Particle Swarm Optimization for Gene Selection in Classifying Cancer Classes

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: The application of microarray data for cancer classification has recently gained in popularity. The main problem that needs to be addressed is the selection of a smaller subset of genes from the thousands of genes in the data that contributes to a disease. This selection process is difficult due to the availability of a small number of samples compared to the huge number of genes, many irrelevant genes, and noisy genes. Therefore, this paper proposes an improved binary particle swarm optimization to select a near-optimal (smaller) subset of informative genes that is relevant for cancer classification. Experimental results show that the performance of the proposed method is superior to the experimental method and other related previous works in terms of classification accuracy and the number of selected genes.

Keywords: Gene selection, Hybrid approach, Microarray data, Particle swarm optimization**.**

I. INTRODUCTION

Microarray technology is a device that can be employed in measuring expression levels of thousands of genes simultaneously. It finally produces microarray data that contain useful information of biological, diagnostic, and prognostic for researchers.[1] Thus, there is a need to select informative genes that contribute to a cancerous state. However, the gene selection process poses a major challenge because of the following characteristics of microarray data: the huge number of genes compared to the small number of samples (higher-dimensional data), irrelevant genes, and noisy data. To overcome this challenge, a gene selection method is used to select a subset of genes that increases the classifier's ability to classify samples more accurately.

Recently, several methods based on particle swarm optimization (PSO) are proposed to select informative genes from microarray data.[2,3,4] PSO is a new evolutionary computation technique proposed by Kennedy and Eberhart.[5] It was motivated from the simulation of social behaviour of organisms such as bird flocking and fish schooling. The work of Shen *et al*. has proposed a hybrid of PSO and tabu search approaches for gene selection.[2] However, the results obtained by using the proposed hybrid method are less significant because the application of tabu approaches in PSO is unable to search into all possible search spaces. Next, an improved binary PSO have been proposed by the work of Chuang *et al*.[3] This approach produced 100% classification accuracy in many data sets, but it used a high number of selected genes to achieve the good result. This is due to all global best particles are reset to the same position when their fitness values does not change after three consecutive iterations. Li *et al*. introduced a hybrid of PSO and GA for the same purpose.[4] Unfortunately, the accuracy result is still not high and many genes selected for cancer classification since there is no probability relations between GA and PSO in the proposed hybrid method. Generally, the proposed methods that based on PSO[2,3,4] are intractable to efficiently produce a near-optimal (smaller) subset of informative genes for higher classification accuracy. This is mainly because the total number of genes in microarray data is too large (higher-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 propose an improved binary PSO to select a smaller (near-optimal) subset of informative genes that is most relevant for the cancer classification. The proposed method is evaluated on two real microarray data sets.

II. METHODS

2.1. The Standard Version of Binary PSO

Binary PSO is initialized with a population of particles. At each iteration, all particles move in the problem space to find the optimal solution. A particle represents a potential solution (gene subset) in an *n*-dimensional space.[6] Each particle has a 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 v_i^d in the range $[0, V_{\max}]$, whereas x_i^d is a binary bit, *i*=1,2,..*m* (*m* is the total number of particles); d=1,2,..*n* (*n* is the dimension of data).

Hence, the vector of particle positions is represented by a binary bit string of length *n*, where *n* is the dimension of data (the total number of genes). Each vector 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:

$$v_i^d(t+1) = w * v_i^d(t) + c_1 r_1 * (pbest_i^d(t) - x_i^d(t))$$
$$+ c_2 r_2 * (gbest^d(t) - x_i^d(t)) \qquad (1)$$

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

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

where w is the inertia weight. c_1 and c_2 are the acceleration constants in the interval [0,2]. r_1, r_2, and r_3 are random values in the range [0,1]. $Pbest_i = (pbest_i^1, pbest_i^2, ..., pbest_i^n)$ and $Gbest = (gbest^1, gbest^2, ..., gbest^n)$ represent the best previous position of the *i*th particle and the global best position of the swarm (all particles), respectively.

2.2. An Improved Binary PSO (IPSO)

In this paper we propose IPSO for gene selection. It is introduced to solve the problems derived from the microarray data, overcome the limitation of the previous works[2,3,4], and inline with the diagnostic goal. IPSO in our work differs from the methods in the previous works in one major part. The major difference is that we modify the existing rule (Eq. 3) for the position update in our proposed IPSO, whereas the previous works used a standard rule (Eq. 3) for the position update in their PSO. Firstly, we analyze the sigmoid function (Eq. 2). This function represents a probability for $x_i^d(t+1)$ to be 0 or 1 ($P(x_i^d(t+1)=0)$ or $P(x_i^d(t+1)=1)$). It has the properties as follows:

$$Sig(v_i^d(t+1)) \in [0,1]$$

$$\lim_{v_i^d(t+1)\to\infty} Sig(v_i^d(t+1)) = 1 \qquad (4)$$

$$\lim_{v_i^d(t+1)\to-\infty} Sig(v_i^d(t+1)) = 0 \qquad (5)$$

if $v_i^d(t+1) = 0$ then

$$P(x_i^d(t+1)=1) = 0.5 \quad \text{or} \quad Sig(0) = 0.5 \qquad (6)$$

if $v_i^d(t+1) < 0$ then

$$P(x_i^d(t+1)=1) < 0.5 \quad \text{or} \quad Sig(v_i^d(t+1) < 0) < 0.5 \qquad (7)$$

if $v_i^d(t+1) > 0$ then

$$P(x_i^d(t+1)=1) > 0.5 \quad \text{or} \quad Sig(v_i^d(t+1) > 0) > 0.5 \qquad (8)$$

$$P(x_i^d(t+1)=0) = 1 - P(x_i^d(t+1)=1) \qquad (9)$$

Also note that the value of $x_i^d(t+1)$ can change even if the value of $v_i^d(t+1)$ does not change, due to the random number r_3 in the Eq. 3. To propose IPSO, the first three items below are suggested:

A. A Simple Modification of the formula of velocity update (Eq. 1)

$$V_i(t+1) = w * V_i(t) + c_1 r_1 * (Pbest_i(t) - X_i(t))$$
$$+ c_2 r_2 * (Gbest(t) - X_i(t)) \qquad (10)$$

where $V_i \in [0, V_{\max}]$.

B. Calculation for the distance of two positions

The number of different bits between two particles relates to the difference between their positions. For example, $Gbest(t) = [1011101001]$ and $X_i(t) = [0100110101]$. The difference between $Gbest(t)$ and $X_i(t)$ is $[1-1110-11-100]$. A value of 1 indicates that compared with the best position, this bit (gene) should be selected, but is not selected, which may decrease classification quality and lead to a lower fitness value. In contrast, a value of -1 indicates that, compared with the best position, this bit should not be selected, but is selected. The selection of irrelevant genes makes the length of the subset longer and leads to a lower fitness value. Assume that the number of 1 is *a*, whereas the number of -1 is *b*. We use the absolute value of $(a-b)$ to express the distance between two positions. Such variation makes particles exhibit the ability of exploration within the solution space. In this example, $(a-b)=4-3$, so the distance between $Gbest(t)$ and $X_i(t)$ is $Gbest(t)-X_i(t)=1$.

C. Modify the existing rule of position update (Eq. 3).

In order to support the diagnostic goal that needs the least number of genes for accurate cancer classification, the rule of position update is simple modified as follows:

$$\text{if } S(V_i(t+1)) > r_3, \text{ then } x_i^d(t+1)=0; \text{ else } x_i^d(t+1)=1 \qquad (11)$$

Please note that the value of $V_i(t+1)$ is always a positive real number. Based on this velocity value, Eq. 2, and Eq. 11, the possibility of $x_i^d(t+1)=1$ is too small. This situation causes a smaller number of genes is selected in order to produce a near-optimal gene subset from higher dimensional data.

D. Fitness function

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

$$fitness(X_i)=w_1 \times A(X_i)+(w_2(M-R(X_i))/M) \qquad (12)$$

in which $A(X_i)\in[0,1]$ is leave-one-out-cross-validation (LOOCV) accuracy on the training set using the only genes in X_i. This accuracy is provided by an SVM classifier. $R(X_i)$ is the number of selected genes in X_i. M is the total number of genes for each sample in the training set. 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 $w_2=1-w_1$.

III. EXPERIMENTS

3.1. Data Sets and Experimental Setup

Two benchmark microarray data sets are used to evaluate IPSO: leukaemia cancer and colon cancer data sets. The leukaemia data set contains the expression levels of 7,129 genes and can be obtained at http://www.broad.mit.edu/cgi-bin/cancer/datasets.cgi. It has 72 samples. For the colon cancer data set, there are 62 samples. It can be obtained at http://chestsurg.org/publications/2002-microarray.aspx.

Firstly, we applied the gain ratio technique to pre-select 500-top-ranked genes. These genes are then used by IPSO in the next process. In this paper, LOOCV is used to measure classification accuracy of a gene subset that produced by IPSO. The implementation of LOOCV is in exactly the same way as did by Chuang *et al.*[3] Two criteria following their importance are considered to evaluate the performance of IPSO: LOOCV accuracy, and the number of selected genes. A near-optimal subset that produces the highest classification accuracy with the smallest number of genes is selected as the best subset. Several experiments are independently conducted 10 times on each data set using IPSO and a standard version of PSO. Next, an average result of the 10 independent runs is obtained.

3.2. Experimental Results

Based on the standard deviations of classification accuracy and the number of selected genes in Table 1, results that produced by IPSO were nearly consistent on both data sets. Interestingly, all runs have achieved 100% LOOCV accuracy on the leukaemia data set with less than 5 selected genes.

According to the Table 2, overall, it is worthwhile to mention that the classification accuracy and the number of selected genes of IPSO are superior to the standard version of binary PSO in terms of the best, average, and standard deviation results.

For an objective comparison, we only compare our work with related previous works that used PSO in their methods.[2,3,4] It is shown in Table 3. For the leukaemia data set, the averages of LOOCV accuracy and the number of selected genes of our work were 100% and 3.5 genes, respectively. The latest previous work also came up with the similar LOOCV result to ours, but they used more than 1,000 genes to obtain the same result.[3] Overall, this work has outperformed the related previous works on both the data sets in terms of LOOCV accuracy and the number of selected genes.

According to Tables 1-3, IPSO is reliable for gene selection since it has produced the near-optimal solution from microarray data. This is due to the modification of position update that causes the selection of a smaller number of genes. Therefore, IPSO yields the optimal gene subset (a smaller subset of informative genes with higher classification accuracy) for cancer classification.

IV. CONCLUSION

In this paper, IPSO has been proposed and tested for gene selection on two real microarray data. Based on the experimental results, the performance of IPSO was superior to the standard version of binary PSO and related previous works. This is due to the fact that the modified rule of position update in IPSO causes a smaller number of genes is selected in each iterative, and finally produce a near-optimal subset of informative genes for better cancer classification. For future work, a combination between a constraint approach and PSO will be proposed to increase the classification accuracy.

Table 1. Classification accuracies for each run using IPSO

Run#	**Leukaemia Data Set**		**Colon Data Set**	
	Classification Accuracy (%)	**#Selected Genes**	**Classification Accuracy (%)**	**#Selected Genes**
1	100	4	93.55	5
2	100	2	93.55	5
3	100	4	96.77	4
4	100	4	93.55	5
5	100	3	93.55	4
6	100	4	95.16	5
7	100	4	93.55	4
8	100	3	95.16	4
9	100	4	93.55	5
10	100	3	93.55	4
Average ± S.D	100 ± 0	3.50 ± 0.71	94.19 ± 1.13	4.5 ± 0.53

Note: Results of the best subsets shown in shaded cells. S.D. denotes the standard deviation, whereas #Selected Genes and Run# represent the number of selected genes and a run number, respectively.

Table 2. A comparison in terms of statistical results of the proposed IPSO and the standard version of PSO

Data	**Method** **Evaluation**	**IPSO**			**The standard version of binary PSO**		
		The Best	**Average**	**S.D**	**The Best**	**Average**	**S.D**
Leukaemia	Classfication Accuracy (%)	100	100	0	98.61	98.61	0
	#Selected Genes	2	3.50	0.71	216	224.70	5.23
Colon	Classfication Accuracy (%)	96.77	94.19	1.13	87.10	86.94	0.51
	#Selected Genes	4	4.50	0.53	214	231	10.19

Note: The best result of each data set shown in shaded cells. S.D. denotes the standard deviation, whereas #Selected Genes represents the number of selected genes.

Table 3. A comparison between our method (IPSO) and other previous methods based on PSO

Data	**Method** **Evaluation**	**This work**	**PSOTS [Shen et al.[2]]**	**IBPSO [Chuang et al.[3]]**	**PSOGA [Li et al.[4]]**
Leukaemia	Classfication Accuracy (%)	(100)	(98.61)	100	(95.1)
	#Selected Genes	(3.5)	(7)	1034	(21)
Colon	Classfication Accuracy (%)	(94.19)	(93.55)	-	(88.7)
	# Selected Genes	(4.50)	(8)	-	(16.8)

Note: The results of the best subsets shown in shaded cells. ‘-‘ means that a result is not reported in the related previous work. A result in ‘( )’ denotes an average result. #Selected Genes represents the number of selected genes.
PSOTS = A hybrid of PSO and tabu search. IBPSO = An improved binary PSO. PSOGA = A hybrid of PSO and GA

REFERENCES

[1] Knudsen S (2002) A biologist’s guide to analysis of DNA Microarray Data. John Wiley & Sons.

[2] 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

[3] Chuang LY, Chang HW, Tu CJ, *et al.* (2008) Improved binary PSO for feature selection using gene expression data. Comput Biol Chem 32:29–38

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

[5] Kennedy J, Eberhart R (1995) Particle swarm optimization. Proceedings of the IEEE International Conference on Neural Networks 4:1942–1948

[6] Kennedy J, Eberhart R (1997) A discrete binary version of the particle swarm algorithm. Proceedings of the IEEE International Conference on Systems, Man, and Cybernetics 5:4104–4108