Informative Gene for Cancer Classification by Using Particle Swarm Optimization

M.S. Mohamad[1,2] S. Omatu[1] S. Deris[2] and M. 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; omatu@cs.osakafu-u.ac.jp; yoshioka@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)*

Abstract: Microarrays technology offers the ability to measure the expression levels of thousands of genes simultaneously in biological organisms. Gene expression data that produced by the technology are expected to be of significant help in the development of efficient cancer diagnoses and classification platforms. The main problem that needs to be addressed is the selection of a small subset of genes from the thousands of genes in the data that contributes to a cancer disease. Therefore, this article proposes particle swarm optimization (PSO) with the constraint of particle's velocities to select a near-optimal (small) subset of informative genes that is relevant for cancer classification. The performance of the proposed method was evaluated by two well-known gene expression data sets and obtained encouraging results as compared with the standard version of binary PSO.

Keywords: Binary particle swarm optimization, Gene selection, Gene expression data, Optimization.

I. INTRODUCTION

Advances in microarray technology allow scientists to measure the expression levels of thousands of genes simultaneously in biological organisms and have made it possible to create databases of cancerous tissues. It finally produces gene expression data that contain useful information of genomic, diagnostic, and prognostic for researchers [1]. Thus, there is a need to select informative genes that contribute to a cancerous state [2]. However, the gene selection process poses a major challenge because of the following characteristics of the data: the huge number of genes compared to the small number of samples (high-dimensional data), irrelevant genes, and noisy data. To overcome this challenge, a gene selection method is used to select a subset of informative genes that maximizes classifier's ability to classify samples more accurately [3]. In computational intelligence domains, gene selection is called feature selection.

Recently, several gene selection methods based on particle swarm optimization (PSO) have been proposed to select informative genes from gene expression data [4],[5]. PSO is a new evolutionary technique proposed by Kennedy and Eberhart [6]. Shen *et al.* [4] have proposed a hybrid of PSO and tabu search approaches for gene selection. However, the results obtained by using the hybrid method are less meaningful since the application of tabu approaches in PSO is unable to search a near-optimal solution in search spaces. Next, Li *et al.* [5] have introduced a hybrid of PSO and genetic algorithms (GA) for the same purpose. Unfortunately, the accuracy result is still not high and many genes are selected for cancer classification since there are no direct probability relations between GA and PSO. Generally, the PSO-based methods are intractable to efficiently produce a small (near-optimal) subset of informative genes for high classification accuracy [4],[5]. This is mainly because the total number of genes in gene expression data is too large (high-dimensional data).

The diagnostic goal is to develop a medical procedure based on the least number of possible genes that needed to detect diseases. Thus, we introduce an enhancement of binary PSO based on the proposed constraint and rule (CPSO) to select a small (near-optimal) subset of informative genes that is most relevant for the cancer classification. The small subset means that it contains the small number of selected genes. In order to test the effectiveness of our proposed method, we apply CPSO to two gene expression data sets.

II. THE STANDARD VERSION OF BINARY PSO (BPSO)

BPSO is initialized with a population of particles. At each iteration, all particles move in a problem space to find the optimal solution. A particle represents a potential solution in an *n*-dimensional space [7]. Each particle has position and velocity vectors for directing its movement. The position vector and velocity vector of the *i*th particle in the *n*-dimension can be represented as $X_i = (x_i^1, x_i^2, ..., x_i^n)$ and $V_i = (v_i^1, v_i^2, ..., v_i^n)$, respectively, where $x_i^d \in \{0,1\}$; *i*=1,2,..*m* (*m* is the total number of particles); and *d*=1,2,..*n* (*n* is the dimension of data). v_i^d is a real number for the *d*-th dimension of the particle *i*, where the maximum v_i^d, $V_{\max} = (1/3)\times n$.

In gene selection, the vector of particle positions is represented by a binary bit string of length *n*, where *n* is the total number of genes. Each position vector (X_i) denotes a gene subset. If the value of the bit is 1, it means that the corresponding gene is selected. Otherwise, the value of 0 means that the corresponding gene is not selected. Each particle in the *t*-th iteration updates its own position and velocity according to the following equations:

$$v_i^d(t+1) = w(t)\times v_i^d(t) + c_1 r_1^d(t)\times(pbest_i^d(t) \\ -x_i^d(t)) + c_2 r_2^d(t)\times(gbest^d(t) - x_i^d(t)) \tag{1}$$

$$Sig(v_i^d(t+1)) = \frac{1}{1+e^{-v_i^d(t+1)}} \tag{2}$$

$$\text{if } Sig(v_i^d(t+1)) > r_3^d(t), \text{ then } x_i^d(t+1) = 1; \\ \text{else } x_i^d(t+1) = 0 \tag{3}$$

where c_1 and c_2 are the acceleration constants in the interval [0,2]. $r_1^d(t), r_2^d(t), r_3^d(t) \sim U(0,1)$ are random values in the range [0,1] that sampled from a uniform distribution. $Pbest_i(t) = (pbest_i^1(t), pbest_i^2(t), ..., pbest_i^n(t))$ and $Gbest(t) = (gbest^1(t), gbest^2(t), ..., gbest^n(t))$ represent the best previous position of the *i*th particle and the global best position of the swarm (all particles), respectively. They are assessed base on a fitness function. $Sig(v_i^d(t+1))$ is a sigmoid function where $Sig(v_i^d(t+1)) \in [0,1]$. $w(t)$ is an inertia weight.

III. AN IMPROVEMENT OF BINARY PSO BASED ON THE CONSTRAINT OF PARTICLE'S VELOCITIES (CPSO)

We propose CPSO for selecting a near-optimal (small) subset of genes. It is proposed to overcome the limitations of BPSO and previous PSO-based methods [4],[5]. CPSO in our work differs from BPSO and the PSO-based methods on two parts: 1) we propose the constraint of elements of particle velocity vectors; 2) we introduce a rule for updating $x_i^d(t+1)$, whereas BPSO and the PSO-based methods have used the original rule (Eq. 3) and no constraint of elements of particle velocity vectors. The constraint and rule are introduced in order to:

- increase the probability of $x_i^d(t+1) = 0$ $(P(x_i^d(t+1) = 0))$.
- reduce the probability of $x_i^d(t+1) = 1$ $(P(x_i^d(t+1) = 1))$.

The increased and decreased probability values cause a small number of genes are selected and grouped into a gene subset. $x_i^d(t+1) = 1$ means that the corresponding gene is selected. Otherwise, $x_i^d(t+1) = 0$ represents that the corresponding gene is not selected. The constraint of elements of particle velocity vectors and the rule are proposed as follows:

$$Sig(v_i^d(t+1)) = \frac{1}{1+e^{-v_i^d(t+1)}} \tag{4}$$

subject to $v_i^d(t+1) \geq 0$

$$\text{if } Sig(v_i^d(t+1)) > r_3^d(t), \text{ then } x_i^d(t+1) = 0; \\ \text{else } x_i^d(t+1) = 1 \tag{5}$$

The constraint of elements of particle velocity vectors and the rule increase $P(x_i^d(t) = 0)$ because the minimum value for $P(x_i^d(t) = 0)$ is 0.5 when $v_i^d(t) = 0$ $(\min P(x_i^d(t) = 0) \geq 0.5)$. Meanwhile, they decrease the maximum value for $P(x_i^d(t) = 1)$ to 0.5 $(\max P(x_i^d(t) = 1) \leq 0.5)$. Therefore, if $v_i^d(t) > 0$, then $P(x_i^d(t) = 0) >> 0.5$ and $P(x_i^d(t) = 1) << 0.5$. For example, the calculations for $P(x_i^d(t) = 0)$ and $P(x_i^d(t) = 1)$ are shown as follows:

if $v_i^d(t) = 1$, then $P(x_i^d(t) = 0) = 0.73$ and $P(x_i^d(t) = 1) = 1 - P(x_i^d(t) = 0) = 0.27$.

if $v_i^d(t) = 2$, then $P(x_i^d(t) = 0) = 0.88$ and $P(x_i^d(t) = 1) = 1 - P(x_i^d(t) = 0) = 0.12$.

The fitness value of a particle (a gene subset) is calculated as follows:

$$fitness(X_i) = w_1 \times A(X_i) + (w_2 \times (n - R(X_i))/n) \tag{6}$$

in which $A(X_i) \in [0,1]$ is leave-one-out-cross-validation (LOOCV) classification accuracy that uses the only genes in a gene subset (X_i). This accuracy is provided by support vector machine classifiers (SVM). $R(X_i)$ is the number of selected genes in X_i. n is the total number of genes for each sample. 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 $w_2 = 1 - w_1$.

IV. EXPERIMENTS

4.1. Data Sets and Experimental Setup

The gene expression data sets used in this study are summarized in Table 1. Experimental results that produced by CPSO are compared with an experimental method (BPSO) for objective comparisons. Additionally, the latest PSO-based methods from previous related works are also considered for comparison with CPSO [4],[5]. Firstly, we applied the gain ratio technique for pre-processing in order to pre-select 500-top-ranked genes. These genes are then used by CPSO and BPSO. Next, SVM is used to measure LOOCV accuracy on gene subsets that produced by CPSO and BPSO. For LOOCV, one sample in the training set is withheld and the remaining samples are used for building a classifier to classify the class of the withheld sample. By cycling through all the samples, we can get cumulative accuracy rates for classification accuracy of methods. In this research, LOOCV is used for measuring classification accuracy due to the small number of samples in gene expression data. Several experiments are independently conducted 10 times on each data set using CPSO and BPSO. Next, an average result of the 10 independent runs is obtained. High LOOCV accuracy, the small number of selected genes, and low running time are needed to obtain an excellent performance.

Table 1: The summary of gene expression data sets.

Data Sets	Number of Samples	Number of Genes	Number of Classes
Leukemia	72	7,129	2
Colon	62	2,000	2

Note:
DB = database.
DB Leukemia: http://www.broad.mit.edu/cgi-bin/cancer/datasets.cgi
DB Colon: http://microarray.princeton.edu/oncology/affydata/index.html

4.2. Experimental Results

Based on the standard deviation of classification accuracies in Table 2, results that produced by CPSO were almost consistent on all data sets. Interestingly, all runs have achieved over 98% LOOCV accuracy with less than 12 selected genes on the leukemia data set. Moreover, at least 88% classification accuracies have been obtained on the colon data set.

Table 2. Experimental results for each run using CPSO

Run no.	Leukemia data set		Colon data set	
	#Acc (%)	No. selected genes	#Acc (%)	No. selected genes
1	100	10	90.32	4
2	100	5	90.32	6
3	100	3	88.71	28
4	98.61	9	91.94	10
5	98.61	9	88.71	8
6	100	31	88.71	8
7	98.61	11	88.71	7
8	98.61	10	88.71	7
9	98.61	8	88.71	5
10	98.61	9	88.71	130
Average ± S.D.	99.17 ± 0.72	10.50 ± 7.61	89.36 ± 1.13	21.30 ± 38.80

Note: The results of the best subsets are shown in the shaded cells. A near-optimal subset that produces the highest classification accuracy with the smallest number of genes is selected as the best subset. #Acc and S.D. denote the classification accuracy and the standard deviation, respectively.

For an objective comparison, CPSO is compared with BPSO. According to the Table 3, it is worthwhile to mention that the classification accuracy and the number of selected genes of CPSO are superior to BPSO in terms of the best, average, and standard deviation results on all the data sets. CPSO also produces smaller numbers of genes and lower running times compared to BPSO on all the data sets. CPSO can reduce its running times because of the following reasons:

- CPSO selects the smaller number of genes compared to BPSO.
- The computation of SVM is fast because it uses the small number of features (genes) that selected by CPSO for classification process.

We also compare our work with previous related works that used PSO-based methods in their proposed methods [4],[5]. It is shown in Table 4. For all the data sets, the averages of the number of selected genes of our work were smaller than the previous works [4],[5]. Our work also have resulted the higher averages of classification accuracies on the leukemia data set compared to the previous works. However, experimental results produced by Shen *et al.* were better than our work on the colon data sets [4]. Running time between CPSO and the previous works cannot be compared because it was not reported in their articles.

Table 3. Comparative experimental results of CPSO and BPSO

Data	Method / Evaluation	CPSO			BPSO		
		Best	#Ave	S.D	Best	#Ave	S.D
Leukemia	#Acc (%)	100	99.17	0.72	98.61	98.61	0
	#Genes	3	10.50	7.16	216	224.70	5.23
	#Time	5.26	6.13	1.44	13.86	13.94	0.03
Colon	#Acc (%)	91.94	89.36	1.13	87.10	86.94	0.51
	#Gene	10	21.30	38.80	214	231	10.19
	#Time	8.78	9.26	0.70	30.58	30.63	0.27

Note: The best result of each data set is shown in the shaded cells. It is selected based on the following priority criteria: 1) the highest classification accuracy; 2) the smallest number of selected genes; 3) the lowest running time.

Table 4. A comparison between our method (CPSO) and previous PSO-based methods

Data	Method / Evaluation	CPSO	PSOTS [4]	PSOGA [5]
Leukemia	#Acc (%)	(99.17)	(98.61)	(95.10)
	#Genes	(10.50)	(7)	(21)
Colon	#Acc (%)	(89.36)	(93.55)	(88.7)
	#Genes	(21.30)	(8)	(16.8)

Note: The result of the best subsets is shown in the shaded cells. It is selected based on the following priority criteria: 1) the highest classification accuracy; 2) the smallest number of selected genes. A result in '( )' denotes an average result.
IBPSO = An improved binary PSO.
PSOGA = A hybrid of PSO and GA.
PSOTS = A hybrid of PSO and tabu search.

V. CONCLUSION

Overall, based on the experimental results, the performance of CPSO was superior to BPSO and previous PSO-based methods in terms of classification accuracy and the number of selected genes. CPSO was excellent because the probability $x_i^d(t+1)=0$ has been increased by the proposed constraint of elements of particle velocity vectors and the introduced rule. The constraint and rule have been proposed in order to yield a near-optimal subset of genes for better cancer classification. CPSO also obtains lower running times because it selects the small number of genes compared to BPSO. For future works, a statistical test will be applied on CPSO in order to test its reliability.

REFERENCES

[1] Knudsen S (2002) A biologist's guide to analysis of DNA microarray data. New York: John Wiley & Sons.

[2] Mohamad MS, Omatu S, Deris S, Misman MF, Yoshioka M (2009) Selecting informative genes from gene expression data by using hybrid methods for cancer classification. Int J Artif Life Robotics 13(2):414-417.

[3] Mohamad MS, Omatu S, Yoshioka M, Deris S (2009) A cyclic hybrid method to select a smaller subset of informative genes for cancer classification. Int J Innovative Comput, Inf, Control 5(8):2189–2202.

[4] Shen Q, Shi WM, Kong W (2008) Hybrid particle swarm optimization and tabu search approach for selecting genes for tumor classification using gene expression data. Comput Biol Chem 32:53–60.

[5] Li S, Wu X, Tan M (2008) Gene selection using hybrid particle swarm optimization and genetic algorithm. Soft Comput 12:1039–1048.

[6] Kennedy J, Eberhart R (1995) Particle swarm optimization. Proc 1995 IEEE Int Conf Neural Networks 4:1942-1948.

[7] Kennedy J, Eberhart R (1997) A discrete binary version of the particle swarm algorithm. Proc 1997 IEEE Int Con Systs, Man, Cybernetics 5:4104–4108.