Applied Soft Computing 24 (2014) 773–780

Contents lists available at ScienceDirect

Applied Soft Computing

journal homepage: www.elsevier.com/locate/asoc

Applying particle swarm optimization-based decision tree classifier for cancer classification on gene expression data

Kun-Huang Chen[a], Kung-Jeng Wang[a,*], Kung-Min Wang[b], Melani-Adrian Angelia[a]

[a] Department of Industrial Management, National Taiwan University of Science and Technology, Taipei 106, Taiwan, ROC
[b] Department of Surgery, Shin-Kong Wu Ho-Su Memorial Hospital, Taipei, Taiwan, ROC

ARTICLE INFO

Article history:
Received 6 April 2013
Received in revised form
21 December 2013
Accepted 13 August 2014
Available online 27 August 2014

Keywords:
Cancer classification
Gene expression
Particle swarm optimization
C4.5

ABSTRACT

Background: The application of microarray data for cancer classification is important. Researchers have tried to analyze gene expression data using various computational intelligence methods.
Purpose: We propose a novel method for gene selection utilizing particle swarm optimization combined with a decision tree as the classifier to select a small number of informative genes from the thousands of genes in the data that can contribute in identifying cancers.
Conclusion: Statistical analysis reveals that our proposed method outperforms other popular classifiers, i.e., support vector machine, self-organizing map, back propagation neural network, and C4.5 decision tree, by conducting experiments on 11 gene expression cancer datasets.

1. Introduction

The rapid development of microarray technology enables researchers to analyze thousands of genes simultaneously and obtain important information about the cell's function. This particular information can be used in cancer diagnosis and prognosis. However, given the characteristics of gene expression data (i.e., high dimension, high noise, and small sample size), the gene selection process remains challenging. A method for choosing the important subset of genes with high classification accuracy is needed to overcome this challenge. Such method would not only save computational costs, but will also enable doctors to identify a small subset of biologically relevant genes with certain cancers and target only a small number of genes in designing less expensive experiments [25]. Moreover, a highly accurate method can also assist in early diagnosis and drug discovery for cancer patients [2].

The method of gene selection generally falls into one of the following three categories: the filter, wrapper, and embedded approaches. The filter approach collects the intrinsic characteristics of genes in discriminating the targeted phenotype class and usually employs statistical methods, such as mutual information, statistical tests (-test, -test), and Wilcoxon's rank test, to directly select feature genes [47]. This approach is easily implemented, but ignores the complex interaction between genes. The "wrapper" approach [41] aims at selecting a subset of feature genes, typically with an induction algorithm to search for an initial gene subset which can then be used for further evaluating new feature gene subsets. The wrapper method is usually superior to the filter one since it involves inter-correlation of individual genes in a multivariate manner. The wrapper method can automatically determine the optimal number of feature genes for a particular classifier. The embedded method is similar to the wrapper method, while multiple algorithms can be combined in the embedded method to perform feature subset selection (Kahavi and John, 1997; [26]). In the embedded method, genetic algorithms (GAs) [58] are generally used as the search engine for feature subset, while other classification methods, such as KNN/GA (K nearest neighbors/genetic algorithms) [23], GA-SVM (genetic algorithms-support vector machine) [24], and so forth, are used to select feature subset. Estimation of distribution algorithm (EDA) [42] is a general framework of GA. Compared to traditional GA that employs crossover and mutation operators to create new population, EDA creates new populations by using a statistical approach to estimate the probability distribution of all promising individual solutions for the previous generation. EDA can also explicitly take into account specific interactions among the variables. When EDA is used to search for feature subsets, classification methods, such as Support vector machine (SVM) [6,36,18,59,13], which can deal with the high-dimension data in a limited sample space, can be used to select feature subsets.

[*] Corresponding author. Tel.: +886 2 2737 6769; fax: +886 2 2 737 6344.
E-mail address: kjwang@mail.ntust.edu.tw (K.-J. Wang).

http://dx.doi.org/10.1016/j.asoc.2014.08.032
1568-4946/© 2014 Elsevier B.V. All rights reserved.

Particle swarm optimization (PSO) is a popular meta-heuristic algorithm developed by [19]. PSO has been widely applied in many fields to solve various optimization problems, including gene selection [25,2,8,45,30]. In PSO, a swarm of particles with randomly initialized positions would move toward the optimal position along the search path that is iteratively updated based on the best particle position and velocity. The position of a particle can be used to represent a candidate solution for the problem. Among them, C4.5 is a decision tree-based classifier listed in the top 10 most influential data-mining algorithms in the research community [55]. Decision trees were a linear method as simple to understand and interpret.

This study proposes a method using the PSO algorithm to optimize the classification accuracy achieved using the C4.5 classifier (denoted as PSOC4.5). This study combines PSO for its excellent search capabilities and C4.5 for its knowledge interpretation advantage. This proposed hybrid technique combining PSO with C4.5 classifier has not been previously investigated by previous researchers. The performance of our proposed method is evaluated by testing the proposed method on 11 micro array datasets, which consist of 1 dataset from cancer patients of the National Health Insurance Research Database in Taiwan [34] and 10 from the Gene Expression Model Selector [15]. Moreover, we compare the performance of our proposed method with other well-known classifier algorithms, i.e., SVM, self-organizing map (SOM), back propagation neural network (BPNN), and C4.5. A statistical test is used to show that the proposed method outperforms other well-known classifiers in terms of classification accuracy.

The rest of the paper is organized as follows. In Section 2, we review the gene selection classification problem and related studies. Section 3 introduces the PSO algorithm and C4.5 classifier as the proposed approach. In Section 4, we present our experimental results and its comparison with those of other methods. Finally, we conclude the study in Section 5.

2. Overview of gene selection classification

2.1. Gene selection

DNA microarray is a technology that allows researchers to measure the expression levels of thousands of genes simultaneously in a single experiment. The method is usually used to compare the gene expression levels in tissues under different conditions, such as wild type versus mutant, or healthy versus diseased [12]. Ref. [33] propose a method for gene microarray classification that combines different feature reduction approaches for improving classification performance using a support vector machine (SVM) as our classifier. Their experiments were performed using several different datasets, and our results (expressed as both accuracy and area under the receiver operating characteristic (ROC) curve) show the goodness of the proposed approach with respect to the state of the art. Park [38] proposes a new approach for inferring combinatorial Boolean rules of gene sets for a better understanding of cancer transcriptome and cancer classification. To reduce the search space of the possible Boolean rules, we identify small groups of gene sets that synergistically contribute to the classification of samples into their corresponding phenotypic groups (such as normal and cancer).

In gene selection, we select the most informative genes, which are most predictive of its related class for classification. The gene selection process includes gene filtering, gene clustering, gene ranking, and gene extraction. Some basic numerical or statistical analysis, such as *t*-test, F-score, and standard deviation (Std.), are applied in filtering genes at the pre-procedure. Gene selection leads to reduced dimensions and improves classification performance.

Given the quantity and complexity of the gene expression data, an expert is unlikely to compute and compare the $n \times m$ gene expression matrix manually. Thus, machine learning and other artificial intelligence techniques have been widely used to classify or characterize gene expression data [5,6].

2.2. Related works

Several researchers have utilized the PSO algorithm to propose solutions for gene selection problems. Alba compared the use of PSO and genetic algorithm (GA), both augmented with SVM, as the classifier for high-dimensional microarray data. A modified PSO, namely, Geometric PSO, has been proposed for comparison with the GA on six public cancer datasets [2]. Li proposed a gene selection method by combining PSO with a GA and adopted SVM as the classifier. Their proposed approach was tested on three benchmark gene expression datasets: leukemia colon cancer, and breast cancer data. The aim of their hybrid method was to guide the proposed algorithm to prevent it from becoming trapped in the local optima [25]. Ref. [30] proposed an improved binary PSO combined with an SVM classifier to select a near-optimal subset of informative genes relevant to cancer classification. In the method by Mohamad et al., the existing rule for updating the particle position and velocity was modified.

Recently, Zhao proposed a novel hybrid framework (NHF) for gene selection and classification of high-dimensional microarray data, which combines information gain (IG), F-score, GA, PSO, and SVM. Three main steps comprise their proposed method. The performance of their proposed method was compared with those of the PSO-based, GA-based, ant colony optimization-based, and simulated annealing (SA)-based methods on five benchmark data sets: leukemia, lung carcinoma, and colon, breast, and brain cancers. The numerical results and statistical analysis showed that their proposed approach is capable of selecting a subset of predictive genes from a large noisy dataset and can capture the correlated structure in the data. Moreover, NHF performs significantly better than the other methods in terms of prediction accuracy with a smaller subset of features [57]. Ref. [51] propose an alternative to the existing methods for functionally annotating genes. The methodology involves building of classification models, validation and graphical representations of the results and reduction of the dimensions of the dataset. Ref. [8] used PSO + 1NN for feature selection. The proposed method was tested on eight benchmark datasets from UC Irvine Machine Learning Repository and then applied to an actual case of obstructive sleep apnea. The experimental results showed that the proposed method is significantly better than BPNN, logistic regression (LR), SVM, and C4.5. Further, [53] utilized a hybrid method combining GA and SVM as the classifier to identify the optimal subset of micro array datasets. The result they obtained from their proposed method was superior to those obtained by microPred and miPred. Heckerling [39] used ANN and genetic algorithms to develop models, based on sets of best predictor variables, for detecting urinary tract infection among women with urinary complaints Ref. [22] proposed a hybrid GA and PSO approach for designing a fuzzy based expert system. In their method, GA is used to find the rules and PSO is used to tune the membership function. They found that the proposed approach generated a compact fuzzy system with high classification accuracy for all six gene expression data sets when compared with other approaches.

Thus, PSO has been successfully applied to solve the gene selection problem. Therefore, we propose a method adopting a combination of PSO and C4.5 decision tree. Moreover, this work investigates the performance of this hybrid type of method, which has seldom been investigated.

3. Proposed method for gene selection

In this section, we describe PSO algorithm and C4.5 as the classifier algorithms (PSOC4.5).

We compare our result to several well-known classifiers, i.e., SVM, SOM, BPNN and C4.5. SVM is a classification method based on the statistical learning theory [10] and considered one of the most powerful machine learning classifiers. However, SVM is a black box system that does not provide insights on the reasons of a classification or explanations similar to artificial neural network (ANN). Moreover, SVM is kernel-based learning. SOM was introduced by Teuvo Kohonen in 1982. SOM is one of the ANN algorithms. SOM is unsupervised learning-based but can be used for supervised learning. BPNN is a common type of ANN, which is modeled on the biological nervous system. BPNN is capable of recognizing complex patterns in data. On the other hand, C4.5 is not a black box system; it is a rule-based classifier that creates a decision tree based on rules.

3.1. Particle swarm optimization

PSO is a population-based stochastic optimization technique inspired by the social behavior of swarm, such as bird flocking or fish schooling, to obtain a promising position to achieve certain objectives [19]. In PSO, each particle has a position, and moves based on an updated velocity. Each particle in a population has a fitness value computed from a fitness function. The main features of a particle in basic PSO are position, velocity, and ability to exchange information with its neighbors, ability to memorize a previous position, and ability to use information to make a decision. Given the easy implementation of PSO (through a few parameter adjustments), it has become a popular optimization algorithm that has been widely used in many fields to solve various problems, including gene selection. The flowchart of PSO algorithm is presented in Fig. 1.

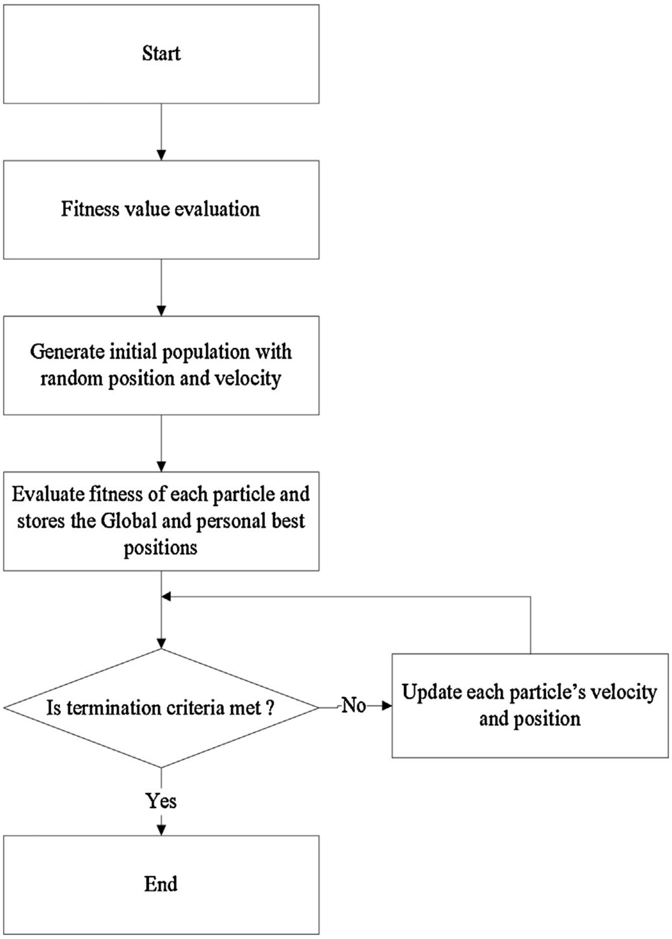
Start
Fitness value evaluation
Generate initial population with random position and velocity
Evaluate fitness of each particle and stores the Global and personal best positions
Is termination criteria met ?
No
Update each particle's velocity and position
Yes
End

Fig. 1. PSO algorithm.

In gene selection, a particle represents a potential solution (i.e., gene subset) in an n-dimensional space. The particle movements are directed by the position vector and velocity vector of each particle. In the n-dimensional space, the vector and velocity vector of the ith particle position are represented as $X_i = [x_i^1, x_i^2, \ldots, x_i^n]$, and $V_i = [v_i^1, v_i^2, \ldots, v_i^n]$, respectively, where x_i^d is a binary bit, i = 1, 2, . . ., m (m is the number of particles), and d = 1, 2, . . ., n (n is the dimension of data). The record of the position of the previous best performance of a particle is $PB_i = [pb_i^1, pb_i^2, \ldots, pb_i^n]$ and the best performance so far in the neighborhood is $GB_i = [gb_i^1, gb_i^2, \ldots, gb_i^n]$. The particle velocity and position is updated based on Eqs. (1) and (2), respectively:

$$v_{id}^{\text{new}} = w \times v_{id}^{\text{old}} + c_1 r_1 (pb_{id}^{\text{old}} - x_{id}^{\text{old}}) + c_2 r_2 (gb_d^{\text{old}} - x_{id}^{\text{old}}), \quad d = 1, 2, \ldots, D \tag{1}$$

$$x_{id}^{\text{new}} = x_{id}^{\text{old}} + v_{id}^{\text{new}}, \quad d = 1, 2, \ldots, D; \quad i = 1, 2, \ldots, N \tag{2}$$

where c_1 and c_2 are positive constant values between 0 and 4, indicating the cognitive and the social learning factors, respectively. The inertia weight (w) has a value between 0.4 and 0.9 and r_1 and r_2 are uniformly distributed numbers whose values are between 0 and 1. The values of the velocities are between $v_{\min}$ and $v_{\max}$. N is the size of the swarms.

3.2. The C4.5 classifier

The decision tree algorithm is well known for its robustness and learning efficiency with a learning time complexity of $O(n \log 2n)$ [50]. C4.5 has been listed in the top 10 algorithms in data mining [55]. It is a popular statistical classifier developed by Ross Quinlan in 1993. Basically, C4.5 is an extension of Quinlan's earlier ID3 algorithm. In C4.5 the Information Gain split criterion is replaced by an Information Gain Ratio criterion which penalizes variables with many states. C4.5 can be used to generate a decision tree for classification. The learning algorithm applies a divide-and-conquer strategy [40] to construct the tree. The sets of instances are accompanied by a set of genes (attributes). This classifier has additional features, such as handling missing values, categorizing continuous attributes, pruning decision trees, deriving rules, and others.

The Information gain (S, A) of a feature A relative to a collection of examples S, is defined as:

$$\text{Gain}(S, A) = \text{Entropy}(S) - \sum_{v \in \text{Value}(A)} \frac{|S_v|}{|S|} \text{Entropy}(S_V) \tag{3}$$

where Values(A) is the set of all possible values for attribute A, and S_v is the subset of S for which feature A has value v (i.e., $S_v = \{s \varepsilon S | A(s) = v\}$).

Note the first term in the equation for Gain is just the entropy of the original collection S and the second term is the expected value of the entropy after S is partitioned using feature A. The expected entropy described by the second term is the direct sum of the entropies of each subset S_v, weighed by the fraction of samples $|S_v|/|S|$ that belong to S_v. Gain (S, A) is therefore the expected

Table 1
Summarization of micro array dataset characteristics.

No.	Data set	# of features	# of samples	# of classes	# of particles
1	11_Tumors	12,601	203	5	126
2	14_Tumors	10,368	50	4	104
3	9_Tumors	10,368	50	4	104
4	Brain_Tumor1	5727	60	9	57
5	Brain_Tumor2	12,534	174	11	125
6	Leukemia2	5328	72	3	53
7	Lung_Cancer	11,226	72	3	112
8	SRBCT	5470	77	2	55
9	Prostate_Tumor	10,510	102	2	105
10	DLBCL	83	2309	4	10

reduction in entropy caused by knowing the value of feature A. The Entropy is given by

$$\text{Entropy}(S) = \sum_{i=1}^{c} -P_i \log_2 P_i \tag{4}$$

3.3. The proposed approach

We integrated PSO algorithm with the C4.5 classifier to address the gene selection problem. Fig. 2 shows the pseudo code of applying PSOC4.5 on gene selection. The important gene was selected using PSO algorithm, and C4.5 was employed as a fitness function of the PSO algorithm to verify the efficiency of the selected genes. The proposed approach is described in more detail below.

```
GENE_SELECTION ($c_1$, $c_2$, $r_1$, $r_2$, $w$, $v_{min}$, $v_{max}$)
Begin
 INITIALIZE_POPULATION();
 While (max iteration or convergence criteria is not met)
  For ($i = 1$ ; $i <$ (numbers of particles) ; $i$++)
   fitness_value = EVALUATE_FITNESS (the particle evaluated by C4.5)
  If (fitness value of $X_i$ is greater than that of $PB_i$)
   $PB_i = X_i$
  End if
  If (fitness value of $X_i$ is greater than that of $GB$)
   $GB = X_i$
  End if
   For ($d = 1$ ; $d <$ (no of genes) ; $d$++ )
    $v_{id}^{new} = w \times v_{id}^{old} + c_1 r_1 (pb_{id}^{old} - x_{id}^{old}) + c_2 r_2 (gb_d^{old} - x_{id}^{old})$
    If $v_{id}^{old} > v_{max}$
      $v_{id}^{new} = v_{max}$
    End if
    If $v_{id}^{old} < v_{max}$
      $v_{id}^{new} = v_{minx}$
    End if
    If $sigmoid(v_{id}^{new}) > U(0,1)$
     Then
      $x_{id}^{new} = 1$
     Else
      $x_{id}^{new} = 0$
    End if
   Next $d$
  Next $i$
 End while
End
```

Fig. 2. Pseudo code of our proposed method for gene selection.

3.3.1. Solution representation

In gene selection, a particle represents a potential solution (i.e., gene subset) in an n-dimensional space. The particles are represented using a binary bit string with length n, where n is the total number of genes. Bit value 1 represents a selected gene, whereas bit 0 represents an unselected gene. For instance, the particle 101,010 with six gene expression data indicates that the first, third, and fifth genes are selected. In addition, we updated the dimension d of particle i according to Eq. (5):

$$x_{id}^{\text{new}} = \begin{cases} 1, & \text{if } \operatorname{sigmoid}(v_{id}^{\text{new}}) > U(0,1) \\ 0, & \text{otherwise} \end{cases} \tag{5}$$

where U is a uniformly distributed random variable, and the sigmoid (v_{id}^{new}) is $1/1 + e^{-v_{id}^{\text{new}}}$.

3.3.2. Population initialization

We used a random or well-adapted PSO to initialize the particle population. However, we assumed that seeding PSO with a good initial can lead to better results. In this study, a probability of 0.5 was randomly assigned to bit values 0 and 1. If $U(0, 1) > 0.5$, then $x_{id}^0 = 1$; otherwise, $x_{id}^0 = 0$.

3.3.3. Particle fitness function

Fitness function plays the most important role in PSO search in this study. This function evaluates the goodness of each particle in a population. The input of the fitness function is a particle and it returns a numerical evaluation representing the goodness of the feature subset.

Fitness function helps us identifying a specific set of features. We use C4.5 decision tree to build a model from the selected feature subset and then evaluate the learned model. When a PSO generates a particular feature subset, we run the C4.5 algorithm. As a result, we have a classifier in the form of a decision tree which is evaluated. We perform this procedure five times using 5-fold cross validation. In the validation, the dataset is randomly divided into five equally sized, mutually exclusive subsets. Each of the subset is used once for testing and the remaining four are used for training. The results of these five runs are then averaged and this average is used as fitness of the feature subset.

In this proposed approach, we used C4.5 error prediction rate (accuracy) as the fitness function of PSO. Accuracy has been the most commonly used metric for assessing the performance of classifiers for years [54]. Accuracy is calculated by Eq. (6):

$$\text{Accuracy} = \frac{\#\text{ of correctly predicted examples}}{\#\text{ of examples}} \tag{6}$$

Table 2
Classification accuracy for 10 micro array datasets (%).

Data set	Fold	SVM	SOM	BPNN	C4.5	PSOC4.5
11_Tumors	1	93.00	77.50	72.50	82.50	92.50
	2	90.68	82.93	68.29	92.68	97.56
	3	95.56	70.73	65.85	95.12	100.00
	4	100.00	90.12	74.07	98.77	100.00
	5	90.00	72.50	65.00	82.50	97.56
	Avg.	93.85	78.76	69.14	90.31	97.52
	(Std.)	(4.07)	(7.93)	(4.01)	(7.46)	(3.06)
14_Tumors	1	70.00	40.00	30.00	60.00	80.00
	2	50.00	30.00	40.00	40.00	60.00
	3	80.00	60.00	30.00	40.00	80.00
	4	70.00	30.00	20.00	40.00	80.00
	5	30.00	40.00	20.00	60.00	70.00
	Avg.	60.00	40.00	28.00	48.00	74.00
	(Std.)	(20.00)	(12.25)	(8.37)	(10.95)	(8.49)
9_Tumors	1	80.00	50.00	60.00	60.00	80.00
	2	60.00	30.00	30.00	50.00	60.00
	3	70.00	30.00	30.00	60.00	80.00
	4	80.00	40.00	10.00	40.00	70.00
	5	90.00	50.00	40.00	50.00	80.00
	Avg.	76.00	40.00	34.00	52.00	74.00
	(Std.)	(11.40)	(10.00)	(18.17)	(8.37)	(8.94)
Brain_Tumor1	1	50.00	0.00	8.33	41.67	60.77
	2	50.00	16.67	16.67	66.67	60.77
	3	56.33	41.67	33.33	58.33	71.67
	4	50.00	25.00	16.67	25.00	45.92
	5	39.67	8.33	0.00	16.67	42.59
	Avg.	49.20	18.33	15.00	41.67	56.34
	(Std.)	(5.99)	(16.03)	(12.36)	(21.25)	(11.96)
Brain_Tumor2	1	77.41	50.00	23.53	70.59	88.24
	2	75.14	51.43	28.57	65.71	78.57
	3	80.86	40.00	22.86	77.14	84.43
	4	92.29	48.57	14.29	82.86	92.22
	5	89.43	60.00	37.14	74.29	85.27
	Avg.	83.03	50.00	25.28	74.12	85.75
	(Std.)	(7.50)	(7.14)	(8.38)	(6.49)	(5.04)
Leukemia2	1	90.86	57.14	50.00	92.86	100.00
	2	100.00	86.67	46.67	93.33	100.00
	3	100.00	53.33	46.67	86.67	100.00
	4	100.00	42.86	57.14	100.00	100.00
	5	78.57	64.29	50.00	78.57	100.00
	Avg.	93.89	60.86	50.10	90.29	100.00
	(Std.)	(9.43)	(16.37)	(4.28)	(8.07)	(0.00)
Lung_Cancer	1	100.00	78.57	71.43	85.71	100.00
	2	100.00	66.67	26.67	86.67	100.00
	3	91.33	60.00	33.33	93.33	100.00
	4	100.00	78.57	35.71	92.86	100.00
	5	100.00	57.14	28.57	92.86	100.00
	Avg.	98.27	68.19	39.14	90.29	100.00
	(Std.)	(3.88)	(10.09)	(18.41)	(3.76)	(0.00)
SRBCT	1	100.00	73.33	73.33	86.67	100.00
	2	100.00	50.00	93.75	68.75	80.00
	3	91.75	75.00	75.00	81.25	100.00
	4	91.33	73.33	80.00	66.67	82.45
	5	100.00	66.67	93.33	86.67	100.00
	Avg.	96.62	67.67	83.08	78.00	92.49
	(Std.)	(4.64)	(10.38)	(9.86)	(9.68)	(10.32)
Prostate_Tumor	1	90.00	65.00	60.00	80.00	95.00
	2	90.48	66.67	76.19	95.24	95.24
	3	83.71	61.90	61.90	90.48	95.24
	4	83.00	75.00	45.00	85.00	90.00
	5	93.00	65.00	40.00	90.00	95.00
	Avg.	88.04	66.71	56.62	88.14	94.10
	(Std.)	(4.43)	(4.94)	(14.44	(5.82)	(2.29)
DLBCL	1	50.00	37.50	18.75	56.25	73.25
	2	92.12	52.94	58.82	70.59	92.42
	3	100.00	76.47	35.29	94.12	100.00
	4	100.00	52.94	35.29	88.24	98.32
	5	100.00	37.50	31.25	87.50	95.43
	Avg.	88.42	51.47	35.88	79.34	91.88
	(Std.)	(21.75)	(15.97)	(14.51)	(15.60)	(10.81)

4. Experiments and results

This section focuses on the results of the proposed method. First, we described the datasets and experimental setups, after which we presented the results of the experiment and the comparison with other well-known classifier algorithms.

4.1. Environment

The performance of our proposed method was evaluated by conducting numerical experiments using 10 micro array cancer datasets with diverse sizes, features, and classes. The 10 datasets from GEMS. The datasets from GEMS included 11_Tumors, 14_Tumors, 9_Tumors, Brain_Tumor1, Brain_Tumor2, Leukemia2, Lung_Cancer, SRBCT, Prostate_Tumor, and DLBCL. According to the data from the Taiwan Cancer Registry [48], these types of cancer belong in the top 10 in terms of cancer incidences and deaths in Taiwan in 2008. We adopted a fivefold cross-validation strategy to guarantee the impartial comparison of the classification results and avoid generating random results. Table 1 summarizes the micro array dataset characteristics. Each dataset is tested independently to determine the occurrence of a specific cancer. The method shows its classification power under different characteristics of micro array gene expression. The datasets contain all the benign and malignant tissue genes, whereas only the malignant tissue genes cause cancers. We intend to select datasets with cancers at different sites in body, such as brain and lung, to justify that our proposed method can be successfully applied to identify high risk genes for cancers and tumors.

Our proposed method was coded in Visual C# 2008 and run in a Core 2 Duo E6600 (2.4 GHz) PC equipped with 2048 MB of RAM under a Windows XP environment. The parameters we used for PSO are stated as follows: the number of particles in the population was set to the number of genes. The c1 and c2 were both set at 2, whereas the lower ($v_{\min}$) and upper bounds ($v_{\max}$) were set at −4 and 4, respectively. The inertia weight (w) was set at 0.4. The process was repeated until either the fitness of the given particle was 1.0 or the number of the iterations was achieved by the default value of T, which was 100. Here, the PSO parameter design was chosen by a survey on several related research articles concerning the utilization of PSO. This parameter configuration has been adopted mostly as PSO parameter settings (e.g., [46,20,17]). Moreover, we conducted many trials to test such parameter setting which gives the best objective value.

4.2. Numerical experiments

The effectiveness of our proposed method was evaluated by comparing its accuracy with the other four popular classification methods, namely, SVM, SOM, BPNN, and C4.5.

Based on the experiments, we evaluated the performance of our proposed algorithm in terms of the accuracy rate as tested on the 11 datasets. Table 2 lists the accuracy of our proposed method compared with the other methods. We applied fivefold cross-validation on the datasets and obtained the average and standard deviations. The proposed method was superior to all the compared methods. Hence, the utilization of PSO combined with the C4.5 classifier yielded a higher accuracy compared with a standard C4.5 classifier. Fig. 3 shows a 95% confidence interval for the mean classification accuracy (with respect to the 11 micro array datasets). The figure clearly shows that PSOC4.5 outperformed the other algorithms. A statistical test was used to compare our proposed method with the other well-known classifier algorithms. We used two-way ANOVA to determine whether the five algorithms were significantly different in terms of average classification accuracy. The classification algorithms were defined as "factor", whereas the datasets were defined as “block”. Table 3 lists the results of ANOVA for average classification accuracy. The results showed significant differences in classification accuracy, and the 11 datasets demonstrated various types of average classification accuracy. Fisher's least significant difference was used to determine if each pair of the five algorithms differed from each other. Table 4 lists the p-value of each pair. The p-values demonstrate that our proposed method exhibits differences in mean classification accuracy compared with the other algorithms, except for SVM.

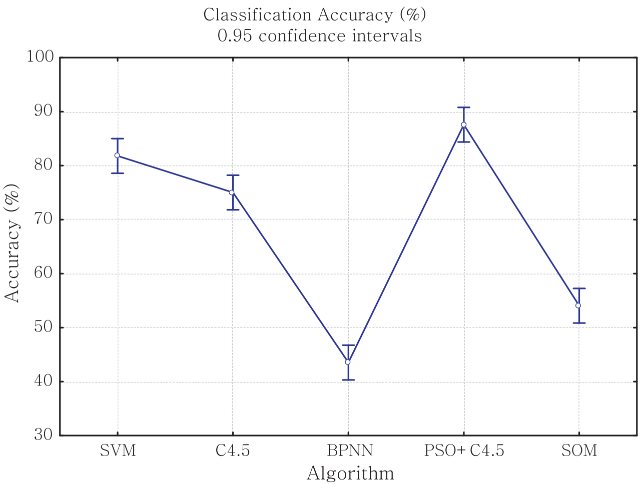

Fig. 3. Confidence interval 95% of the mean for classification accuracy.

4.3. A clinical case study

The clinical practice data consisted of 13 cancers on 5335 samples from actual top cancer cases in the [48]. Raw intensity data (CEL files) generated using Affymetrix HG-U133A and HG-U133 plus 2.0 platforms were retrieved from Gene expression omnibus (GEO) and ArrayExpress. Arrays performed using samples other than human clinical specimens, such as cell lines, primary cells, and transformed cells, were excluded.

All microarray raw data were pre-processed using three different algorithms: Affymetrix Microarray Suite 5 (MAS5), robust multi-chip average (RMA), and GC-robust multi-chip average (GCRMA) as implemented in the Bioconductor packages. RMA and GCRMA processed data on a multi-array basis; therefore all arrays of the same platform were uniformly pre-processed to reduce variance. The samples of each cancers is shown in Table 5. The cancer micro array = consisted of 13 cancer types, namely, bladder, blood,

Table 3
ANOVA for average classification accuracy.

Source	DF	SS	MS	F	p
Algorithm	4	77,003	19,251	132.268	0.00
Dataset	10	410.00	6,245	42.911	0.00
Error	260	37,841	146		
Total	274	177,298			

Table 4
p-value of multiple comparison for average classification accuracy.

	SVM	SOM	BPNN	C4.5	PSOC4.5
SVM		0.000000	0.000000	0.001410	0.066826
SOM			0.000007	0.000000	0.000000
BPNN				0.000000	0.000000
C4.5					0.000001
PSO + C4.5					

Table 5
The arrays of cancers.

No.	Cancer	# of samples
1	Bladder	94
2	Blood	503
3	Bone marrow	676
4	Brain	250
5	Breast	1817
6	Cervix uteri	148
7	Colon	437
8	Kidney	301
9	Liver	107
10	Lung	222
11	Lymph node	280
12	Ovary	331
13	Prostate	166

bone marrow, brain, breast, cervix uterus, colon, kidney, liver, lung, lymph node, ovary, and prostate.

Table 6 shows the test results on the arrays. The obtained accuracy for PSOC4.5 and SVM was 97.26 and 72.46, respectively. The results indicated that PSOC4.5outperformed the SVM and other methods.

4.4. Biological roles of selected genes

Five independent sets of cDNA clones (refer to Supplementary Tables 1–5 of Appendix) were selected to perform a fivefold cross-validation. We confirm a small number of genes that were selected multiple times (more than 4) (refer to Supplementary Tables 1 and 2). The expression levels of these genes provide a high discrimination power for the tumors of different anatomical origin. Thus, these genes are likely to be tissue-specific genes. Further, such expression differences may result from organ or tissue-specific malignant transformation. The findings are consulted with medical experts and doctors, as well as confirmed by academic literatures. The details are given as follows.

Among the 18 genes (or 19 cDNA clones) PAX8, HSPA6, SERPINB3, and KLK2 genes were identified. We found that two independent cDNA clones, 207,924_x_at and 120_at, which correspond to the same gene (PAX8) were selected four and five times, respectively. PAX8 exhibits relatively high levels of expression in normal human thyroid, ovary and kidney tissues; therefore, it has been identified as a tissue-specific gene by previous reports [37,56,7]. PAX8 was identified to involve in several cancers such as thyroid [4], ovarian tumor [29], and neuroendocrine tumor [27]. It has also been shown that PAX8 expression is a reliable marker that can effectively distinguish between tumors of different anatomical origin [14,44,49]. Similar findings have been reported on SERPINB3 and KLK2 genes [31,52]. KLK2 has been found in prostate cancer [32,9,21] and HSPA6 was identified to involve in colon cancer [35]. Further, we also noticed that three genes on this 18 gene-list, TIMM17A, RBL2, and ACTR2, were concluded to be stably expressed housekeeping genes [7,56]. Although these three genes were constitutively expressed in normal tissues with low expression variations, prior studies reported that their expression levels in different types and/or stages of tumors are significantly different. TIMM17A is highly expressed in breast cancer tissues and its expression level is significantly associated with unfavorable clinical outcomes [43]. RBL2/p130 gene was identified as a member of the retinoblastoma family which plays crucial roles in tumor suppression [11]. Ref. [3] reported that RBL2 is ubiquitously expressed in normal tissue, but its expression has an inverse correlation with the histological grading of lung cancer. ACTR2 encodes a protein which is a major constituent of Apr2/3 complex that plays key roles in actin polymerization [28]. Ref. [1] revealed an elevated expression of ACRT2 in head and neck squamous cancer cell lines relative to normal cells. Therefore, the expression patterns of TIMM17A, RBL2, and ACTR2 genes were markedly altered during the process of malignant transformation. Whether this expression alteration is related to specific organs remained to be elucidated.

Table 6
Classification accuracy for 13 cancer micro array (%).

Data set	Fold	SVM	SOM	BPNN	C4.5	PSOC4.5
13 cancer	1	72.16	50.89	72.50	92.60	97.84
	2	73.29	54.83	34.02	93.63	97.38
	3	73.25	54.64	37.77	93.63	96.72
	4	71.42	50.98	34.49	93.16	97.09
	5	72.16	51.64	34.11	92.69	97.28
	Avg.	72.46	52.60	42.58	93.14	97.26
	(Std.)	(0.80)	(1.98)	(16.80)	(0.49)	(0.41)

5. Conclusion

In this paper, we proposed a novel method in gene selection using a PSO algorithm combined with a C4.5 classifier. Our proposed PSOC4.5 method was presented and compared with other well-known classifiers. Eleven cancer datasets were used to test the performance of the proposed method. We used a fivefold cross-validation method to justify the performance of our proposed method. Our proposed method achieved a higher accuracy compared with all the other methods. Two-way ANOVA was utilized to prove statistically that the performance of our proposed method was superior over all methods.

This research have four contributions. First, this study pioneered the utilization of a hybrid method combining PSO and C4.5 classifier and its application to microarray cancer datasets in gene selection. Second, the PSO algorithm was successfully applied to the binary applications and combined with C4.5 as classifier, providing a solution to problems of classification and gene selection. Third, in this research, we compared our proposed method with other well-known algorithms using a variety of datasets with diverse sizes and numbers of classes and features. Our proposed method outperformed all the other methods. Fourth, we found several types of genes that were selected multiple times, i.e., PAX8, HSPA6, SERPINB3 and KLK2. These genes have been identified to have high risk on several cancers and tumors.

Further investigation on PSO parameter adjustment and the local optima trapping problem needs to be carried out. A hybrid method with GA utilizing the mutation operator can help diversify the particles or introduce the craziness concept as a solution to the local optima drawback.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.asoc.2014.08.032.

References

[1] M.T. Abraham, M.A. Kuriakose, P.G. Sacks, H. Yee, L. Chiriboga, E.L. Bearer, M.D. Delacure, Motility-related proteins as markers for head and neck squamous cell cancer, Laryngoscope 111 (7) (2001) 1285–1289.

[2] E. Alba, J. García-Nieto, L. Jourdan, E.G. Talbi, Gene selection in cancer classification using PSO/SVM and GA/SVM hybrid algorithms, in: Proceedings of IEEE Congress on Evolutionary Computation, CEC 2007, 2007, pp. 284–290.

[3] A. Baldi, V. Esposito, A. De Luca, Y. Fu, I. Meoli, G.G. Giordano, M. Caputi, F. Baldi, A. Giordano, Differential expression of Rb2/p130 and p107 in normal human tissues and in primary lung cancer, Clin. Cancer Res. 3 (1) (1997) 1691–1697.

[4] J.A. Bishop, R. Sharma, W.H. Westra, PAX8 immunostaining of anaplastic thyroid carcinoma: a reliable means of discerning thyroid origin for undifferentiated tumors of the head and neck, Hum. Pathol. 42 (12) (2011) 1873–1877.

[5] A. Brazma, J. Vilo, Gene expression data analysis, FEBS Lett. 480 (1) (2000) 2–16.

[6] M.P. Brown, W.N. Grundy, D. Lin, N. Cristianini, C.W. Sugnet, T.S. Furey, M. Ares Jr., D. Haussler, Knowledge-based analysis of microarray gene expression data by using support vector machines, Proc. Natl. Acad. Sci. 97 (1) (2000) 262–267.

[7] C.W. Chang, W.C. Cheng, C.R. Chen, W.Y. Shu, M.L. Tsai, C.L. Huang, I.C. Hsu, Identification of human housekeeping genes and tissue-selective genes by micro-array meta-analysis, PLoS ONE 6 (7) (2011) e22859.

[8] L.F. Chen, C.T. Su, K.H. Chen, P.C. Wang, Particle swarm optimization for feature selection with application in obstructive sleep apnea diagnosis, Neural Comput. Appl. 21 (8) (2011) 2087–2096.

[9] C.H. Chiang, C.J. Hong, Y.H. Chang, L.S. Chang, K.K. Chen, Human kallikrein-2 gene polymorphism is associated with the occurrence of prostate cancer, J. Urol. 173 (2005) 429–432.

[10] C. Cortes, V. Vapnik, Support-vector networks, Mach. Learn. 20 (1995) 273–297.

[11] G. DeFalco, A. Giordano, pRb2/p130: a new candidate for retinoblastoma tumor formation, Oncogene 25 (38) (2006) 5333–5340.

[12] S. Dudoita, J. Fridlyanda, T.P. Speeda, Comparison of discrimination methods for the classification of tumors using gene expression data, J. Am. Stat. Assoc. 97 (457) (2002) 77–86.

[13] L. Evers, C.M. Messow, Sparse kernel methods for high-dimensional survival data, Bioinformatics 24 (14) (2008) 1632–1638.

[14] M. Fujiwara, J. Taube, M. Sharma, T.H. McCalmont, J. Kim, PAX8 discriminates ovarian metastases from adnexal tumors and other cutaneous metastases, J. Cutan. Pathol. 37 (9) (2010) 938–943.

[15] GEMS Dataset, 2012, http://www.gems-system.org/

[17] C.W. Hsu, C.J. Lin, A comparison of methods for multi-class support vector machines, IEEE Trans. Neural Netw. 12 (2002) 415–425.

[18] S. Hua, Z. Sun, A novel method of protein secondary structure prediction with high segment overlap measure: support vector machine approach, J. Mol. Biol. 308 (2) (2001) 397–407.

[19] J. Kennedy, R. Eberhart, Particle swarm optimization, Proc. IEEE Int. Conf. Neural Netw. 4 (1995) 1942–1948.

[20] J. Kennedy, R. Eberhart, Y. Shi, Swarm Intelligence, Morgan Kaufmann, San-Mateo, CA, 2001.

[21] M. Kohli, P.G. Rothberg, C. Feng, E. Messing, J. Joseph, S.S. Rao, A. Hendershot, D. Sahsrabudhe, Exploratory study of a KLK2 polymorphism as a prognostic marker in prostate cancer, Cancer Biomark. 7 (2) (2010) 101–108.

[22] P.G. Kumar, T.A.A. Victoire, P. Renukadevi, D. Devaraj, Design of fuzzy expert system for microarray data classification using a novel genetic swarm algorithm, Expert Syst. Appl. 39 (2) (2012) 1811–1821.

[23] L. Li, T.A. Darden, C.R. Weinberg, A.J. Levine, L.G. Pedersen, Gene assessment and sample classification for gene expression data using a genetic algorithm/k-nearest neighbor method, Comb. Chem. High Throughput Screen. 4 (8) (2001) 727–739.

[24] L. Li, W. Jiang, X. Li, K.L. Moser, Z. Guo, L. Du, Q. Wang, E.J. Topol, Q. Wang, S. Rao, A robust hybrid between genetic algorithm and support vector machine for extracting an optimal feature gene subset, Genomics 85 (1) (2005) 16–23.

[25] S. Li, X. Wu, M. Tan, Gene selection using hybrid particle swarm optimization and genetic algorithm, Soft Comput. 12 (11) (2008) 1039–1048.

[26] X. Li, S. Rao, Y. Wang, B. Gong, Gene mining: a novel and powerful ensemble decision approach to hunting for disease genes using microarray expression profiling, Nucleic Acids Res. 32 (9) (2004) 2685–2694.

[27] P.I. Lorenzo, C.M. Jimenez Moreno, I. Delgado, N.C. Vuilleumier, R. Meier, L.G. Izquierdo, T. Berney, R.G. Carbonero, A. Rojas, B.R. Gauthier, Immunohistochemical assessment of Pax8 expression during pancreatic islet development and in human neuroendocrine tumors, Histochem. Cell Biol. 136 (5) (2011) 595–607.

[28] L.M. Machesky, K.L. Gould, The Arp2/3 complex: a multifunctional actin organizer, Curr. Opin. Cell Biol. 11 (1) (1999) 117–121.

[29] R. McKnight, C. Cohen, M.T. Siddiqui, Utility of paired box gene 8 (PAX8) expression in fluid and fine-needle aspiration cytology: an immunohistochemical study of metastatic ovarian serous carcinoma, Cancer Cytopathol. 118 (2010) 298–302.

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

[31] R.K. Nam, W.W. Zhang, L.H. Klotz, J. Trachtenberg, M.A. Jewett, J. Sweet, A. Toi, S. Teahan, V. Venkateswaran, L. Sugar, A. Loblaw, K. Siminovitch, S.A. Narod, Variants of the hK2 protein gene (KLK2) are associated with serum hK2 levels and predict the presence of prostate cancer at biopsy, Clin. Cancer Res. 12 (21) (2006) 6452–6458.

[32] R.K. Nam, W.W. Zhang, J. Trachtenberg, M.A. Jewett, M. Emami, D. Vesprini, W. Chu, M. Ho, J. Sweet, A. Evans, A. Toi, M. Pollak, S.A. Narod, Comprehensive assessment of candidate genes and serological markers for the detection of prostate cancer, Cancer Epidemiol. Biomark. Prev. 12 (12) (2003) 1429–1437.

[33] L. Nanni, S. Brahnam, A. Lumini, Combining multiple approaches for gene microarray classification, Bioinformatics 28 (8) (2008) 1151–1157.

[34] National Health Insurance Research Database, 2012, http://nhird.nhri.org.tw/

[35] E.J. Noonan, R.F. Place, D. Pookot, S. Basak, J.M. Whitson, H. Hirata, C. Giardina, R. Dahiya, miR-449a targets HDAC-1 and induces growth arrest in prostate cancer, Oncogene 28 (14) (2009) 1714–1724.

[36] J.H. Oh, J. Gao, A kernel-based approach for detecting outliers of high-dimensional biological data, BMC Bioinformatics 10 (4) (2009) S7.

[37] M.P. di Magliano, R. Di Lauro, M. Zannini, Pax8 has a key role in thyroid cell differentiation, Proc. Natl. Acad. Sci. U. S. A. 97 (24) (2000) 13144–13149.

[38] I. Park, K.H. Lee, D. Lee, Inference of combinatorial Boolean rules of synergistic gene sets from cancer microarray datasets, Bioinformatics 26 (12) (2010) 1506–1512.

[39] P.S. Heckerling, G.J. Canaris, S.D. Flach, T.G. Tape, R.S. Wigton, B.S. Gerber, Predictors of urinary tract infection based on artificial neural networks and genetic algorithms, Int. J. Med. Inform. 76 (4) (2007) 289–296.

[40] J.R. Quinlan, C4.5: Programs for Machine Learning, Morgan Kaufmann, Los Altos, 1993.

[41] R. Kohavi, G.H. John, Wrappers for feature subset selection, Artif. Intell. 97 (1–2) (1997) 273–324.

[42] Y. Saeys, S. Degroeve, D. Aeyels, P. Rouzé, Y. Van de Peer, Feature selection for splice site prediction: a new method using EDA-based feature ranking, BMC Bioinformatics 5 (64) (2004) 1–11.

[43] M. Salhab, N. Patani, W. Jiang, K. Mokbel, High TIMM17A expression is associated with adverse pathological and clinical outcomes in human breast cancer, Breast Cancer 19 (2) (2012) 153–160.

[44] A.R. Sangoi, R.S. Ohgami, R.K. Pai, A.H. Beck, J.K. McKenney, R.K. Pai, PAX8 expression reliably distinguishes pancreatic well-differentiated neuroendocrine tumors from ileal and pulmonary well-differentiated neuroendocrine tumors and pancreatic acinar cell carcinoma, Mod. Pathol. 24 (3) (2011) 412–424.

[45] Q. Shen, M.W. Shi, W. Kong, Hybrid particle swarm optimization and tabu search approach for selecting genes for tumor classification using gene expression data, Comput. Biol. Chem. 32 (1) (2008) 53–60.

[46] Y. Shi, R.C. Eberhart, A Modified Particle Swarm Optimizer, in: Proceeding of IEEE International Conference on Evolutionary Computation, Anchorage, 1998, pp. 69–73.

[47] Y. Su, T.M. Murali, V. Pavlovic, M. Schaffer, S. Kasif, RankGene: identification of diagnostic genes based on expression data, Bioinformatics 19 (12) (2003) 1578–1579.

[48] Taiwan Cancer Registry, 2012, http://tcr.cph.ntu.edu.tw

[49] D. Tacha, D. Zhou, L. Cheng, Expression of PAX8 in normal and neoplastic tissues: a comprehensive immunohistochemical study, Appl. Immunohistochem. Mol. Morphol. 19 (4) (2011) 293–299.

[50] A.C. Tan, D. Gilbert, Ensemble machine learning on gene expression data for cancer classification, Appl. Bioinform. 2 (3) (2003) S75–S83.

[51] T. Theodosiou, L. Angelis, A. Vakali, G.N. Thomopoulos, Gene functional annotation by statistical analysis of biomedical articles, Int. J. Med. Inform. 76 (8) (2007) 601–613.

[52] C. Turato, M.A. Buendia, M. Fabre, M.J. Redon, S. Branchereau, S. Quarta, M. Ruvoletto, G. Perilongo, M.A. Grotzer, A. Gatta, P. Pontisso, Over-expression of SERPINB3 in hepatoblastoma: a possible insight into the genesis of this tumour? Eur. J. Cancer 48 (8) (2012) 1219–1226.

[53] Y. Wang, X. Chen, W. Jiang, L. Li, W. Li, L. Yang, M. Liao, B. Lian, Y. Lv, S. Wang, S. Wang, X. Li, Predicting human microRNA precursors based on an optimized feature subset generated by GA-SVM, Genomics 98 (2) (2011) 73–78.

[54] I.H. Witten, E. Frank, Data Mining: Practical Machine Learning Tools and Techniques, Morgan Kaufmann, San-Mateo, CA, 2005.

[55] X. Wu, V. Kumar, J.R. Quinlan, J. Ghosh, Q. Yang, H. Motoda, G.F. McLachlan, A. Ng, B. Liu, P.S. Yu, Z.H. Zhou, M. Steinbach, D.J. Hand, D. Steinberg, Top 10 algorithms in data mining, Knowl. Inform. Syst. 14 (1) (2008) 1–37.

[56] S.J. Xiao, C. Zhang, Q. Zou, Z.L. Ji, TiSGeD: a database for tissue-specific genes, Bioinformatics 26 (9) (2010) 1273–1275.

[57] W. Zhao, G. Wang, H.B. Wang, H.L. Chen, H. Dong, Z.D. Zha, A novel framework for gene selection, Int. J. Adv. Comput. Technol. 3 (2011) 184–191.

[58] X.M. Zhao, Y.M. Cheung, D.S. Huang, A novel approach to extracting features from motif content and protein composition for protein sequence classification, Neural Netw. 18 (2005) 1019–1028.

[59] Y. Zhu, X. Shen, W. Pan, Network-based support vector machine for classification of microarray samples, BMC Bioinformatics 10 (1) (2009) S21.