An Improved Binary Particle Swarm Optimisation for Gene Selection in Classifying Cancer Classes

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

[1] Department of Computer Science and Intelligent Systems, Graduate School of Engineering, Osaka Prefecture University, Sakai, Osaka 599-8531, Japan
mohd.saberi@sig.cs.osakafu-u.ac.jp,
{omatu, 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
safaai@utm.my, anazida@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 because of the availability of the 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 optimisation 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 a standard version of particle swarm optimisation 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 optimisation.

1 Introduction

Microarray 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 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 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 [3].

Recently, several methods based on particle swarm optimisation (PSO) are proposed to select informative genes from microarray data [4],[5],[6]. PSO is a new evolutionary technique proposed by Kennedy and Eberhart [7]. It is motivated from the simulation of social behaviour of organisms such as bird flocking and fish schooling.

S. Omatu et al. (Eds.): IWANN 2009, Part II, LNCS 5518, pp. 495–502, 2009.

Shen *et al*. have proposed a hybrid of PSO and tabu search approaches for gene selection [4]. However, the results obtained by using the hybrid method are less significant because 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*. [5]. This approach produced 100% classification accuracy in many data sets, but it used a higher number of selected genes to achieve the higher accuracy. It uses the higher number because of all global best particles are reset to the same position when their fitness values do not change after three consecutive iterations. Li *et al*. have introduced a hybrid of PSO and GA for the same purpose [6]. Unfortunately, the accuracy result is still not high and many genes are selected for cancer classification since there is no direct probability relation between genetic algorithms (GA) and PSO. Generally, the proposed methods that based on PSO [4],[5],[6] 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 (IPSO) to select a smaller (near-optimal) subset of informative genes that is most relevant for the cancer classification. The proposed method is evaluated on three real microarray data sets.

2 Methods

2.1 A Standard Version of Binary PSO (BPSO)

Binary PSO (BPSO) is initialised 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 (gene subset) in an *n*-dimensional space [8]. 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 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).

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 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 a generation updates its own position and velocity according to the following equations:

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

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

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

where w is the inertia weight. The value of this weight is chosen based on several preliminary runs. 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 ith particle and the global best position of the swarm (all particles), respectively. They are assessed base on a fitness function. $Sig(v_i^d)$ is a sigmoid function where $Sig(v_i^d) \in [0,1]$.

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 related previous works [4],[5],[6], 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, whereas the previous works used a standard rule (Eq. 3). Firstly, we analyse the sigmoid function (Eq. 2). This function represents a probability for x_i^d to be 0 or 1 ($P(x_i^d = 0)$ or $P(x_i^d = 1)$). It has the properties as follows:

$$\lim_{v_i^d \to \infty} Sig(v_i^d) = 1. \tag{4}$$

$$\lim_{v_i^d \to -\infty} Sig(v_i^d) = 0. \tag{5}$$

$$\text{if } v_i^d = 0 \text{ then } P(x_i^d = 1) = 0.5 \text{ or } Sig(0) = 0.5. \tag{6}$$

$$\text{if } v_i^d < 0 \text{ then } P(x_i^d = 1) < 0.5 \text{ or } Sig(v_i^d < 0) < 0.5. \tag{7}$$

$$\text{if } v_i^d > 0 \text{ then } P(x_i^d = 1) > 0.5 \text{ or } Sig(v_i^d > 0) > 0.5. \tag{8}$$

$$P(x_i^d = 0) = 1 - P(x_i^d = 1). \tag{9}$$

Also note that the value of x_i^d can change even if the value of v_i^d does not change, due to the random number r_3 in the Eq. 3. To propose IPSO, the following approaches are suggested:

2.2.1 Modifying 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:

$$\text{If } S(V_i) > r_3, \text{ then } x_i^d = 0; \text{ else } x_i^d = 1. \tag{10}$$

The value of particle velocity, V_i in the modified formula (Eq. 10) represents the whole of elements of a particle velocity vector, whereas the standard formula represents a single element. Moreover, V_i is also a positive real number. Based on this positive velocity value, Eq. 2, and Eq. 10, the possibility of $x_i^d = 1$ 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 (microarray data).

2.2.2 A Simple Modification of the Formula of Velocity Update (Eq. 1)

In this formula, the calculation of the value of velocity is completely based on the whole of bits of a particle position vector, whereas the original formula (Eq. 1) is based on a single bit.

$$V_i = w * V_i + c_1 r_1 * (Pbest_i - X_i) + c_2 r_2 * (Gbest - X_i). \tag{11}$$

2.2.3 Calculation for the Distance of Two Positions

The number of different bits between two particles relates to the difference between their positions. For example, $Gbest = [0011101000]$ and $X_i = [1100110100]$. The difference between $Gbest$ and X_i is $[-1-1110-11-100]$. A value of 1 indicates that compared with the best position, this bit (gene) should be selected, but it 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 it 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 $(a-b)$, $|a-b|$ to express the distance between two positions. In this example, $|a-b| = |3-4| = 1$, so the distance between $Gbest$ and X_i is $Gbest - X_i = 1$.

2.2.4 Fitness Function

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

$$fitness(X_i) = w_1 \times A(X_i) + (w_2 (M - R(X_i)) / M). \tag{12}$$

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 support vector machine classifiers (SVM). $R(X_i)$ 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 $w_2 = 1 - w_1$.

3 Experiments

3.1 Data Sets and Experimental Setup

Three benchmark microarray data sets are used to evaluate IPSO: leukaemia, lung, and mixed-lineage leukaemia (MLL) cancer data sets. The leukaemia data set contains 72 samples of the expression levels of 7,129 genes. It can be obtained at http://www.broad.mit.edu/cgi-bin/cancer/datasets.cgi. For the lung cancer data set, there are 181 samples. It can be downloaded at http://chestsurg.org/publications/2002-microarray.aspx. The MLL cancer data set has three leukaemia classes: acute lymphoblastic leukaemia (ALL), acute myeloid leukaemia (AML), and MLL. There are 12,582 genes in each sample. This data set contains 72 samples and can be downloaded at http://www.broad.mit.edu/cgi-bin/cancer/datasets.cgi.

Firstly, we applied the gain ratio technique to pre-select 500-top-ranked genes. These genes are then used by IPSO and a standard version of binary PSO (BPSO). In this paper, LOOCV is used to measure classification accuracy of a gene subset. The implementation of LOOCV is in exactly the same way as did by Chuang *et al.* [5] 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 BPSO. 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 all data sets. Interestingly, all runs have achieved 100% LOOCV accuracy with less than 10 selected genes on the data sets. This means that IPSO has efficiently selected and produced a near-optimal gene subset from higher-dimensional data (microarray data).

Table 1. Experimental results for each run using IPSO

Run#	Leukaemia Data Set		Lung Data Set		MLL Data Set	
	Classification Accuracy (%)	#Selected Genes	Classification Accuracy (%)	#Selected Genes	Classification Accuracy (%)	#Selected Genes
1	100	4	100	9	100	7
2	100	2	100	6	100	6
3	100	4	100	6	100	7
4	100	4	100	5	100	6
5	100	3	100	6	100	8
6	100	4	100	8	100	4
7	100	4	100	4	100	5
8	100	3	100	5	100	7
9	100	4	100	7	100	8
10	100	3	100	6	100	9
Average	100	3.50	100	6.20	100	6.70
± S.D	± 0	± 0.71	± 0	± 1.48	± 0	± 1.50

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.

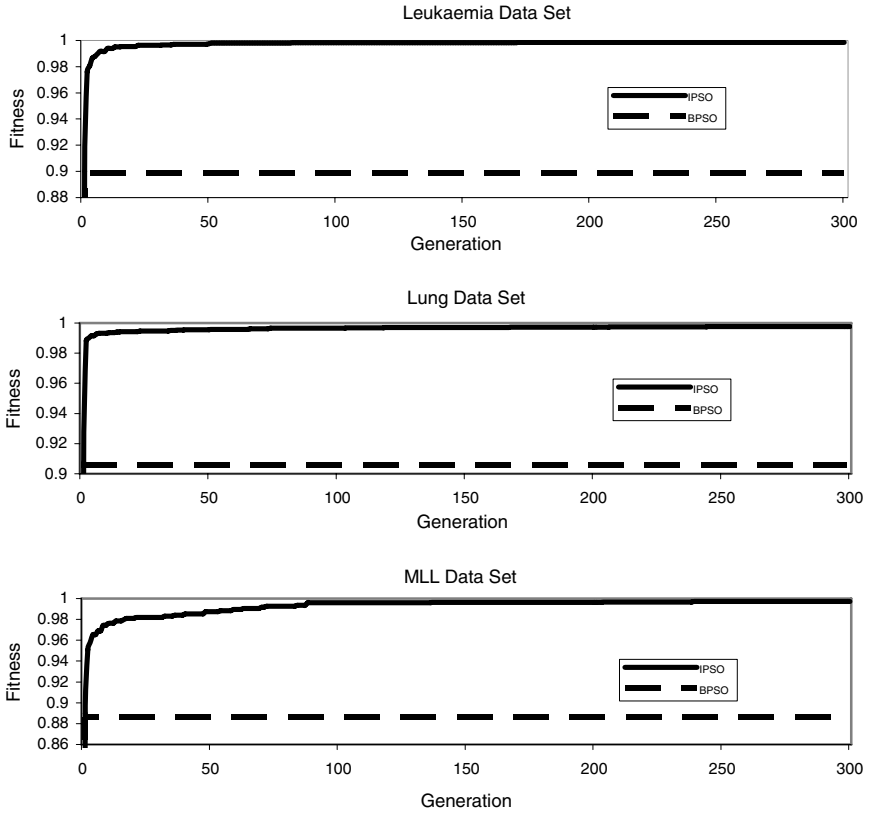

Fig. 1. The relation between the average of fitness values (10 runs on average) and the number of generations for IPSO and BPSO

Figure 1 shows that the average of fitness values of IPSO increases dramatically after a few generations in all the data sets. The higher fitness produces a smaller subset of selected genes with higher classification rate. The condition of velocity that should always be positive real numbers provided in the initialisation method, and the new rule of position update provoke the early convergence of IPSO. In contrast, the average of fitness values of BPSO was no improvement until the last generation.

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 BPSO in terms of the best, average, and standard deviation results on all the data sets.

For an objective comparison, we compare our work with related previous works that used PSO in their methods [4],[5],[6]. It is shown in Table 3. For all the data sets, the averages of LOOCV accuracy and the number of selected genes of our work were 100% and less than seven selected 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 [5]. Overall, this work has outperformed the related previous works on all data sets in terms of LOOCV accuracy and the number of selected genes.

Table 2. A comparison in terms of statistical results of the proposed IPSO and BPSO

Data	Method / Evaluation	IPSO			The standard version of binary PSO (BPSO)		
		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
Lung	Classfication Accuracy (%)	100	100	0	99.45	99.39	0.18
	#Selected Genes	4	6.20	1.48	219	223.33	4.24
MLL	Classfication Accuracy (%)	100	100	0	97.22	97.22	0
	#Selected Genes	4	6.70	1.50	218	228.11	4.86

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	IPSO [Our work]	PSOTS (Shen *et al.* [4])	IBPSO (Chuang *et al.*[5])	PSOGA (Li *et al.* [6])
Leukaemia	Classification Accuracy (%)	(100)	(98.61)	100	(95.1)
	#Selected Genes	(3.5)	(7)	1034	(21)
Lung	Classification Accuracy (%)	(100)	-	-	-
	# Selected Genes	(6.20)	-	-	-
MLL	Classification Accuracy (%)	(100)	-	100	-
	# Selected Genes	(6.70)	-	1292	-

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.

According to Fig. 1 and 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.

4 Conclusions

In this paper, IPSO has been proposed for gene selection on three 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 generation, and finally produce a near-optimal subset of genes for better cancer classification. For future works, a combination between a constraint approach and PSO will be proposed to minimise the number of selected genes.

References

1. Knudsen, S.: A Biologist's Guide to Analysis of DNA Microarray Data. John Wiley & Sons, New York (2002)
2. Mohamad, M.S., Omatu, S., Deris, S., Misman, M.F., Yoshioka, M.: Selecting Informative Genes from Microarray Data by Using Hybrid Methods for Cancer Classification. J. Artif. Life Rob. 13(2), 414–417 (2009)
3. Mohamad, M.S., Omatu, S., Deris, S., Hashim, S.Z.M.: A Model for Gene Selection and Classification of Gene Expression Data. J. Artif. Life Rob. 11(2), 219–222 (2007)
4. Shen, Q., Shi, W.M., Kong, W.: Hybrid Particle Swarm Optimization and Tabu Search Approach for Selecting Genes for Tumor Classification Using Gene Expression Data. Comput. Biol. Chem. 32, 53–60 (2008)
5. Chuang, L.Y., Chang, H.W., Tu, C.J., Yang, C.H.: Improved Binary PSO for Feature Selection Using Gene Expression Data. Comput. Biol. Chem. 32, 29–38 (2008)
6. Li, S., Wu, X., Tan, M.: Gene Selection Using Hybrid Particle Swarm Optimization and Genetic Algorithm. Soft Comput. 12, 1039–1048 (2008)
7. Kennedy, J., Eberhart, R.: Particle Swarm Optimization. In: 1995 IEEE International Conference on Neural Networks, vol. 4, pp. 1942–1948. IEEE Press, Los Alamitos (1995)
8. Kennedy, J., Eberhart, R.: A Discrete Binary Version of the Particle Swarm Algorithm. In: 1997 IEEE International Conference on Systems, Man, and Cybernetics, vol. 5, pp. 4104–4108. IEEE Press, Los Alamitos (1997)