ISSN : 0974 - 7435 Volume 8 Issue 11

BioTechnology
An Indian Journal

FULL PAPER

BTAIJ, 8(11), 2013 [1483-1489]

The classification of multiclass tumor gene expression data based on two-layer particle swarm optimization

Yajie Liu [1,2]*, Xinling Shi[1], Changxin Gou1, Baolei Li[1] and Lian Gao[1]

[1]Information School, Yunnan University, North of Cuihu Road, No.2, Kunming, Yunnan, (CHINA)

[2]Center of Medical Equipment, Third Affiliated Hospital of Kunming Medical University (Yunnan Tumor Hospital), Kunzhou Road, No.518, Kunming, Yunnan, (CHINA)

E-mail: lyjcome@qq.com

ABSTRACT

The classification of gene expression data to determine different type of tumor samples is significantly important to research tumors in molecular biology level for making further treatment plan of the patient. Particle swarm optimization (PSO) has employed as a solution for classification and clustering in bioinformatics. In this study, a classifier based on the two layer particle swarm optimization (TLPSO) algorithm is established to classify the uncertain training sample sets obtained from gene expression data of breast, prostate, lung and colon tumor samples. Compared with PSO and K-means algorithm in validation, the classification stability and accuracy based on the proposed TLPSO algorithm is improved significantly, which may provide more information to clinicians for choosing more appropriate treatment.

KEYWORDS

Classification;
Gene;
Tumor;
TLPSO;
PSO.

INTRODUCTION

In last decades, it has been increasingly recognized that targeting specific therapies to distinct tumor subtypes can help maximizing the treatment efficacy and minimizing the toxicity to normal organs[1]. So an accurate cancer classification becomes a necessity. However, conventional cancer classification largely relies on a complex and inexact combination of clinical and histopathological data[2]. These classic methods cannot provide an accurate classification when dealing with atypical tumors or morphologically indistinguishable tumor subtypes.

Advances in the area of gene microarray technologies have led to promise of cancer diagnosis using new molecular based approaches[3]. It offers hope that cancer classification can be objective and highly accurate, which could provide clinicians with the information to choose the most appropriate forms of treatment.

Prediction of the diagnostic category of a tissue sample in identified categories is known as classification. A challenge in prediction the diagnostic categories using microarray data is that the number of genes is usually much greater than the number of tissue samples available[4].

Multiclass classification techniques can be roughly divided into three types. The first type is the binary classification algorithm with limited application for two class

FULL PAPER

problems, including weighted voting scheme[5], K nearest neighbors[6], support vector machine[7], deterministic forest[8]. The second one is the decompositions of multiclass problems into binary ones combining with other scheme methods, such as the one-versus-rest and the one-versus-one[9] method. The last type is directly classification of multiclass expression data, including genetic programming[10] with no global search ability and particle swarm optimization usually with unstable prediction results[11].

In this paper, the two-layer particle swarm optimization (TLPSO) algorithm is applied to multiclass tumor sample classification. The multiclass gene expression data, which contains breast, lung, prostate and colon tumor data, is used as sample data. In order to evaluate the performance of the proposed approach, the particle swarm optimization (PSO) and K-means algorithm is also applied to the same gene expression data to compare the results of them.

METHOD

Two-layer particle swarm optimization

The TLPSO is a novel evolutionary algorithm from PSO algorithm, whose block diagram can be shown in Figure. 1. In PSO, each particle moves around in a D-dimensional search space simultaneously based on its own memory and knowledge gained by the swarm as a whole to find the best solution[12]. In TLPSO, there is a two layer structure: top layer and bottom layer[13]. The whole particles *N* are divided into *M* swarms, each swarm contains *N/M* particles in the bottom layer, *M* swarms constitute the top layer. Each global best position in each swarm of the bottom layer is set to be the position of the particle in the swarm of the top layer. Therefore, the global best position in the swarm of the top layer influences indirectly the particles of each swarm in the bottom layer. Furthermore, a mutation operation is added into particles of each swarm in the bottom layer. Consequently, the diversity of the population in the TLPSO increases so that the TLPSO has the ability to avoid trapping into a local optimum.

Initially, *M* swarms of *N* particles, x^{jk}, $j=1,2,\ldots,M$, $k=1, 2,\ldots, N$, are randomly generated in the bottom layer, where x^{jk} is the position of the *k*-th particle in the

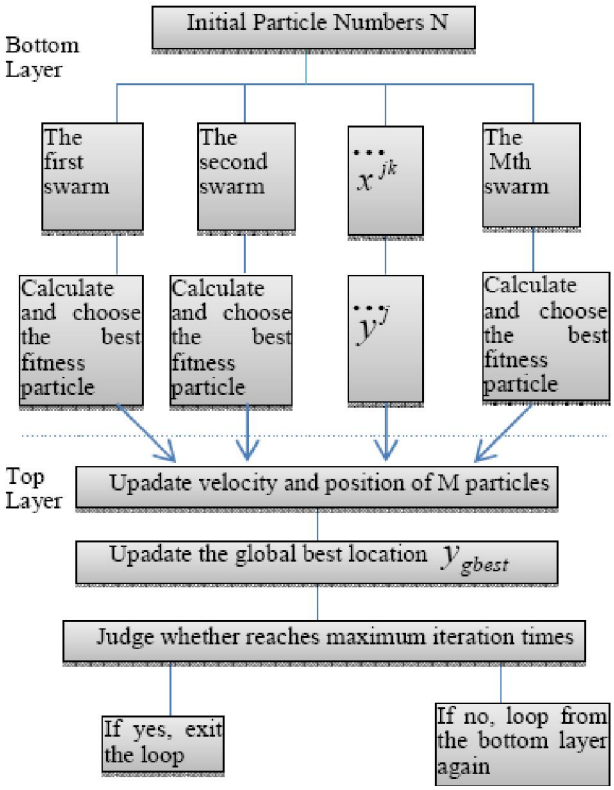

Figure 1 : The Block diagram of the TLPSO

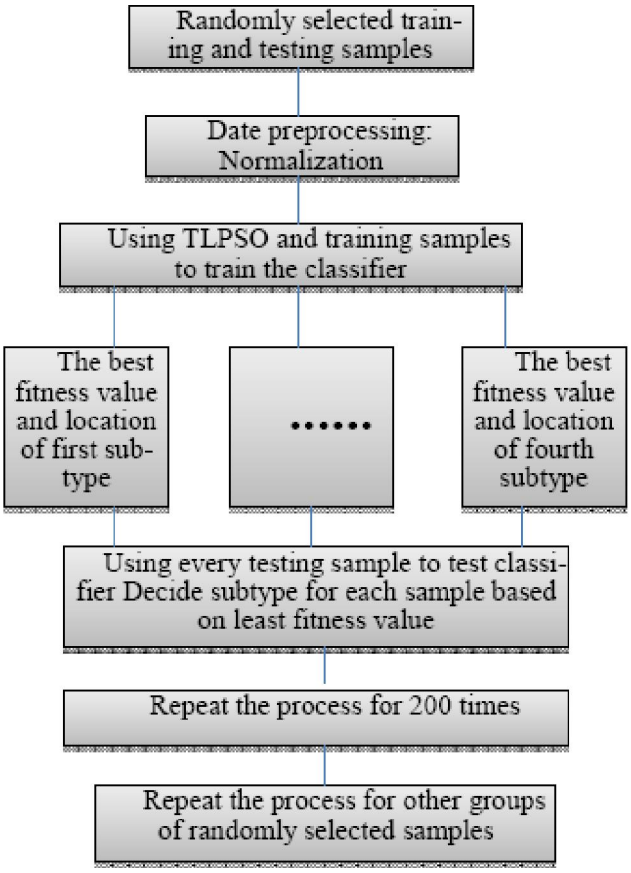

Figure 2 : The flow chart of classification

j-th swarm of the bottom layer, the global best position of the j-th swarm, $y^j, j \in \{1, 2, \ldots, M\}$, is determined. Here, the global best position of the swarm in the bottom layer is set to be the position of the particles in the swarm of the top layer. That is, the particles of the swarm, y^j in the top layer are determined. Then, the global best position of the swarm, y_{gbest}, in the top layer is determined according to the fitness values of the particles in the top layer.

Data description

The multiclass expression data used in this study contains 4 tumor types, including breast, prostate, lung and colon as they are the most well-known and common tumor disease in adults, comprising almost 50% of all tumors[14]. Each tumor sample used in this experiment contains 103 samples and 5521 genes produced by HG-U94A platform[15].

Uncertain sample set

As the composition and quantity of training and testing samples usually plays a fundamental role in classification, researches have been accustomed to using existing or divided training and testing datasets. However, divided training and testing samples will not always be existing especially in multiclass tumor gene expression data. The other difficulty is that it is not an easy way to confirm which composition of training and testing samples will be best for classification. In order to get an objective and comprehensive evaluation for classification results, the following method is designed. First, training samples are randomly selected from different subtypes. Then testing samples are acquired by removing training samples from the whole dataset. Finally, repeat the process by training samples increasing from 5 to 20 with an interval of 5 samples to obtain 8 different combinations of training and testing samples.

Classification procedure

Data oreprocessing

To avoid overfitting, the data are simply normalized to have zero-mean and one standard deviation.

Initial particle population

M particles are generated randomly with independent position and velocity. Dimensionality was 5521 corresponding to number of genes.

Establish classifier

A classifier based on four clustering centrals are expressed by fitness and position values and established by applying the TLPSO algorithm to training samples of different subtypes respectively for breast, prostate, lung, and colon.

Validity

The classifier is used to validate testing samples.

Repeat

There are 100 realizations of the classifier by using randomly selected training and testing samples to produce a series of distribution of classification results.

Change samples and repeat

The process described above for other subgroups is repeated with randomly selected training and testing samples. The results are compared to evaluate the performance of the proposed approach in this paper. Figure.2 shows the flow chat of classification.

RESULTS AND DISCUSSION

Multiclass tumor gene expression data contains 26 breast samples, 26 prostate samples, 28 lung samples and 23 colon samples. Each sample consists of 5521 genes. Initially, 5 samples are selected randomly from each tumor type to obtain totally 20 training samples and 83 testing samples for training the classifier based on TLPSO algorithm. The classifier is described as the form of two variables, which are global best fitness value and its position of each sample. Finally, testing samples are used to validate the classifier which could be evaluated by counting the correct predicted testing samples.

The classification process is repeated for 100 times. However, longer time is needed for the classifying process compared with our last research of less gene numbers. And different results for randomly selected samples are obtained each time because randomly selected training and testing samples are not always the same. It is difficult to confirm which composition of training and testing samples will produce the best classification result. In order to make a comprehensive and comparative evaluation, classification numbers are enlarged to

FULL PAPER

100 times with 5 training samples randomly selected for each time. The distribution of classification results is considered to be an influencing factor of performance. To better understand which combination of training samples will produce the best result, training samples increase from 5, 10, and 15 to 20. Finally, each subgroup of training sample is run for 100 times and totally 400 times classification is performed. Particle numbers and iteration times of the first subgroups are also reduced to 60 and 200 times to shorten the performance time. TABLE 1 shows different composition of training and testing samples. Training samples are 20 in the second column and third row because there are four tumor types in the dataset and 5 samples are randomly selected for each one.

TABLE 1 : Different composition of training and testing samples

Subgroups Results	1	2	3	4
Samples of Each Subtype	5	10	15	20
Training Samples	20	40	60	80
Testing Samples	83	63	43	23
Total samples	103	103	103	103
Genes	5521	5521	5521	5521

TABLE 2 shows the initial conditions, best prediction results and distribution of 100 times classification for each subgroup and algorithm. Results in column TLPSO* and TLPSO come from the same algorithm. However, the value of each initial variable is unvaried for each subgroup in TLPSO* and gradually increases in TLPSO and PSO. Subgroups and compositions are consistent in the three algorithms. The Best Result means the best classification result which contains the least wrongly predicted samples. It is clearly that different combination of training and testing samples can produce different results, which means one time or two times classification results could not be considered as an evaluation way even though the numbers of best prediction samples are so close or already equal to total testing one. 90%~100% means the ratio of correctly predicted samples to whole testing samples of each sample composition is greater than 90%.

Obviously, the classification results are unexpectedly poor and widely distributed in TLPSO*. The best prediction rate is all below 80% in three subgroups. The best prediction rates decreases with the increase of training samples and the classification results is relatively good in subgroup of less training samples. It could be considered as the enlargement of training samples have greatly accelerated the computation complexity and make it difficult to convergence.

In order to improve the computation performance, the particle numbers and iteration times are tried to rise gradually for different subgroups of TLPSO. TLPSO in TABLE 2 shows the best prediction result and distribution of 100 times classification results for each sub-

TABLE 2 : The initialization, best prediction results and distribution of TLPSO*, TLPSO and PSO.

Algorithms Results	TLPSO*				TLPSO				PSO			
Subgroups	1	2	3	4	1	2	3	4	1	2	3	4
Particle Numbers	60	60	60	60	60	90	150	300	60	90	150	300
Iterations	200	200	200	200	200	250	300	600	200	250	300	600
Best Prediction Samples	77	49	26	16	77	57	35	22	80	60	43	23
Best Prediction Rate%	92.77	77.78	60.47	69.57	92.77	90.47	81.40	95.65	96.39	95.24	100	100
90%~100%	2	0	0	0	2	1	0	1	40	50	52	67
80%~90%	15	0	0	0	15	4	1	0	55	45	44	31
70%~80%	20	6	0	0	20	12	1	1	5	5	4	2
60%~70%	35	19	1	8	35	18	11	6	0	0	0	0
50%~60%	16	25	9	6	16	27	11	10	0	0	0	0
40%~50%	10	28	20	14	10	19	28	20	0	0	0	0
30%~40%	2	17	37	50	2	14	25	37	0	0	0	0
20%~30%	0	5	33	20	0	5	22	22	0	0	0	0
10%~20%	0	0	0	2	0	0	1	3	0	0	0	0
0~10%	0	0	0	0	0	0	0	0	0	0	0	0

group with different particles numbers and iteration times. Compared to fixed initialize condition in TLPSO*, the difference is the relatively better performance when the same point is that it also decreases with the increase of training samples, which demonstrates that the classification result is so sensitive to the numbers of training samples. However, it is important to note that the classification ratio could always surpass 80% at least in one time classification for each subgroup which provides a crucial point for further comparison.

Figure 3 shows the box plots of classification results for each sample composition with different initial conditions and respective results. On each box, the central line is the median and the edges of the box are the 25th and 75th percentiles. The whiskers extend to the most extreme data points not considered outliers, and outliers are plotted individually. Shorter and upper of the box means the better of the performance. The particle numbers and iteration time are unchangeable in subgroups of the TLPSO*, while it increases gradually with the expansion of training samples in TLPSO. It is clear that both the best and average prediction result of TLPSO is better than TLPSO*.

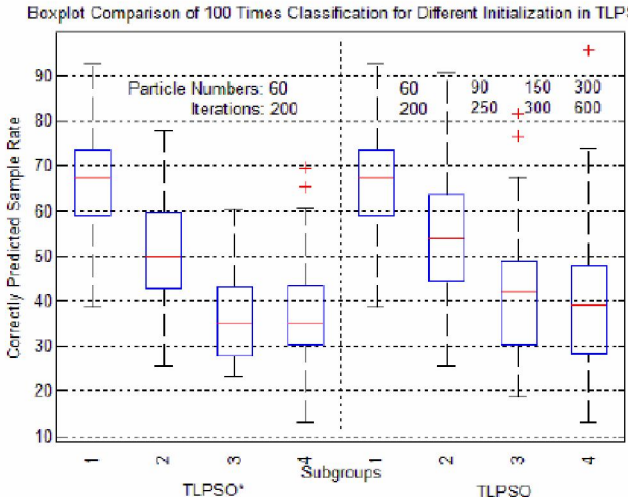

Figure 3 : Boxplot comparison of 100 times classification for TLPSO* and TLPSO.

In addition, basic PSO algorithm is used to make a comparison to further validate TLPSO. All parameters and initialization included in PSO are the same as in the TLPSO excepting the value of M because it is the special for TLPSO. The corresponding data of PSO in TABLE 2 shows the results and distributions of PSO. There is no clear difference for the best prediction results between PSO and TLPSO, while the distribution of correctly predicted samples is completely different. In PSO, all the classification ratios surpass 70% and some of them beyond 90%. The performance is improved with the increase of training samples for the TLPSO* and TLPSO. For the indicator of distribution, the performance of PSO is more excellent than the TLPSO. It is also clear in Figure.4 that boxes of PSO are more sizable and symmetrical, while boxes of TLPSO vary with their length and height.

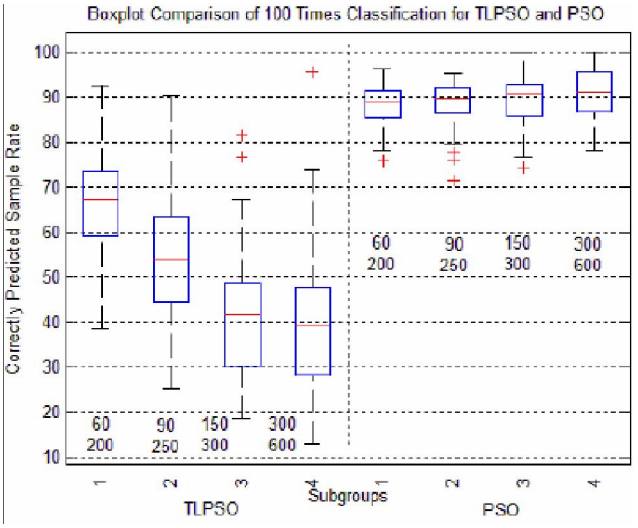

Figure 4 : Boxplot comparison of 100 times classification for TLPSO and PSO.

Finally, K-means classification algorithm is used to make a comparison. As an unsupervised learning algorithm, the K-means is used directly on all samples during the classification procedure, in which the number of clustering centrals is 4 determined by prior knowledge. However, the classification results of K-means are different and vary greatly each time. So, it can be conclude from TABLE 3 that the K-means is not a stable and accurate classification algorithm for samples used in this study.

The superiority of PSO algorithm is the improved accuracy of its classification as all prediction ratios surpass 70% in 100 times classification. However, the stability of the algorithm cannot deduced from those experiments results for the reason that randomly selected combinations of training and testing samples used for PSO and TLPSO make it difficult to further comparison. The same point for the three algorithms is the preset clustering numbers, while the difference is the processing way and their performance.

FULL PAPER

In order to make a comprehensive comparison of the three algorithms for better understanding their stability and accuracy, a subgroup of fixed sample combination which produces the best prediction result and accuracy in 100 times classification is used to make further classification. Particle numbers and iteration times are enlarged to 1000 and 2000 respectively and are unchangeable in each subgroup. For convenience of comparison the time required for these algorithms, just 10 times classification is performed in each subgroup of each algorithm.

As shown in TABLE 3, the classification results are almost the same in TLPSO of each subgroup, but there is great different for PSO and K-means. It can be seen that although the results based on PSO is excellent for subgroup 4, the numbers and names of wrongly predicted samples vary greatly each time from subgroup 1 to subgroup 3. It means that there should be more training for the algorithm if better result need to be got. It is clear that both accuracy and stability of the results based on K-means is no good that the least wrongly predicted samples are 4 while the most ones are more than 20.

TABLE 3 : Wrongly predicted samples of the TLPSO, PSO and K-means in 10 times classification

	Wrongly predicted samples								
NO	Subgroup 1		Subgroup 2		Subgroup 3		Subgroup 4		All the samples
	TLPSO	PSO	TLPSO	PSO	TLPSO	PSO	TLPSO	PSO	K-means
1	66	60,70	15,16,66	2,7,14	13,16	7,16	15	23	13-16,53-80
2	66	60,70	15,16,66	14,37,52,62,90,92,93	13,16	60,70	15	0	13,15,16,67
3	66	60,70	15,16,66	14,57,61,62,73	13,16	60,70	15	20,23	6,13-16,53-80,81,84,87,89,90
4	66	60,70	15,16,66	14,61,62,73,93	0	60,70	15	23	13,15,16,55,62-66,68-70,81,84
5	66	1,38,60,70	15,16,66	2,7,14	13,16	70	15	0	6,13-16,59,66,
6	66	38,60,70	15,16,66	14,93	13,16	70	15	0	6,13-16,53-80,81,84,87,89,90
7	66	60,70	15,16,66	2,14	13	7,18,85,89,92	15	20,23	13,15,16,67
8	66	14,60,70	15,16,66	2,7,14,16	13	7	15	0	13,15,16,55,62-66,68-70,81,84
9	66	60,70	15,16,66	2,7,14	13,16	0	0	23	6,13-16,59,66,
10	66	60,70	15,16,66	2,14	13,16	70	15	0	13-16,53-80

However, the numbers and names of wrongly predicted samples are almost the same in each subgroup of TLPSO. In subgroup 1 and 4, just one of them appears as wrongly predicted sample and even none of them arises in one time. The classification result is also stable in subgroup 2 and 3, though there are more than one mistaken results. Besides, there is no mistake in one time classification in subgroup 3 and 4 for large numbers of training sample, which is employed in this experiment. It is also clear from Figure.5 that both the stability and accuracy of TLPSO is super than other two algorithms.

CONCLUSIONS

In this paper, a classification procedure based on TLPSO algorithm is established and evaluated by handling the tumor gene expression data of 103 samples which contain 5521 genes for each sample and belong to 4 different type. These data are divided into 4 training and testing subgroups based on randomly selection strategy during the experiment.

100 times classification for each subgroup is performed to demonstrate the performance of TLPSO. The indicator of best prediction rate almost beyond 80%, while the distribution of prediction accuracy differs greatly compared to PSO. It is because that the particle number and iteration time is not large enough for TLPSO. Then a comparison with PSO and K-means algorithm in 10 times classification for subgroups with fixed sample combination is carried out. The best prediction results show that the TLPSO outperforms the PSO and K-means in both of the stability and accuracy during the 10 times classification. Consider the characteristic of PSO, which is more likely trapped in local optimum, so it could not always reach global optimum in each time. For K-means, numbers of wrongly predicted samples vary greatly and there is no referenced value. In conclusion, both of the stability and accuracy of TLPSO outperforms the two algorithms.

Our further research work includes stability validation and performance improvement in datasets with larger number classes and genes. We will also make more detailed research to compare the different classification results among tumor diseases.

ACKNOWLEDGMENT

This work was supported by the National Natural Science Foundation of China under Grants 61062005.

REFERENCES

[1] Xiaohong Huang, Wei Pan; Linear regression and two-class classification with gene expression data, BIOINFORMATICS, **19**, 2072-2078 **(2003)**.

[2] Jane Jijun Liu, Gene Cutler, Wuxiong Li; Multiclass cancer classification and biomarker discovery using GA-based algorithms, BIOINFORMATICS, **21**, 2691-2697 **(2005)**.

[3] S.Bicciato, A.Luchini, C.Di Bello; PCA disjoint models for multiclass cancer analysis using gene expression data, BIOINFORMATICS, **19**, 571-578 **(2003)**.

[4] Ka Yee Yeung, Roger E.Bumgarner, Adrian E.Raftery; Bayesian model averaging development of an improved multi class gene selection and classification tool for microarray data, BIOINFORMATICS, **21**, 2394-2402 **(2005)**.

[5] A.M.Bagirov, B.Ferguson, S.Ivkovic; New algorithm for multiclass cancer diagnosis using tumor gene expression signatures, BIOINFORMATICS, **19**, 1800-1807 **(2003)**.

[6] Sridhar Ramaswamy, Pablo Tamayo, Ryan Rifkin; Multiclass cancer diagnosis using tumor gene expression signatures, PNAS, **98**, 15149-15154 **(2001)**.

[7] T.R.Golub, D.K.Slonim, P.Tamayo et al; Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring, Science, **286**, 531-537 **(1999)**.

[8] Heping Zhang, Chang-Yung Yu, Burton Singer; Cell and tumor classification using gene expression data: Construction of forests, PNAS, **100**, 4168-4172 **(2003)**.

[9] Alexander Statnikov, Constantin F.Aliferis, Ioannis Tsamardinos et al; A comprehensive evaluation of multicategory classification methods for microarray gene expression cancer diagnosis, BIOINFORMATICS, **21**, 631-643 **(2005)**.

[10] Jianjun Yu, Jindan Yu, Arpit A.Almal, et al; Feature Selection and Molecular Classification of Cancer Using Genetic Programming, Neoplasia, **9**, 292-303 **(2007)**.

[11] Yajie Liu, Xinling Shi, Zhenzhou An; Classification of Leukemia Gene Expression Data Using Particle Swarm Optimization, The Sixth International Conference on Genetic and Evolutionary Computing, Kitakyushu, Japan, **(2012)**.

[12] J.Kennedy, R.C.Eberhart; Particle Swarm Optimization, Proceedings of the 1995 IEEE International Conference on Neural Networks, **4**, 1942-1948 **(1995)**.

[13] Chia-Chong Chen, Two-layer particle swarm optimization for unconstrained optimization problems, Applied Soft Computing, 295-304 **(2011)**.

[14] Ahmedin Jemal, Freddie Bray, Melissa M.Center and et al, Global cancer statistics, CA CANCER, **61**, 69-90 **(2011)**.

[15] Yujin Hoshida, Jean-Philippe Brunet, Pablo Tamayo, et al, Subclass Mapping Identifying Common Subtypes in Independent Disease Data Sets, PLoS ONE, 1-8 **(2007)**.