GENE SUBSET SELECTION FOR LUNG CANCER CLASSIFICATION USING A MULTI-OBJECTIVE STRATEGY

Mohd Saberi Mohamad[1,2], Sigeru Omatu[1], Safaai Deris[2] and Michifuci Yoshioka[1]

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

[2]Department of Software Engineering,
Faculty of Computer Science and Information System,
Universiti Teknologi Malaysia, 81310 Skudai, Johor, Malaysia

Email: mohd.saberi@sig.cs.osakafu-u.ac.jp, {sigeru,yoshioka}@.cs.osakafu-u.ac.jp,
safaai@utm.my

Abstract: A microarray machine offers the ability to measure the expression levels of thousands of genes simultaneously. It is used to collect the 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 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 classification accuracy. Hence, this paper proposes a solution to the problems by using a multi-objective strategy in genetic algorithms. This approach is experimented on one gene expression data set, namely the lung cancer. It obtains encouraging result on the data set as compared with an approach that uses single-objective strategy in genetic algorithms.

Keywords: 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 synthesised 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 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 tissue have produced remarkable results and indicated that gene expression data could

significantly aid in the development of an efficient cancer classification [3]. 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 a data set. Many of the genes are also 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, e.g., cancer versus normal [4]. A gene selection process is used to reduce the number of genes used in 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, 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 [3]. The application of hybrid approaches using genetic algorithm (GA) with a classifier has grown in recent years. From the previous works, the GA performed well but only on data that have a number of features that is less than 1,000.

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

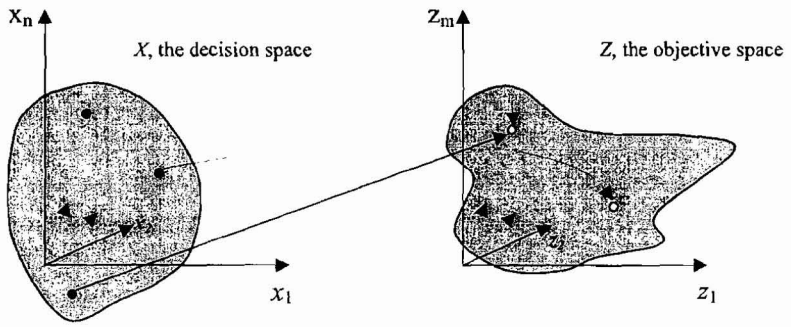

Figure 1. The n-dimensional decision space maps to the m-dimensional objective space.

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.

2. A MULTI-OBJECTIVE STRATEGY IN GA

MOGASVM is developed to improve the performance of GASVM that uses single-objective [3]. All information of GASVM such as flowchart, algorithm, chromosome representation, fitness function, and parameter values are available in Mohamad *et al.* [3].

In the sense of classification accuracy maximisation and gene subset size minimisation, a 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 $f(x)$:

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

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 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 (see Fig. 1). If all objective functions are for maximisation, a subset x is said to dominate than another x (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 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 as the Pareto front [2].

Pareto front in this research is defined as the set of non-dominated gene subsets. MOGASVM is one of promising approaches to find or approximate the Pareto front. The role of this approach is 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 a specialised fitness function. The ultimate goal of 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 the fitness function. Therefore, the fitness of an individual is calculated such equation (4):

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

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

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

where $A(x) \in [0,1]$ is the leave-one-out-cross-validation (LOOCV) accuracy on the training data using the only 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 priority weights corresponding to the importance of accuracy and the number of selected genes, respectively, where $w_1 \in [0.1, 0.9]$ and $w_2 = 1 - w_1$. f_2 is calculated such above in order to support the maximisation function of minimisation of gene subset size. In this paper, the accuracy is more important than the number of selected genes (gene subset size).

Ambroise and Mclachlan (2002) have indicated that testing results could be overoptimistic, caused by the "selection bias", if the test samples were not excluded from the classifier building process in a hybrid approach [1]. Therefore, the proposed MOGASVM is totally excluded the test samples from the classifier building process in order to avoid the influence of bias.

3. EXPERIMENTAL RESULTS

3.1 Data Sets

One gene expression data set is used to evaluate the proposed approach, namely the lung 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 the test set. Each sample is described by 12,533 genes. It can be obtained at http://chestsurg.org/publications/2002-microarray.aspx.

3.2 Experimental setup

Three criteria following its important are used to evaluate MOGASVM performances: the test accuracy, the 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 for reducing the number of genes and achieving better classification of gene expression data. The second objective is to show that MOGASVM is better than the original version of GASVM [3] that use a single-

objective approach. To achieve these objectives, several experiments are conducted 10 times 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 [3] are also experimented in this research as a comparison with MOGASVM.

3.3 Result analysis and discussion

Table 1 displays results of the experiments for the lung cancer data set 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, classification accuracy and the number of selected genes sets were 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 lung samples were 75.31% and 85.84%, respectively, using $w_1 = 0.7$ and $w_2 = 0.3$.

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

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

Weight		Average for the Lung Data Set		
w_1	w_2	Accuracy (%)		Number of Selected Genes
		LOOCV	Test	
0.1	0.9	75 ± 0	84.43 ± 4.16	4,416.5 ± 17.90
0.2	0.8	75 ± 0	85.24 ± 4.68	4,421.3 ± 21.53
0.3	0.7	75 ± 0	84.16 ± 3.79	4,416.6 ± 13.59
0.4	0.6	75 ± 0	81.75 ± 4.30	4,410.3 ± 26.30
0.5	0.5	75 ± 0	84.10 ± 4.78	4,415.7 ± 25.40
0.6	0.4	75 ± 0	84.90 ± 4.04	4,423.2 ± 19.62
0.7	0.3	75.31 ± 0.99	85.84 ± 3.97	4,418.5 ± 50.19
0.8	0.2	75 ± 0	83.22 ± 4.86	4,419 ± 15.25
0.9	0.1	75 ± 0	83.83 ± 4.30	4,423.3 ± 19.66

Note: Best result shown in shaded cells.

Table 2. The result of the best subset in 10 runs (using $w_1 = 0.8$ and $w_2 = 0.3$).

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

Table 2 shows that the best performances (LOOCV and test accuracies) were 78.13% and 93.29%, respectively using 4433 genes. The best performances have been found in the seventh experiment.

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

Method	Lung Data Set (Average; The Best)		
	Number of Selected Genes	Accuracy (%)	
		LOOCV	Test
MOGASVM	(4,418.5 ± 50.19; 4,433)	(75.31 ± 0.99; 78.13)	(85.84 ± 3.97; 93.29)
GASVM (single-objective)	(6,267.8 ± 56.34; 6,342)	(75.00 ± 0; 75.00)	(84.77 ± 2.53; 87.92)
SVM	(12,533 ± 0; 12,533)	(65.63 ± 0; 65.63)	(85.91 ± 0; 85.91)

Note: Best result shown in shaded cells.

In table 3, the LOOCV accuracy, the test accuracy, and the number of selected genes are written in the parenthesis; the first and second parts are the average result and showcased the best result, respectively. This table shows that the performance of MOGASVM was better than GASVM and SVM in terms of LOOCV accuracy, test accuracy, and the number of selected genes on average and the best results. In general, MOGASVM has reduced about three-quarters of the total number of genes, whereas about a half of GASVM. This is due to the ability of MOGASVM to simultaneously search different regions of a solution space and therefore it is possible to find a diverse set of solution in a 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 the multi-objective approach is needed for disease classification of gene expression data.

4. CONCLUSIONS

MOGASVM has been proposed, developed, and analysed to solve the gene selection problems. By performing experiments, this research found that classification accuracy and the number of selected genes were more fluctuating and not equal 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.

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

ACKNOWLEDGEMENTS

This study was supported and approved by Universiti Teknologi Malaysia, Osaka Prefecture University, and Malaysian Ministry of Higher Education. The authors gratefully thank the referees for the helpful suggestions.

REFERENCES

[1] Ambroise, C. and McLachlan, G. J., "Selection bias in gene extraction on the basis of microarray gene-expression data", Proceedings of the National Academy of Science of the USA, Washington, USA, Volume 99, Issue 10, pp 6562–6566, 2002.

[2] Handl, J., Kell, D. B. and Knowles, J, "Multi-objective optimisation in bioinformatics and computational biology", IEEE/ACM Transaction on Computational Biology & Bioinformatics, Volume 4, Issue 2, pp. 279–292, 2007.

[3] Mohamad, M. S., Deris, S. and Illias, R. M., "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, Volume 5, pp.91–107, 2005.

[4] Mohamad, M. S., Omatu, S., Deris, S. and Hashim, S. Z. M., "A model for gene selection and classification of gene expression data", International Journal of Artificial Life & Robotics, Volume 11, Issue 2, pp.219–222, 2007.