A Constraint and Rule In an Enhancement of Binary Particle Swarm Optimization To Select Informative Genes For Cancer Classification

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

[1]Artificial Intelligence and Bioinformatics Research Group, Faculty of Computer Science and Information Systems, Universiti Teknologi Malaysia, Skudai, 81310 Johor, Malaysia.
{saberi*, safaai}@utm.my

[2]Department of Computer Science and Intelligent Systems, Graduate School of Engineering, Osaka Prefecture University, Sakai, Osaka 599-8531, Japan
{omatu, yoshioka} @cs.osakafu-u.ac.jp

*Corresponding author

Abstract

Gene expression data have been analyzing by many researchers by using a range of computational intelligence methods. From the gene expression data, selecting a small subset of informative genes can do cancer classification. Nevertheless, many of the computational methods face difficulties in selecting small subset since the small number of samples needs to be compared to the huge number of genes (high-dimension), irrelevant genes and noisy genes. Hence, to choose the small subset of informative genes that is significant for the cancer classification, an enhanced binary particle swarm optimization is proposed. Here, the constraint of the elements of particle velocity vectors is introduced and a rule for updating particle's position is proposed. Experiments were performed on five different gene expression data. As a result, in terms of classification accuracy and the number of selected genes, the performance of the introduced method is superior compared to the conventional version of binary particle swarm optimization (BPSO). The other significant finding is lower running times compared to BPSO for this proposed method.

Keywords: Binary particle swarm optimization, gene selection, gene expression data, optimization.

1 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 [3]. Thus, there is a need to select informative genes that contribute to a cancerous state [5]. 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 [6]. 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],[7]-[10]. PSO is a new evolutionary technique proposed by Kennedy and Eberhart [1]. Shen et al. have proposed a hybrid of PSO and tabu search approaches for gene selection [7]. 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. have introduced a hybrid of PSO and genetic algorithms (GA) for the same purpose [4]. 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.

Next, Chuang *et al.* proposed an improved binary PSO [8]. 100% classification accuracy in many data sets had been yielded by using the proposed method, but it utilized a large number of selected genes (large gene subset) to obtain the high accuracy. This method used a large number of genes because the global best particle was reset to the zero position when its fitness values did not change after three consecutive iterations. Chuang *et al*. [9],[10] introduced a combination of tabu search and PSO for gene selection [9], and currently they proposed a hybrid of BPSO and a combat GA for the same purpose [10]. However, both proposed approaches still need a high number of selected to result high classification accuracy. A significant weakness was found resulting from the combination of PSO and tabu search or a combat GA which did not share probability significance in their processes. Generally, the PSO-based methods are intractable to efficiently produce a small (near-optimal) subset of informative genes for high classification accuracy [4],[7]-[10]. 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 five gene expression data sets, including binary-classes and multi-classes data sets.

This paper is organized as follows. In Section 2 and Section 3, we briefly describe the conventional version of binary PSO and CPSO, respectively. Section 4 presents data sets used and experimental results. Section 5 summarizes this paper by providing its main conclusions and addresses future developments.

2 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. Each particle has position and velocity vectors for directing its movement. The position vector and velocity vector of the ith 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) [2]. v_i^d represent an element of particle velocity vectors. It is a real number for the *d*-th dimension of the particle *i*, where the maximum v_i^d, $V_{\max} = (1/3) \cdot n$.

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:

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

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

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

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 and initialized with 1.4. It is updated as follows:

$$w(t+1) = \frac{(w(t)-0.4) \times (MAXITER - Iter(t))}{(MAXITER + 0.4)} \quad (4)$$

where *MAXITER* is the maximum iteration (generation) and $Iter(t)$ is the current iteration.

3 An enhancement of Binary PSO (CPSO)

Almost all previous works of gene expression data researches have selected a subset of genes to obtain excellent cancer classification. Therefore, in this article, 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],[7]-[10]. 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:

1. increase the probability of $x_i^d(t+1) = 0$ $(P(x_i^d(t+1) = 0))$.
2. 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:

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

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

$$\text{if } Sig(v_i^d(t+1)) > r_3^d(t), \text{ then } x_i^d(t+1) = 0; \text{ else } x_i^d(t+1) = 1 \quad (6)$$

Theorem 1. 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 $v_i^d(t)=0$ $(\min P(x_i^d(t)=0)=0.5)$. Mean while, they decrease the maximum value for $P(x_i^d(t)=1)$ to 0.5 $(\max P(x_i^d(t)=1)\le 0.5)$. Therefore, if $v_i^d(t)>0$, then $P(x_i^d(t)=0)>>0.5$ and $P(x_i^d(t)=1)<<0.5$.

Proof. ( ) Figure 1 shows that a) The constraint of elements of particle velocity vectors and the rule in CPSO increase $P(x_i^d(t)=0)$; b) Equations (1-3) in BPSO yield $P(x_i^d(t)=0)=P(x_i^d(t)=1)=0.5$. For example, the calculations for $P(x_i^d(t)=0)$ and $P(x_i^d(t)=1)$ in Fig. 2(a) 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:

$$fitness(X_i)=w_1 A(X_i)+(w_2(n-R(X_i))/n) \tag{7}$$

in which $A(X_i)$ 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 $[0.1,0.9]$ and $w_2=1-w_1$.

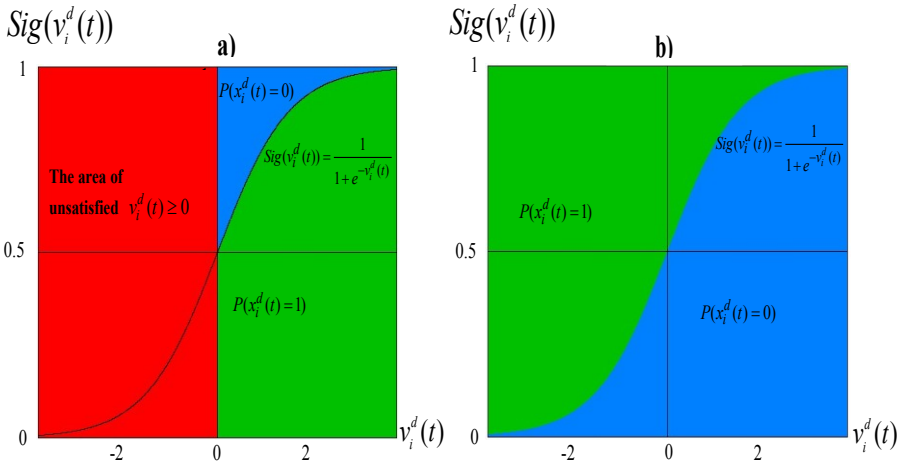

Fig. 1. The areas of $P(x_i^d(t)=0)$ and $P(x_i^d(t)=1)$ based on sigmoid functions in a) CPSO; b) BPSO. The blue and green colors show the areas for $P(x_i^d(t)=0)$ and $P(x_i^d(t)=1)$ respectively, and whereas the red color indicates the part of unsatisfied $v_i^d(t)$ □ 0

4 Experiments

4.1 Data Sets and Experimental Setup

The gene expression data sets used in this study are summarized in Table 1. They included binary-classes and multi-classes data sets. 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],[7]-[10]. 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 set	No. classes	No. samples	No. genes	Source
Leukemia	2 (ALL and AML)	72 (67 ALL and 25 AML)	7,129	http://www.broad.mit.edu/cgi-bin/cancer/datasets.cgi
Lung	2 (MPM and ADCA)	181 (31 MPM and 150 ADCA)	12,533	http://chestsurg.org/publications/2002-microarray.aspx.
MLL	3 (ALL, MLL, and AML)	72 (24 ALL, 20 MLL, and 28 AML)	12,582	http://www.broad.mit.edu/cgi-bin/cancer/datasets.cgi
SRBCT	4 (EWS, RMS, NB, and BL)	82 (28 EWS, 25 RMS, 18 NB, and 11 BL)	2,308	http://research.nhgri.nih.gov/microarray/Supplement/
Colon	2 (Normal and tumor)	62 (22 normal and 40 tumor)	2,000	http://microarray.princeton.edu/oncology/affydata/index.html

Note:
MPM = malignant pleural mesothelioma.
ADCA = adenocarcinoma.
ALL = acute lymphoblastic leukemia.
MLL = mixed-lineage leukemia.
AML = acute myeloid leukemia.
SRBCT = small round blue cell tumors.

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 100% LOOCV accuracy with less than 50 selected genes on the SRBCT data set. Moreover, over 97% classification accuracies have been obtained on other data sets, except for the colon data set. This means that CPSO has efficiently selected and produced a near-optimal gene subset from high-dimensional data (gene expression data).

Table 2. Experimental results for each run using CPSO on the leukemia, colon, lung, MLL, and SRBCT data sets

Run#	Leukemia		Colon		Lung		MLL		SRBCT	
	#Acc (%)	# Select-ed Genes	#Acc (%)	#Select-ed Genes	#Acc (%)	# Selected Genes	#Acc (%)	#Select-ed Genes	#Acc (%)	#Select-ed Genes
1	100	10	90.32	4	99.45	9	97.22	32	100	20
2	100	5	90.32	6	99.45	9	98.61	113	100	48
3	100	3	88.71	28	99.45	7	97.22	38	100	42
4	98.61	9	91.94	10	99.45	30	97.22	28	100	50
5	98.61	9	88.71	8	99.45	8	97.22	6	100	21
6	100	31	88.71	8	99.45	9	95.83	6	100	37
7	98.61	11	88.71	7	98.90	8	97.22	11	100	32
8	98.61	10	88.71	7	99.45	5	97.22	37	100	27
9	98.61	8	88.71	5	99.45	15	97.22	88	100	21
10	98.61	9	88.71	130	99.45	13	97.22	33	100	50
Aver-age ± S.D.	99.17 ± 0.72	10.50 ± 7.61	89.36 ± 1.13	21.30 ± 38.80	99.39 ± 0.15	11.30 ± 7.17	97.22 ± 0.66	39.20 ± 35.04	100 ± 0	34.80 ± 12.30

Note: Results of the best subsets is shown in 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, whereas #Selected Genes and Run# represent the number of selected genes and a run number, respectively.

Figure 2 shows that the averages of fitness values of CPSO increase dramatically after a few generations on all the data sets. A high fitness value is obtained by a combination between a high classification rate and a small number (subset) of selected genes. The condition of the proposed constraint of elements of particle velocity vectors that should always be positive real numbers started in the initialization method, and the new rule for updating particle's positions provoke the early convergence of CPSO. In contrast, the averages of fitness values of BPSO was no improvement until the last generation due to $P(x_i^d(t)=0)=P(x_i^d(t)=1)=0.5$.

For an objective comparison, CPSO is compared with BPSO. According to the Table 3, overall, 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. The classification accuracies of BPSO and CPSO were same on the lung and SRBCT data sets. However, the number of selected genes of BPSO was higher than CPSO to achieve the same accuracy.

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.

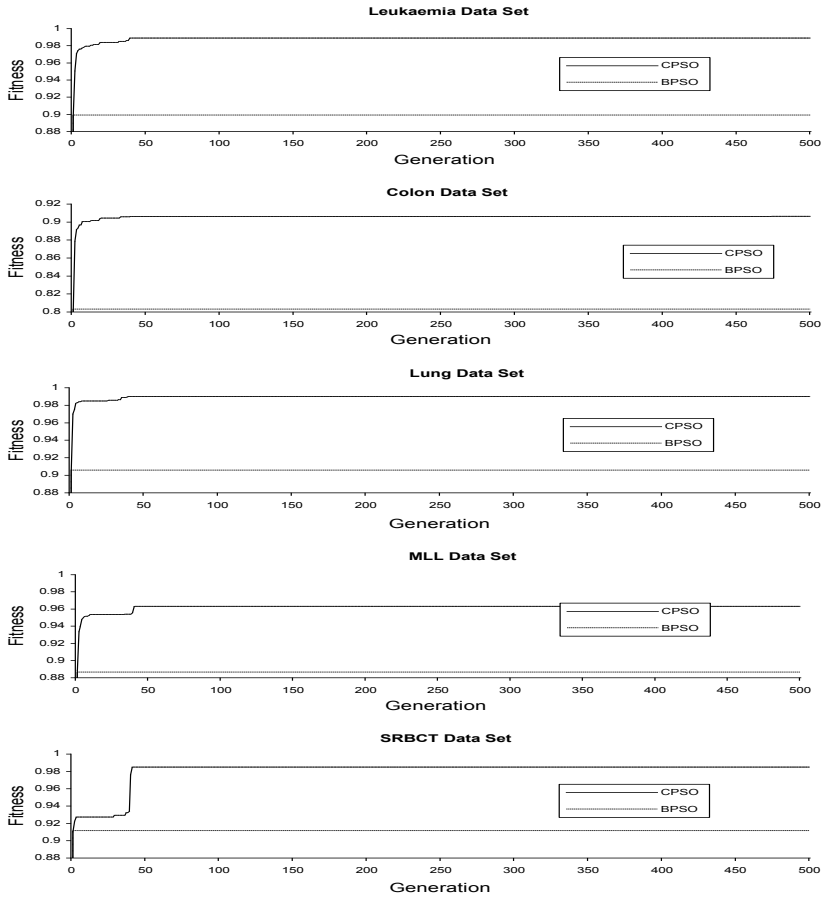

Fig. 2. The relation between the average of fitness values (10 runs on average) and the number of generations for CPSO and BPSO

We also compare our work with previous related works that used PSO-based methods in their proposed methods [4],[7]-[10]. 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. Our work also have resulted the higher averages of classification accuracies on the leukemia and SRBCT data sets compared to the previous works. However, experimental results produced by Shen et al. were better than our work on the colon data set [7]. This is due to the incorporation of tabu search (TS) as a local improvement procedure enables the algorithm HPSOTS to overleap local optima and show satisfactory performance in classifying cancer classes and reducing the number

of genes. Running time between CPSO and the previous works cannot be compared because it was not reported in their articles.

According to Fig. 2 and Tables 2-4, CPSO is reliable for gene selection since it has produced the near-optimal solution from gene expression data. This is due to the proposed constraint of elements of particle velocity vectors and the introduced rule increase the probability $x_i^d(t+1)=0$ $(P(x_i^d(t+1)=0))$. The increased probability value for $x_i^d(t+1)=0$ 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.

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
	#Genes	10	21.30	38.80	214	231	10.19
	#Time	8.78	9.26	0.70	30.58	30.63	0.27
Lung	#Acc (%)	99.45	99.39	0.18	99.45	99.39	0.18
	#Genes	5	11.30	7.17	219	223.33	4.24
	#Time	63.53	64.40	0.87	110.71	111.07	0.23
MLL	#Acc (%)	98.61	97.22	0.66	97.22	97.22	0
	#Genes	113	39.20	35.04	218	228.11	4.86
	#Time	9.51	11.64	4.98	19.37	19.90	0.35
SRBCT	#Acc (%)	100	100	0	100	100	0
	#Genes	20	34.80	12.30	206	221.30	7.35
	#Time	21.67	21.76	1.32	44.86	44.88	0.01

Note: The best result of each data set is shown in shaded cells. It is selected based on the following priority criteria: 1) the highest classification accuracy; 2) the smallest number of selected genes. #Acc and S.D. denote the classification accuracy and the standard deviation, respectively, whereas #Genes and #Ave represent the number of selected genes and an average, respectively. #Time stands for running time in the hour unit.

Table 4. A Comparison Between Our Method (CPSO) and previous PSO-Based Methods

Data	Method / Evaluation	CPSO	PSOTS [7]	PSOGA [4]	IBPSO [8]	TS-BPSO [9]	BPSO-CGA [10]
Leu-kemia	Average #Acc (%)	(99.17)	(98.61)	(95.10)	-	-	-
	Best #Acc (%)	100	-	-	100	100	100
	Average #Genes	(10.50)	(7)	(21)	-	-	-
	Best #Genes	3	-	-	1034	2577	300
Colon	Average #Acc (%)	(89.36)	(93.55)	(88.7)	-	-	-
	Best #Acc (%)	91.94	-	-	-	-	-
	Average #Genes	(21.30)	(8)	(16.8)	-	-	-
	Best #Genes	10			-	-	-
Lung	Average #Acc (%)	(99.39)	-	-	-	-	-
	Best #Acc (%)	99.45	-	-	-	-	-
	Average #Genes	(11.30)	-	-	-	-	-
	Best #Genes	5	-	-	-	-	-
MLL	Average #Acc (%)	(97.22)	-	-	-	-	-
	Best #Acc (%)	98.61	-	-	-	-	-
	Average #Genes	(39.20)	-	-	-	-	-
	Best #Genes	113	-	-	-	-	-
SRBC T	Average #Acc (%)	(100)	-	-	-	-	-
	Best #Acc (%)	100	-	-	100	100	100
	Average #Genes	(34.80)	-	-	-	-	-
	Best #Genes	20	-	-	431	1084	880

Note: '-' means that a result is not reported in the related previous work. A result in '( )' denotes an average result.

PSOGA = A hybrid of PSO and GA. PSOTS = A hybrid of PSO and tabu search. IBPSO = An improved binary PSO.

TS-BPSO = A combination of tabu search and BPSO. BPSO-CGA = A hybrid of BPSO and a combat genetic algorithm.

5 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 modified representation of particle's positions in PSO will be proposed to reduce the number of genes subsets in solution spaces.

Acknowledgement

We would like to thank Malaysian Ministry of Science, Technology and Innovation for supporting this research by an e-science research grant (Grant number: 06-01-06-SF1029). This research is also supported by a GUP research grant (Grant number: Q.J130000.2507.04H16) that was sponsored by Universiti Teknologi Malaysia.

References

1. J. Kennedy and R. Eberhart, Particle swarm optimization, Proceedings of the 1995 IEEE International Conference on Neural Networks 4 (1995), 1942-1948.
2. J. Kennedy and R. Eberhart, A discrete binary version of the particle swarm algorithm, Proceedings of the 1997 IEEE International Conference on Systems, Man, and Cybernetics 5 (1997), 4104-4108.
3. S. Knudsen, A Biologist's Guide to Analysis of DNA Microarray Data, John Wiley & Sons, New York, 2002.
4. S. Li, X. Wu and M. Tan, Gene selection using hybrid particle swarm optimization and genetic algorithm, Soft Comput. 12 (2008), 1039-1048.
5. M. S. Mohamad, S. Omatu, S. Deris, M. F. Misman and M. Yoshioka, Selecting informative genes from gene expression data by using hybrid methods for cancer classification, Int. J. Artif. Life & Rob. 13(2) (2009), 414-417.
6. M. S. Mohamad, S. Omatu, M. Yoshioka and S. Deris, A cyclic hybrid method to select a smaller subset of informative genes for cancer classification, Int. J. Innovative Comput., Inf. & Control 5(8) (2009), 2189-2202.
7. Q. Shen, W. M. Shi and W. Kong, Hybrid particle swarm optimization and tabu search approach for selecting genes for tumor classification using gene expression data, Comput. Biol. & Chem. 32 (2008), 53-60.
8. L. Y. Chuang, H. W. Chang, C. J. Tu and C. H. Yang, Improved binary PSO for feature selection using gene expression data, Comput. Biol. & Chem. 32 (2009), 29-38.
9. L. Y. Chuang, C. H. Yang and C. H. Yang, Tabu search and binary particle swarm optimization for feature selection using microarray data, J. Comput. Biol. 16(12) (2009), 1689-1703.
10. L. Y. Chuang, C. H. Yang, J. C. Li and C. H. Yang, A hybrid BPSO-CGA approach for gene selection and classification of microarray data, J. Comput. Biol. 18 (2011): 1-14.