The Fifteenth International Symposium on Artificial Life and Robotics 2010 (AROB 15th '10),
B-Con Plaza, Beppu,Oita, Japan, February 4-6, 2010

Particle Swarm Optimization with a Modified Sigmoid Function for Gene Selection from Gene Expression Data

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: In order to select a small subset of informative genes from gene expression data for cancer classification, recently, many researchers are analyzing gene expression data using various computational intelligence methods. However, due to the small number of samples compared to the huge number of genes (high-dimension), irrelevant genes, and noisy genes, many of the computational methods face difficulties to select the small subset. Thus, we propose an enhancement of binary particle swarm optimization to select a small subset of informative genes that is relevant for classifying cancer samples more accurately. In this proposed method, three approaches have been introduced to increase the probability of bits in particle's positions to be zero. By performing experiments on two gene expression data sets, we have found that the performance of the proposed method is superior to previous related works, including the conventional version of binary particle swarm optimization (BPSO) in terms of classification accuracy and the number of selected genes. The proposed method also produces lower running times compared to BPSO.

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

I. INTRODUCTION

Recent advances in microarrays 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. However, the gene selection process poses a major challenge because of the following characteristics of gene expression 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 usually used to select a subset of informative genes that maximizes classifier's ability to classify samples more accurately [2]. The advantages of gene selection has been reported in Mohamad *et al*. [2].

Recently, several gene selection methods based on particle swarm optimization (PSO) have been proposed to select informative genes from gene expression data [3],[4],[5]. PSO is a new optimization technique proposed by Kennedy and Eberhart [6]. It is motivated from the simulation of social behavior of organisms such as bird flocking and fish schooling. Shen *et al*. [3] 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, an improved binary PSO have been proposed by Chuang *et al*. [4]. This approach produced 100% classification accuracy in many data sets, but it used a high number of selected genes (large gene subset) to achieve the high accuracy. It uses the high number because of the global best particle is reset to zero position when its fitness values do not change after three consecutive iterations. After that, 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 [3],[4],[5] are intractable to efficiently

produce a small (near-optimal) subset of informative genes for high classification accuracy. This is mainly because the total number of genes in gene expression data is too large (high-dimensional data).

Therefore, we propose an enhancement of binary PSO (EPSO) to select a small (near-optimal) subset of informative genes that is most relevant for classifying cancer classes more accurately. In order to test the effectiveness of our proposed method, we apply EPSO to two gene expression data sets, including binary-classes and multi-classes data sets.

II. METHODS

2.1. The Conventional 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:

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

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

if $Sig(v_i^d(t+1)) > r_3^d(t)$, then $x_i^d(t+1) = 1;$

else $x_i^d(t+1) = 0$ (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 ith 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.

2.2. An Enhancement of Binary PSO (EPSO)

In this article, we propose EPSO for selecting a near-optimal (small) subset of genes. It is proposed to overcome the limitations of BPSO and previous PSO-based methods [3],[4],[5]. EPSO in our work differs from BPSO and the PSO-based methods on three parts: 1) we introduce a scalar quantity that called particles' speed (s); 2) we propose a rule for updating $x_i^d(t+1)$; 3) we modify the existing sigmoid function, whereas BPSO and the PSO-based methods have used the original rule (Eq. 3) and the standard sigmoid function (Eq.2), and no particles' speed implementation. The particles' speed, rule, and sigmoid function 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 particles' speed, rule, and sigmoid function are proposed as follows:

$$s_i(t+1) = w(t) \times s_i(t) + c_1 r_1(t) \times dist(Pbest_i(t) - X_i(t)) + c_2 r_2(t) \times dist(Gbest(t) - X_i(t)) \quad (4)$$

$$Sig(s_i(t+1)) = \frac{1}{1+e^{-5 s_i(t+1)}} \quad (5)$$

subject to $s_i(t+1) \geq 0$

if $Sig(s_i(t+1)) > r_3^d(t)$, then $x_i^d(t+1) = 0;$

else $x_i^d(t+1) = 1$ (6)

where $s_i(t+1)$ represents the speed of the particle i for the t+1 iteration, whereas in BPSO and previous PSO-based methods (Eq. 1, Eq. 2, and Eq. 3), $v_i^d(t+1)$ represents a single element of a particle velocity vector for the particle i. In EPSO, Eq. 4, Eq. 5, and Eq. 6 are used to replace Eq. 1, Eq. 2, and Eq. 3, respectively. $s_i(t+1)$ is the rate at which the particle i changes its position. The most important property of $s_i(t+1)$ is $s_i(t+1) \geq 0$. Hence, $s_i(t+1)$ is used instead of

$v_i^d(t+1)$ so that its positive value can increase $P(x_i^d(t+1)=0)$. In Mohamad *et al*. [8], there is an explanation on how to calculate the distance between two positions of two particles, e.g., $dist(Gbest(t)-X_i(t))$ in Eq. 4.

Equations (4-6) and $s_i(t)\geq 0$ increase $P(x_i^d(t)=0)$ because the minimum value for $P(x_i^d(t)=0)$ is 0.5 when $s_i(t)=0$ $(\min P(x_i^d(t)=0)\geq 0.5)$. 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 $s_i(t)>0$, 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 $s_i(t)=1$, then $P(x_i^d(t)=0)=0.993307$ and $P(x_i^d(t)=1)=1-P(x_i^d(t)=0)=0.006693$.

if $s_i(t)=2$, then $P(x_i^d(t)=0)=0.999955$ and $P(x_i^d(t)=1)=1-P(x_i^d(t)=0)=0.000045$.

A. Fitness functions

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

$$fitness(X_i)=w_1\times A(X_i)+(w_2\times(n-R(X_i))/n) \quad (7)$$

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 $w_2=1-w_1$.

III. EXPERIMENTS

3.1. Data Sets and Experimental Setup

Two real microarrays data sets are used to evaluate EPSO and BPSO: leukemia cancer and mixed-lineage leukemia (MLL) data sets. The leukemia 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 two cancer classes: acute lymphoblastic leukemia (ALL) and acute myeloid leukemia (AML). In this data set, bone marrow and blood samples were taken from 72. There are also 72 samples in the MLL cancer data. It has three tumor classes (MLL, ALL, and AML) and can be obtained at http://www.broad.mit.edu/cgi-bin/cancer/datasets.cgi.

In order to avoid selection bias, the implementation of LOOCV is in exactly the same way as did by Chuang *et al*. [4] where the only one cross-validation cycle (outer loop), namely LOOCV is used.

3.2. Experimental Results

Based on the standard deviation of classification accuracy in Table 1, results that produced by EPSO were consistent on all data sets. Interestingly, all runs have achieved 100% LOOCV accuracy with less than 71 selected genes on the Leukemia data set. Over 97% classification accuracies have been obtained on the MLL data set. This means that EPSO has efficiently selected and produced a near-optimal gene subset from high-dimensional data (gene expression data).

According to the Table 2, overall, it is worthwhile to mention that the classification accuracy of EPSO are superior to BPSO in terms of the best, average, and standard deviation results on all the data sets. Moreover, EPSO also produces a smaller number of genes compared to BPSO. The running times of EPSO are lower than BPSO in all the data sets. EPSO can reduce its running times because of the following reasons:

For an objective comparison, we compare our work with previous related works that used PSO-based methods in their proposed methods [3],[4],[5]. It is shown in Table 3. For the leukemia data set, the averages of classification accuracies of our work were higher than the previous works. Our work also have resulted the smaller averages of the number of selected genes on all the data sets compared to the previous works.

Table 1. Experimental results for each run using PSO

Run#	Leukemia		MLL	
	#Acc (%)	#Selected Genes	#Acc (%)	#Selected Genes
1	100	55	100	131
2	100	65	100	123
3	100	65	100	117
4	100	70	100	113
5	100	51	100	116
6	100	62	100	109
7	100	58	100	116
8	100	61	100	114
9	100	63	100	111
10	100	67	100	111
Average	100	61.70	100	116.10
± S.D.	± 0	± 5.72	± 0	± 6.56

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; 3) the lowest running time. #Acc and S.D. denote the classification accuracy and the standard deviation, respectively, whereas #Selected Genes and Run# represent the number of selected genes and a run number, respectively. #Time stands for running time.

Table 2. Comparative experimental results of EPSO and BPSO

Data	Method / Evaluation	EPSO			BPSO		
		Best	#Ave	S.D	Best	#Ave	S.D
Leukemia	#Acc (%)	100	100	0	98.61	98.61	0
	#Genes	51	61.70	5.72	3488	3528.75	26.83
	#Time	7.52	7.46	0.67	261.34	261.41	0.18
MLL	#Acc (%)	100	100	0	95.83	95.83	0
	#Genes	109	116.10	6.56	6101	6153.1	31.62
	#Time	13.51	13.83	0.18	236.759	239.00	1.34

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 3. A comparison between our method (EPSO) and previous PSO-based methods

Data	Method / Evaluation	EPSO	IBPSO [4]	PSOTS [3]	PSOGA [5]
Leukemia	#Acc (%)	(100)	100	(98.61)	(95.10)
	#Genes	(61.70)	1034	(7)	(21)
MLL	#Acc (%)	(100)	100	-	-
	#Genes	(116.10)	1292	-	-

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. '-' means that a result is not reported in the previous related work. 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.

The latest previous work also came up with the similar LOOCV results (100%) to ours on the Leukemia data sets but they used many genes obtain the same results [4]. Moreover, they could not have statistically meaningful conclusions because their experimental results were obtained by only one independent run on each data set, and not based on average results. The average results are important since their proposed method is a stochastic approach. Additionally, in their approach, the global best particles' position is reset to zero position when its fitness values do not change after three successive iterations.

According to Tables 1-3, EPSO is reliable for gene selection since it has produced the near-optimal solution from gene expression data. This is due to the proposed particles' speed, the introduced rule, and the modified sigmoid function increase the probability $x_i^d(t+1)=0$ $(P(x_i^d(t+1)=0))$. This high probability causes the selection of a small number of informative genes and finally produces a near-optimal subset (a small subset of informative genes with high classification accuracy) for cancer classification. The particles' speed is introduced to provide the rate at which a particle changes its position, whereas the rule is proposed to update particle's positions. The sigmoid function is modified for increasing the probability of bits in particle's positions to be zero.

IV. CONCLUSION

In this paper, EPSO has been proposed for gene selection on two gene expression data sets. Overall, based on the experimental results, the performance of EPSO was superior to BPSO and PSO-based methods that proposed by previous related works in terms of classification accuracy, the number of selected genes, and running times. EPSO was excellent because the probability $x_i^d(t+1)=0$ has been increased. For future works, a modified representation of particle's positions in PSO will be proposed to reduce the number of genes subsets in solution spaces.

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, 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.

[3] 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.

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

[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.

[8] Mohamad MS, Omatu S, Deris S, Yoshioka M, (2009) Particle swarm optimization for gene selection in classifying cancer classes. Int J Artif Life Robotics 14(1):16–19.