46	92.58 ± 0.88	443.6 ± 443.60
0.9	0.1	95.26 ± 1.18	84.12 ± 2.63	2211.6 ± 31.18	92.26 ± 0.72	433.8 ± 433.80

Note: Best result shown in shaded cells.

All LOOCV results of Leukemia data set were much higher than the test results due to the problem of overfitting. The data set properties, i.e., thousand of genes with less than hundred of samples in the training sets can possibly cause the overfitting which learning a decision surface that performs well on the training data but bad on the testing data.

Table 2. Results of the best subset in five runs

Data set	LOOCV %	Test %	Experiment No.	Number of Genes
Leukemia	97.37	88.24	4	2252
Colon	93.55	-	All	438,467,438, 473,460

Table 2 shows that the best performances (LOOCV and test accuracies) were 97.37% and 88.24% respectively for Leukemia data set using 2252 genes. For Colon data set, the highest LOOCV accuracy was 93.55 % using 438, 467, 473 or 460 genes. The best performance for Leukemia data set has been found in the fourth experiment, while for Colon data set the best performances are found in all experiments.

In table 3, LOOCV accuracy, test accuracy and number of selected genes are written in the parenthesis; the first and second parts are average and showcased the best results respectively. This table 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 quarter of 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 high 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 is needed for disease classification of gene expression 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 analysed to solve the issues on two benchmark gene expression data sets. By performing experiments, this research found that classification accuracy and number of selected genes for both data sets were more fluctuating and not equal when using different values of *w1* and *w2*. 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 strategy in it can find a diverse solution in Pareto optimal set. However, MOGASVM did not achieve higher accuracy, and the number of selected genes was still higher. MOGASVM can also be extended to other applications such as pattern recognition, computer vision and cognitive science.

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 Intell Appl 5:1–17

[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(2):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(2):279–292

[4] Ambroise C, McLachlan GJ (2002) Selection bias in gene extraction on the basis of microarray gene-expression data. Proc Nat Aca Sci USA 6562–6566

Table 3. 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	(2225.4, 2199)	(95.79, 97.37)	(85.88, 88.24)	(455.2, 438)	(93.55, 93.55)
GASVM (Single-objective)	(3580.6, 3535)	(94.74, 94.74)	(82.94, 85.29)	(991.4, 957)	(91.61, 91.94)
SVM	(7129, 7129)	(94.74, 94.74)	(85.29, 85.29)	(2000, 2000)	(85.48, 85.48)

Note: Best result shown in shaded cells.