IET Systems Biology

Research Article

IET Journals
The Institution of Engineering and Technology

Gene selection approach based on improved swarm intelligent optimisation algorithm for tumour classification

ISSN 1751-8849
Received on 29th April 2015
Revised on 1st December 2015
Accepted on 3rd December 2015
doi: 10.1049/iet-syb.2015.0064
www.ietdl.org

Cong Jin[1] ✉, *Shu-Wei Jin*[2]

[1]*School of Computer, Central China Normal University, Wuhan 430079, People's Republic of China*
[2]*Département de Physique, École Normale Supérieure, 24, rue Lhomond 75231 Paris Cedex 5, France*
✉ *E-mail: jincong@mail.ccnu.ed.cn*

Abstract: A number of different gene selection approaches based on gene expression profiles (GEP) have been developed for tumour classification. A gene selection approach selects the most informative genes from the whole gene space, which is an important process for tumour classification using GEP. This study presents an improved swarm intelligent optimisation algorithm to select genes for maintaining the diversity of the population. The most essential characteristic of the proposed approach is that it can automatically determine the number of the selected genes. On the basis of the gene selection, the authors construct a variety of the tumour classifiers, including the ensemble classifiers. Four gene datasets are used to evaluate the performance of the proposed approach. The experimental results confirm that the proposed classifiers for tumour classification are indeed effective.

1 Introduction

The medical experts [1, 2] agree that early diagnosis of tumour is of great benefit; however, it is very difficult for traditional tumour detection, such as X-ray imaging. In recent 10 years, in order to achieve early diagnosis of tumour, the gene expression profiles (GEP) have already attracted a large number of medical workers and computer scientists [3–8]. Recent studies demonstrated that the genes in GEP dataset could provide useful information for tumour classification. However, the classification based on GEP dataset is very different from previous classification problems in that the number of genes (typically tens of thousands) greatly exceeds the number of samples (typically a few hundred or less), resulting in the known problem of 'curse of dimensionality' and over-fitting of the training data. Therefore, to drastically reduce the dimensionality of tumour data, how to select important gene subsets from thousands of genes in GEP dataset is a very key step to address tumour classification.

Generally, tumour classification can be considered as a problem consisting of two tasks: gene selection and classification. Gene selection is the recognition of important genes from thousands of highly correlated GEP to capture the informative genes for tumour classification. Classification requires constructing a model, which predicts the class or category under given conditions. In the past few years, many classification algorithms [2, 3, 5–12] have been applied to tumour classification based on the gene expression data. For instance, rough set theory was used for cancer prediction [9], data clustering with optimisation was applied for cancer classification [10], and random forest was used for analysing acute leukaemia [12], etc.

Recently, various gene selection approaches have been proposed [9–12]. Most of them have been proven helpful for improving the classification accuracy (CA) of disease and providing useful information for biologists and medical experts. Typically, the gene selection approaches may be grouped into three teams: filter [1, 13], wrapper [14, 15] and hybrid techniques [16–19]. The filter technique relies on the general characteristics of data to evaluate and select gene subsets without involving classification algorithm. The wrapper technique first implements an optimising algorithm that adds or removes genes to produce various gene subsets, and then employs a classification algorithm to evaluate these gene subsets [14]. The hybrid technique attempts to take advantage of the filter and wrapper techniques by exploiting their complementary strengths [17]. However, these gene selection approaches deal with gene sequentially one by one [20], which does not seem appropriate for GEP dataset with the large number of genes.

Swarm intelligence optimisation is a series of relatively optimisation algorithms including ant colony optimisation [21, 22], particle swarm optimisation (PSO) [23, 24] and so on. It simulates the swarm behaviour of the social individuals, using the information interchange and cooperation between individuals to achieve optimisation. Compared with the other swarm intelligence algorithms, PSO is a simple algorithm with desirable properties including fewer control parameters, better convergence performance and so on. It is mainly used to solve some non-linear complex optimisation problems. In addition, in the implementation process of PSO, it does not need any assumptions of the gene expression data, only uses the characteristics of the data itself. PSO has attracted many researchers and has emerged as the most popular tool for intelligent optimal problems. Because of these advantages, we use PSO to select genes in the paper. The proposed gene selection approach belongs to the filter type, which differs from the existing work, and obtains selected genes from the optimal solution of PSO.

An ensemble classification [25, 26] is the process by which multiple classifiers are strategically generated and combined in order to solve a particular machine learning problem. It is primarily used to improve the classification or prediction performance of a model, or to reduce the likelihood of a poor or an unfortunate selection [27]. Recently, ensemble classification algorithm has been used in bioinformatics [28, 29].

In this paper, we propose an improving gene selection approach based on PSO. The main contributions of this work are the following:

(i) The proposed gene selection approach is solely based on given GEP data. The number of selected genes is decided by improving PSO (IPSO) automatically, which does not require predetermining.
(ii) The proposed IPSO algorithm can effectively maintain the diversity of the population, and the experiment results confirm that the most informative genes can be selected from the whole gene space.
(iii) The proposed ensemble classifier (EC) can obtain the highest CA in all compared classifiers, and the experiment results confirm

that the ensemble technology is effective for tumour classification and CA is stable and reliable.

The rest of this paper is organised as follows. Section 2 introduces the standard PSO, IPSO, and binary IPSO (BIPSO), which will be used in the gene selection research. The proposed gene selection approach is described in Section 3. The proposed EC is discussed in Section 4. Section 5 presents the experimental studies including the experimental datasets, parameters initialisation and experimental results. Finally, the conclusions are given in Section 6.

2 Standard and IPSO

2.1 Standard PSO

PSO is a population-based optimisation algorithm [30, 31]. It is initialised with a population of random potential solutions and the algorithm searches for satisfying performance. The potential solutions, called particles, are flow through the problem space. Each particle i has a position represented by a vector $\boldsymbol{X}_i$. A swarm of particles moves through a d-dimensional space, with the velocity of each particle represented by a vector $\boldsymbol{V}_i$. The velocity and position of a particle satisfy the following equations, respectively

$$\begin{aligned}\boldsymbol{V}_i(t+1) = \omega \cdot \boldsymbol{V}_i(t) + c_1 \text{rand}_1(P_{i,\text{best}}(t) - \boldsymbol{X}_i(t)) \\ + c_2 \text{rand}_2(P_{\text{global}}(t) - \boldsymbol{X}_i(t))\end{aligned} \tag{1}$$

$$\boldsymbol{X}_i(t+1) = \boldsymbol{X}_i(t) + \boldsymbol{V}_i(t+1) \tag{2}$$

where, t is the current iteration number, ω is inertia weight, c_1 and c_2 are positive constants, rand_1 and rand_2 are uniformly distributed random numbers in the interval [0, 1]. $P_{i,\text{best}}$ and P_{global} are the best previously visited position of the particle i and the best value of all individual particle position values, respectively. The initial velocities of the particles are probabilities limited to the interval [0, 1]. We let $\boldsymbol{X}_i(t) = (X_{i1}(t), X_{i2}(t), \ldots, X_{id}(t))$ and $\boldsymbol{V}_i(t) = (V_{i1}(t), V_{i2}(t), \ldots, V_{id}(t))$. Generally, ω, c_1 and c_2 are predefined by the user.

The fitness value of particle i, at iteration t, is calculated as follows

$$F(\boldsymbol{X}_i(t)) = \frac{1}{N}\sum_{k=1}^{N}\sum_{j=1}^{d}(X_{ij}^{(k)}(t) - \hat{X}_{ij}^{(k)}(t))^2 \tag{3}$$

where, for particle i, $X_{ij}^{(k)}$ is the jth output component of the kth sample, $\hat{X}_{ij}^{(k)}$ the jth actual component of the kth observation sample, and N total number of the samples. For the minimisation problem, the fitness value is better as long as F is smaller. The best position $P_{i,\text{best}}$ of the particle i is determined by following formula

$$P_{i,\text{best}}(t+1) = \begin{cases} \boldsymbol{X}_i(t), & \text{if} \quad F(\boldsymbol{X}_i(t)) < F(P_{i,\text{best}}(t)) \\ P_{i,\text{best}}(t), & \text{if} \quad F(\boldsymbol{X}_i(t)) \geq F(P_{i,\text{best}}(t)) \end{cases} \tag{4}$$

At update steps of standard PSO, the velocity of each particle is calculated according to (1) and the position is updated according to (2). When a particle finds a better position than the previously best position, this better position will be stored in the memory. PSO algorithm goes on until a satisfactory solution is found or the predefined number of iterations is met.

2.2 Improving PSO

We note that, during searching process, some of the particles are no longer moving after several iterations. The usual reason is that the best particle has fallen into a local minimum or the velocity of the particle has been reduced to a very small value, which means that the premature convergence has occurred because of the loss of population diversity.

To maintain the diversity of the population, in this paper, we try to use the average position $P_{i,\text{ave}}$ of several particles to generate the current position of each particle [32]. The particle velocity update equation is given as follows

$$\begin{aligned}\boldsymbol{V}_i(t+1) = \omega \cdot \boldsymbol{V}_i(t) + c_1 \text{rand}_1(P_{i,\text{ave}}(t) - \boldsymbol{X}_i(t)) \\ + c_2 \text{rand}_2(P_{\text{global}}(t) - \boldsymbol{X}_i(t))\end{aligned} \tag{5}$$

where $P_{i,\text{ave}} = (P_{i,1} + P_{i,2} + \cdots + P_{i,T})/T$ is the average value of the first T beat positions of the particles, and T is the number of the particles predefined by the user.

We also note that, if the weight ω is large, the particle will have difficulty converging to P_{global}, and vice versa. Usually, ω was assumed to be linearly decreasing during the iteration process, which shows that ω is the weight factor of variable velocity and can restrain the maximum velocity automatically for accelerating the convergence. However, if PSO finds the optimal solution with linear decreasing ω at the beginning, the search will jump out of the optimal solution because a larger ω value results in the lowering of the search capability. Therefore, in this paper, the value of ω is modified according to $\omega(t+1) = 1 - \lambda\omega^2(t)$, where λ is a parameter given by the user.

At the primary phase of iterations, if $\omega(t)$ is large, the whole space will be searched for finding the optimum solution. As the iteration increases, the search space will be smaller until the optimal solution is found. Hence, we can obtain an IPSO. The particle velocity update equation is as follows

$$\begin{aligned}\boldsymbol{V}_i(t+1) = \omega(t) \cdot \boldsymbol{V}_i(t) + c_1 \text{rand}_1(P_{i,\text{ave}}(t) - \boldsymbol{X}_i(t)) \\ + c_2 \text{rand}_2(P_{\text{global}}(t) - \boldsymbol{X}_i(t))\end{aligned} \tag{6}$$

3 Gene selection approach

Assume that $D = (S, G)$ is a GEP dataset with N samples and d genes, where $G = \{g_1, \ldots, g_d\}$ and $S = \{S_1, \ldots, S_N\}$ are the gene and the sample sets, respectively. ith sample is represented as $S_i = \{g_{i1}, g_{i2}, \ldots, g_{id}\}^{\text{T}}$. $C = \{c_1, \ldots, c_L\}$ refers to tumour type set.

When using IPSO for selecting genes, we need to use the BIPSO. Generally, the maximum and minimum velocity vectors, namely $\boldsymbol{V}_{\max}$ and $\boldsymbol{V}_{\min}$, will be restrained in order to control velocity. Wherever $\boldsymbol{V}_i$ exceeds $\boldsymbol{V}_{\max}$, its velocity is set to $\boldsymbol{V}_{\max}$, whereas when $\boldsymbol{V}_i$ is less than $\boldsymbol{V}_{\min}$, its velocity is set to $\boldsymbol{V}_{\min}$. Then, the expression level for each velocity was normalised to [0, 1]. The normalisation procedure is given as follows

$$V_i' = \frac{\boldsymbol{V}_i - \boldsymbol{V}_{\min}}{\boldsymbol{V}_{\max} - \boldsymbol{V}_{\min}} \tag{7}$$

For any random number, *rand*, in the interval [0, 1], if $\text{rand} < V'_i$, then $\boldsymbol{X}_i = 1$, else $\boldsymbol{X}_i = 0$. Hence, for the proposed gene selection approach, the position of particle i is represented by a binary (0 or 1) string, for example, $\boldsymbol{X}_i = \{X_{i1}, X_{i2}, \ldots, X_{id}\}$. 1 represents a selected gene and 0 a non-selected one. The velocity of particle i is represented by $\boldsymbol{V}_i = \{V_{i1}, V_{i2}, \ldots, V_{id}\}$. In BIPSO, once an optimal solution P_{global} is found, the selected genes can be achieved. The method is such that, if the ith position value of P_{global} is 1, then the corresponding gene is selected. In contrary, the gene is not selected if the ith position value of P_{global} is 0.

The detail gene selection algorithm is shown as follows: (see Fig. 1).

So far, from P_{global}, we can get a set G_1, and its elements are to be selected genes.

4 Ensemble tumour classifier

In this section, we discuss the tumour classification problem in which $\{(S_i, c_i), i = 1, 2, \ldots, N\}$, where $S_i \in R^I$ and $c_i \in C = \{c_1, \ldots, c_L\}$, each

IET Syst. Biol., 2016, Vol. 10, Iss. 3, pp. 107–115

Algorithm 1

Input A GEP dataset $D = (S, G)$ and tumor type set C. Initialize a population of the particles with random positions and velocities. Each particle is treated as a point in a d-dimensional space. Normalized gene set is still denoted by G.

Output A set G_1 with selected l genes, where $l < d$.

Do

Evaluate $P_{i,best}$ of each particle i according to equation (4).

Update $P_{i,best}$ if $P_{i,best}$ is better than P_{global}.

Determine P_{global}. Choose the particle with the best value $P_{i,best}$ of all.

For each particle:

Calculate particle's new velocity according to (6);

Let the new velocity value be binary.

Let the new position value be binary.

Calculate particle's new position according to (2).

While a sufficiently good P_{global} or a maximum number of iterations are not yet attained.

Fig. 1 *BIPSO-based gene selection algorithm*

tumour type being described by a set of l selected genes ($l < d$) and GEP samples i was represented as $S_i = \{g_{i1}, g_{i2}, \ldots, g_{il}\}^{\mathrm{T}}$.

4.1 Some common classifiers

There are many kinds of classifiers with different performances. Some classifiers are very effective for some, while not ideal for others. Therefore, using a single classifier for classification task is not a good strategy. If the classification results of several classifiers are integrated, the EC will have better performance [27, 28, 33]. In this paper, three classifiers are briefly introduced as follows.

Linear discriminant analysis (LDA, namely h_1) [34] is a conventional classification approach, which determines linear decision boundaries between t classes while taking into account between-class and within-class variances. If the error distributions for each class are the same (identical covariance matrices), LDA constructs the optimal linear decision boundary between the classes. LDA is a useful linear classifier.

k nearest neighbour (kNN, namely h_2) [35] is a type of instance-based learning, where the function is only approximated locally and all computation is deferred until classification. kNN is the simplest in all machine learning algorithms: an object is classified by a majority vote of its neighbours, with the object

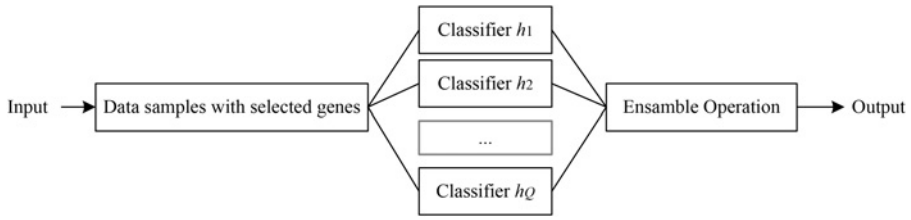

Fig. 2 *General ensemble framework of several classifiers*

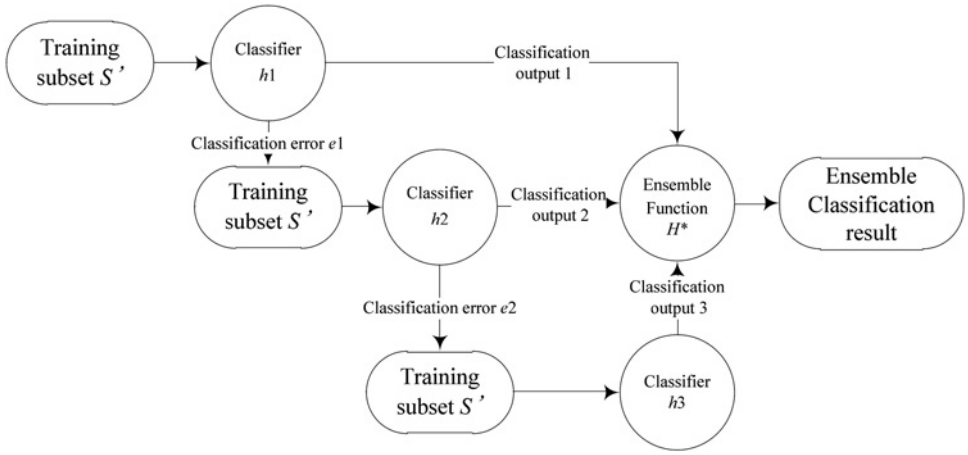

Fig. 3 *Boosting ensemble framework of several classifiers*

Algorithm 2

Input: A GEP dataset $D_1 = (S_1, G_1)$ with l selected genes and tumor type set C. Q is the number of classifiers.

Output: Boosting EC classifier.

Divide sample set S_1 into two subsets, *i.e.*, training set S' with N_1 samples and testing set S'' with N_2 samples according to a certain ratio, and $N = N_1 + N_2$. Let training set $S' = \{(S_1, c_1), (S_2, c_2), \ldots, (S_{N_1}, c_{N_1})\}$.

(1) Initialize $w_i^{(1)} = \frac{1}{N_1}$ for all $i = 1, 2, \ldots, N_1$.

(2) For $r = 1$ to Q

(3) Train classifier with respect to the weighted sample set $\{S', w^{(r)}\}$ and obtain hypothesis:

$$h_r: S' \rightarrow \{+1, -1\}, \text{ i.e., } h_r = M(S', w^{(r)})$$

(4) Calculate the weighted training error ε_r of h_r:

$$\varepsilon_r = \sum_{i=1}^{N_1} w_i^{(r)} I(c_i \neq h_r(S_i))$$

Where, I is an index function such that $I(c_i \neq h_r(S_i)) = 1$, $I(c_i = h_r(S_i)) = 0$.

(5) Let $\alpha_r = \frac{1}{2}\log\frac{1-\varepsilon_r}{\varepsilon_r}$.

(6) Update weights $w_i^{(r+1)} = \frac{w_i^{(r)}}{Z_r}\exp\{-\alpha_r c_i h_r(S_i)\}$. Where, Z_r is a normalization constant, such that $\sum_{i=1}^{N_1} w_i^{(r+1)} = 1$.

(7) Break if $\varepsilon_r = 0$ or $\varepsilon_r \geq \frac{1}{2}$ and set $Q = r - 1$.

(8) EC result $H^*(S') = \sum_{r=1}^{Q} \frac{\alpha_r}{\sum_{j=1}^{Q}\alpha_j} h_r(S')$.

(9) Using the testing set S'', we compute CA of H^* and obtain the performance evaluation using x-fold cross validation.

Fig. 4 *Boosting EC algorithm*

being assigned to the class most common among its k nearest neighbours (k is a positive integer, typically small).

Support vector machine (SVM, namely h_3) [36] is supervised learning model with associated learning algorithms that analyse data and recognise patterns, used for classification and regression analysis. The basic SVM takes a set of input data and predicts, for each given input, which of two possible classes form the output, making it a non-probabilistic binary linear classifier.

4.2 Ensemble classifier

As an effective approach, the ensemble operation is available for improving classification performance of biomedical data, and therefore gaining more and more attention in biomedicine and computer communities. The simple ensemble operation is max, average, median, or backpropagation neural network (BPNN), etc. Fig. 2 shows the framework of general EC.

In general EC, in order to achieve the ensemble classification, let $h_j(S_i)$ be the class result of classifier h_j for a sample S_i. Consider an EC H* by consisting of Q single classifiers, namely $j = \{1, 2, \ldots, Q\}$. For each sample S_i, its output is

$$H^*(S_i) = \arg \underset{j=1,2,\ldots,Q}{\text{ensamble operation}}\, h_j(S_i) \tag{8}$$

In addition to these simple ensemble operations above, boosting is also an effective ensemble operation [37], and its framework is described in Fig. 3.

In this paper, x-fold cross validation [38] was used to evaluate the classification performance of the classifiers. Algorithm 2 (see Fig. 4) describes the implementation process of a boosting EC.

4.3 Measure the diversity of population

Hamming distance (HD) is used in this paper since HD can describe the difference between the particles. When HD is larger, the difference between particles is larger, whereas the difference between particles is smaller. The difference between the particles is calculated as $\text{Ham} = \sum_{X_i, X_j \in X} \left|\text{Sgn}(X_i - X_j)\right|$, where, X is a binary population, X_i and X_j are the positions of particle i and j represented using binary, respectively, and X_{ik} is the kth bit value of $\boldsymbol{X}_i$. The binary population X can be expressed as

$$X = \begin{pmatrix} x_{11} & x_{12} & \ldots & x_{1l} \\ x_{21} & x_{22} & \ldots & x_{2l} \\ & \ldots & \ldots & \\ x_{n1} & x_{n2} & \ldots & x_{nl} \end{pmatrix}$$

where n is the population size. Let p_{ij}^k be the difference of X_i and X_j in

the kth bit, namely $p_{ij}^k = \begin{cases} 1, & \text{if } X_{ik} \neq X_{jk} \\ 0, & \text{other} \end{cases}$. Therefore, the diversity of the population X in the kth bit is $p^k = \sum_{i=1}^{n} \sum_{j=1}^{l} p_{ij}^k$, and the diversity measure of the population X is calculated as follows

$$\text{Div}(X) = \frac{\sum_{k=1}^{l} p^k}{n(n-1)l}$$

5 Experiments and discussions

5.1 Datasets

To validate the performances of the proposed gene selection approach and EC, four publicly available gene microarray datasets were used and summarised in Table 1. The first two benchmark datasets (Leukemia [39] and Colon tumour [40]) are two-class, while the other two benchmark datasets (SRBCT [41] and Lymphoma [42]) are multi-class. We divide each benchmark dataset into two subsets: the training and test sets, whose descriptions are also summarised in Table 1.

The expression levels for each gene of all samples were normalised to [0, 1]. The normalisation procedure is given in the following equation

$$G = \frac{\text{Value} - \text{Value}_{\min}}{\text{Value}_{\max} - \text{Value}_{\min}} \tag{9}$$

where $\text{Value}_{\max}$ is the maximum original value and $\text{Value}_{\min}$ the minimum original value.

5.2 Parameters initialisation

In the experiments, the training set is used to build the classifiers and the testing set is used to evaluate the classification capability of the classifiers. For comparing the performance of the proposed approach, three classifiers (namely LDA, kNN and SVM) and their ensemble are practiced in the four given GEP datasets, respectively. The reason of choosing them is that they are relatively fast and reliable. The three classifiers will be initialised as follows:

- LDA: Using proportional probabilities.
- kNN: The number of observations (k) in the set of closest neighbour was set to 5. Euclidean distance is used to measure the similarity of the samples.
- SVM: The radial basis function (RBF) kernel is used in the experiments. The regularisation parameter was set to 1, the bandwidth of the kernel function was set to 0.5.
- BIPSO: The number of the particles n is set to 100, $c_1 = c_2 = 2$, and the max iteration is set to 500. In IPSO, T is set to 4. Let λ be set to 1.99 and $\omega(0) = 0.256$.

In addition, for the ensemble operation, BPNN is also initialised.

- BPNN: The initial connection weights for BPNN were randomly chosen between −1.0 and 1.0. The learning rate and momentum term for training BPNN were chosen as 0.1–0.15 and 0.8–0.9, respectively. The number of training epochs of BPNN was chosen as 50. The training error threshold was set to 0.002. Of course, these parameters were not meant to be optimal. The construction of BPNN was one input layer, one hidden layer and one output layer. There are 3 and 1 neurons in input and output layers, respectively. The most common winner-takes-all method was used as the output of BPNN. The hidden and output neuron functions were defined by the logistic sigmoid function $f(x) = 1/(1 + \exp(-x))$. In the hidden layer, the number of the neurons is determined by $(M_1 + M_2)^{1/2} + \vartheta$, where ϑ is a constant of 1–10, M_1 and M_2 are the number of the input and output neurons, respectively.

Table 1 Description of four benchmark datasets

	Dataset	# samples	# genes	# of classes	# training set	# test set
1	Leukaemia	72	7129	2 (ALL/AML)	38	34
2	Colon tumour	62	2000	2 (tumour/ normal)	40	22
3	SRBCT	83	2308	4 (EWS/BL/ NB/RMS)	63	20
4	Lymphoma	62	4026	3 (B-CLL/FL/ DLBCL)	42	20

5.3 Classification accuracy

In this study, CA in test set S'' is measured as follows:

$$\text{Accuracy}(S'') = \frac{\sum_{i=1}^{|S''|} \text{Assess}(S_i)}{|S''|}, \quad S_i \in S''$$

$$\text{Assess}(S_i) = \begin{cases} 1, & \text{if classifier } h(S_i) = \text{a.c} \\ 0, & \text{otherwise} \end{cases}$$

where, a.c is the actual class of the sample S_i, and *classifier* $h(S_i)$ is return class of S_i by the classifier h.

5.4 Results and discussions

To observe the effect of IPSO for maintaining population diversity, we calculate the DIV values of PSO and IPSO. We experiment 20 times on each dataset, and observe the average DIV value of all experimental results. The experimental results of 200 iterations are shown in Fig. 5.

From Fig. 5, the average DIV value of PSO converges quickly to 0 after less iteration, which shows that the population diversity is very low after iterating about 20 times. For IPSO, the population diversity is well maintained with a non-zero value. The average DIV value of IPSO is higher than PSO's, which shows that IPSO is superior than PSO for maintaining population diversity.

To evaluate the performance of the proposed gene selection approach for two-class, SVM is used as a classifier. We compare the proposed approach to two existing approaches, namely gene ranking [5] and shrinkage [43], on the given first two benchmark datasets and CAs are summarised in Table 2.

Results from Leukaemia: When randomly dividing dataset into the training and test sets accordance to given number, 0.9302 and 0.8875 are the best CA values of gene ranking and shrinkage after running 20 times, respectively, and 0.9824 is the best CA value of BIPSO. Therefore, the proposed approach is better than the other two approaches.

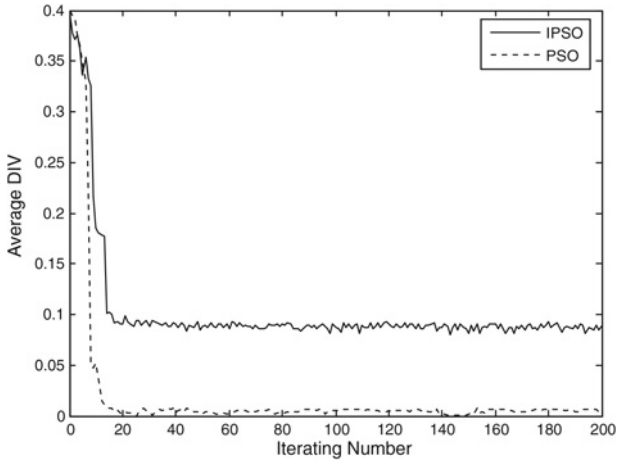

Fig. 5 *DIV values of IPSO and PSO*

Table 2 Comparison of CA results on Leukaemia and Colon tumour datasets

Datasets	Method of gene selection	Average # of selected genes	CA (%)
Leukaemia	Gene ranking	51.62	0.9302
	Shrinkage	45.85	0.8875
	BIPSO	40.31	0.9824
Colon tumour	Gene ranking	49.84	0.9480
	Shrinkage	43.69	0.8874
	BIPSO	25.27	0.9902

Table 3 Minimum number, the percentages for selected genes and their CA values

Gene selection approach		Minimum # of selected genes	The percentage of selected genes, %	CA (%)
Leukaemia	Gene ranking	42	0.5611	0.7416
	Shrinkage	36	0.3928	0.6032
	BIPSO	8	0.1122	0.9045
Colon tumour	Gene ranking	31	1.5500	0.7127
	Shrinkage	24	1.2000	0.6329
	BIPSO	6	0.3000	0.9218

Results from Colon tumour: Similar to the Leukemia dataset, 0.9480 and 0.8874 are the best CA values of gene ranking and shrinkage after running 20 times, respectively, and 0.9902 is the best value CA of BIPSO. Therefore, the proposed approach is better than the other two approaches.

In addition, when the minimum gene number is selected, we compare their CA results using three approaches, respectively. The percentage of the selected minimum gene number is listed in Table 3.

Table 3 shows that, when the minimum gene number is selected, the proposed BIPSO approach still can get the best CA value in all three approaches, which shows that the most informative genes may be selected from the whole gene space by BIPSO.

For multi-class, according to above approaches, we compare the proposed approach to other existing approaches, namely EPSO [44], mRMR-PSO [45] and mRMR-GA [46], on given the last two benchmark datasets and the CA values are summarised in Table 4.

Results from SRBCT: When randomly dividing dataset into the training and test sets according to the given number, 0.9964 and 0.9397 are the best CA values of EPSO and mRMR-PSO after running 20 times, respectively, and 1.0000 is the best CA value of BIPSO. Therefore, the proposed approach is better than the other two approaches.

Results from lymphoma: Similar to the SRBCT dataset, 0.9393 and 0.9696 are the best CA values of mRMR-GA and mRMR-PSO after running 20 times, respectively, and 0.9998 is the best CA value of BIPSO. Therefore, the proposed approach is better than the other two approaches.

In addition, when the minimum gene number is selected, we compare their CA results using three approaches, respectively. The percentage of the selected minimum gene number is listed in Table 5.

Table 5 shows that, when the minimum gene number is selected, the proposed approach BIPSO still can get the best CA in all three approaches, which shows that the most informative genes may be selected from the whole gene space by BIPSO.

Table 4 Comparison of CA results on SRBCT and Lymphoma datasets

Datasets	Method of gene selection	Average # of selected genes	CA (%)
SRBCT	EPSO	14.90	0.9964
	mRMR-PSO	68.00	0.9397
	BIPSO	8.74	1.0000
Lymphoma	mRMR-GA	43.00	0.9393
	mRMR-PSO	82.00	0.9696
	BIPSO	7.25	0.9998

Table 5 Minimum number, the percentages for selected genes and their CA values

Gene selection approach		Minimum # of selected genes	The percentage of selected genes (%)	CA (%)
SRBCT	EPSO	12	0.5119	0.9847
	mRMR-PSO	68	2.9463	0.9397
	BIPSO	7	0.3033	1.0000
Lymphoma	mRMR-GA	43	1.0681	0.9393
	mRMR-PSO	82	2.0368	0.9696
	BIPSO	5	0.1242	0.9994

Table 6 Average CA values for 5-fold cross-validation

Methods	Leukaemia	Colon tumour	SRBCT	Lymphoma
BIPSOSVM	0.9358	0.9495	0.9976	0.9784
BIPSOLDA	0.9195	0.9287	0.9771	0.9653
BIPSO*k*NN	0.9102	0.9309	0.9667	0.9604
EC1	0.9389	0.9546	0.9965	0.9758
EC2	0.9278	0.9342	0.9819	0.9682
EC3	0.9337	0.9479	0.9924	0.9749
EC4	0.9612	0.9685	0.9986	0.9825
EC5	0.9890	0.9779	1.0000	1.0000

As shown in Tables 2–5, for the two-class and multi-class problems, CA values of all BIPSO on the four benchmark datasets are the highest. We also note that the number of the selected genes by BIPSO on the four benchmark datasets is the least, which shows that BIPSO is very effective for the gene selection. We still note that, the compared gene selection approaches, namely EPSO and mRMR-PSO, are based on PSO, which shows that the improving operation of the proposed BIPSO is more suitable than EPSO and mRMR-PSO for the gene selection.

Using selected genes, we compare the performance of the differential single classifiers. In this study, we designed three single classifiers, which are defined as follows:

BIPSO + SVM = BIPSOSVM
BIPSO + LDA = BIPSOLDA
BIPSO + *k*NN = BIPSO*k*NN

Table 7 Comparison results of different existing classifiers on the four benchmark datasets

Datasets	Classification algorithms	Source	CA
Leukaemia	Representative gene vectors + ANN ensemble	[47]	0.9590
	SVM ensemble	[48]	0.9861
	Classical rough set + SVM-RBF ensemble	[49]	0.9722
	Proposed EC5		0.9890
Colon tumour	Representative gene vectors + ANN ensemble	[47]	0.8790
	SVM ensemble	[48]	0.9040
	Classical rough set + SVM-RBF ensemble	[49]	0.8710
	Proposed EC5		0.9779
SRBCT	The significance of gene to tumour + ANN ensemble	[50]	1.0000
	Evolutionary algorithm + *k*NN ensemble	[51]	1.0000
	Wavelet packet transforms + Neighbourhood rough sets + SVM	[52]	1.0000
	Proposed EC5		1.0000
Lymphoma	mRMR-PSO + SVM classifier	[45]	0.9696
	mRMR-GA + SVM classifier	[46]	0.9500
	GA with dynamic parameter (GADP) + SVM classifier	[53]	1.0000
	Proposed EC5		1.0000

IET Syst. Biol., 2016, Vol. 10, Iss. 3, pp. 107–115

Table 8 Times of CA value using differential classifiers

Approaches	CA									
	<0.80		≤0.80–<0.85		≤0.85–<0.90		≤0.90–<0.95		≥0.95	
BIPSOSVM	4	1.33%	24	8%	68	22.67%	80	26.66%	124	41.34%
BIPSOLDA	16	5.33%	35	11.67%	84	28%	68	22.66%	97	32.34%
BIPSO*k*NN	14	4.67%	21	7%	63	21%	74	24.66%	128	42.67%
EC1	7	2.33%	18	6%	51	17%	58	19.33%	166	55.34%
EC2	10	3.33%	25	8.33%	52	17.33%	64	21.34%	149	49.67%
EC3	8	2.67%	22	7.33%	58	19.33%	49	16.34%	163	54.33%
EC4	3	1%	9	3%	41	13.66%	42	14%	205	68.34%
EC5	0	0%	3	1%	23	7.67%	38	12.67%	236	78.66%

Their EC is defined as follows

BIPSO + Ensemble of (LDA, *k*NN, SVM) according to max = EC1
BIPSO + Ensemble of (LDA, *k*NN, SVM) according to average = EC2
BIPSO + Ensemble of (LDA, *k*NN, SVM) according to median = EC3
BIPSO + Ensemble of (LDA, *k*NN, SVM) according to BPNN = EC4
BIPSO + Ensemble of (LDA, *k*NN, SVM) according to boosting = EC5

After running 20 times, the average CA values of the test results are summarised in Table 6.

From Table 6, we find that, for any benchmark dataset, the average CA value of each single classifier, except BIPSOSVM, is lower than EC's. In other words, the five ECs achieve overall better classification performance, which shows that, as a single classifier, SVM has the best classification performance in the three single classifiers. For the ECs, EC5 achieves the highest average CA value on all four benchmark datasets, and therefore EC5 is an ideal EC. We also note that, EC4 does not obtain the highest average CA value, but its average CA value is not only much higher than any single classifier, but also higher than EC1, EC2 and EC3, therefore EC4 also is an ideal EC.

We also compare classifier EC5 to other existing classifiers for tumour classification, the classification results are listed in Table 7.

Table 7 shows that, on the SRBCT dataset, the proposed EC5 is as good as other approaches in CA. On the other three datasets, the proposed EC5 achieves the best CA in all approaches, which confirms that the proposed EC5 is an ideal classifier for tumour classification.

Because the division of training and test sets is a random choice, the CA value of a classifier at each run is not necessarily the same. To evaluate the classification ability and reliability of the classifiers derived by selected genes set by BIPSO, the total run number is 300 times, and it runs 100 times on each dataset. Table 8 and Figs. 6–13 show the distributions of CA values for testing set over 300 times by BIPSOSVM, BIPSOLDA, BIPSO*k*NN, EC1, EC2, EC3, EC4 and EC5, respectively.

From Table 8 and Figs. 6–13, we note that eight classifiers can achieve satisfactory CA values, and their classification results are stable and reliable. In the experiments, the proportions of over 90% CA values for eight classifiers are 68%, 55%, 67.33%, 74.67%, 71%, 70.67%, 82.33%, and 91.33%, respectively. In addition, from the whole, the classification result of any EC is

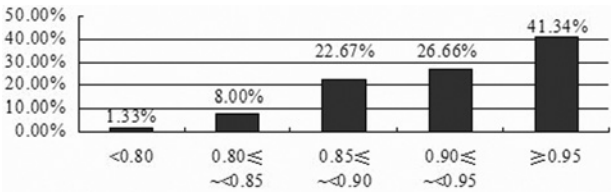

Fig. 6 *Distribution of BIPSOSVM*

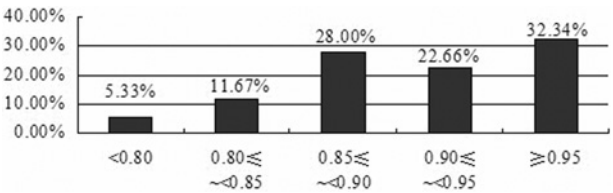

Fig. 7 *Distribution of BIPSOLDA*

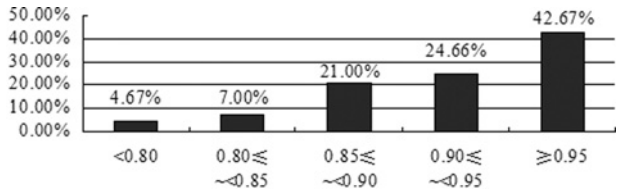

Fig. 8 *Distribution of BIPSOkNN*

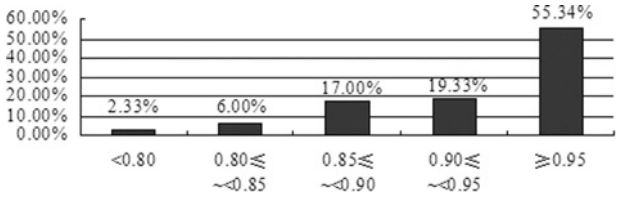

Fig. 9 *Distribution of EC1*

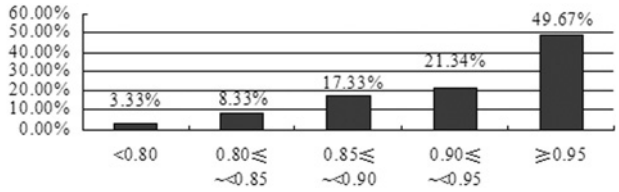

Fig. 10 *Distribution of EC2*

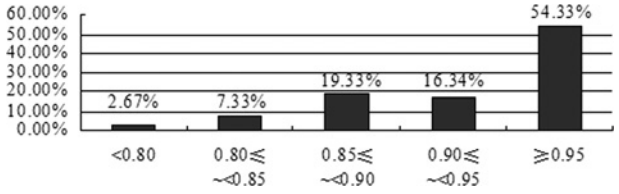

Fig. 11 *Distribution of EC3*

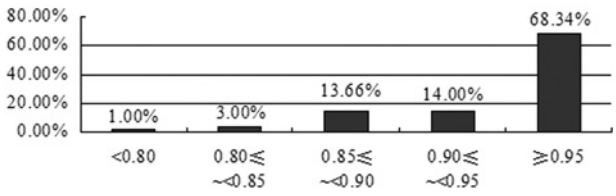

Fig. 12 *Distribution of EC4*

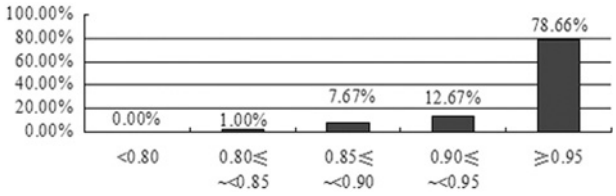

Fig. 13 *Distribution of EC5*

better than all single classifiers. In the ECs, the classification result of EC5 is better than the other four ECs.

6 Conclusions

In this paper, we proposed a gene selection approach for tumour classification. To show the performance of the proposed approach, we used the four tumour datasets in the experiments. The experimental results show that the proposed approach has better performance than existing approaches in many aspects. Even for different classifiers, the proposed approach can still achieve the desired classification performance. The proposed gene selection approach has the following advantages:

(i) The proposed gene selection approach is solely based on given datasets. The number of the selected genes is decided automatically by BIPSO.
(ii) The advantage of the proposed IPSO algorithm is that it can effectively maintain the diversity of the population, and the experiment results confirm that the performance of the tumour classification algorithm using selected genes is satisfactory.
(iii) The ensemble classifier EC5 can obtain the highest CA value in all compared classifiers, which shows that the ensemble technology is effective for tumour classification and CA value is stable and reliable.

7 References

1 Lazar, C., Taminau, J., Meganck, S., *et al.*: 'A survey on filter techniques for feature selection in gene expression microarray analysis', *IEEE/ACM Trans. Comput. Biol. Bioinf.*, 2012, **9**, (4), pp. 1106–1119
2 Reboiro, J.M., Arrais, J.P., Oliveira, J.L., *et al.*: 'Gene committee: a web-based tool for extensively testing the discriminatory power of biologically relevant gene sets in microarray data classification', *BMC Bioinf.*, 2014, **15**, (1), p. 31
3 Chandra, B., Babu, K.V.N.: 'Classification of gene expression data using spiking wavelet radial basis neural network', *Expert Syst. Appl.*, 2014, **41**, (4), pp. 1326–1330
4 Verhaegh, W., van Ooijen, H., Inda, M.A., *et al.*: 'Selection of personalized patient therapy through the use of knowledge-based computational models that identify tumor-driving signal transduction pathways', *Cancer Res.*, 2014, **74**, (11), pp. 2936–2945
5 Golub, T., Slonim, D., Tamayo, P., *et al.*: 'Molecular classification of cancer: class discovery and class prediction by gene expression monitoring', *Science*, 1999, **286**, pp. 531–537
6 Thanh, N., Abbas, K., Douglas, C., *et al.*: 'Hidden Markov models for cancer classification using gene expression profiles', *Inf. Sci.*, 2015, **316**, pp. 293–307
7 Li, B.Q., Huang, T., Liu, L., *et al.*: 'Identification of colorectal cancer related genes with mRMR and shortest path in protein-protein interaction network', *PloS One*, 2012, **7**, (4), p. e33393
8 Ashwinikumar, K., Naveen, K.B.S.C., Vadlamani, R., *et al.*: 'Colon cancer prediction with genetics profiles using evolutionary techniques', *Expert Syst. Appl.*, 2011, **38**, (3), pp. 2752–2757
9 Maji, P., Paul, S.: 'Rough set based maximum relevance-maximum significance criterion and gene selection from microarray data', *Int. J. Approx. Reason.*, 2011, **52**, (3), pp. 408–426
10 Elyasigomari, V., Mirjafari, M.S., Screen, H.R.C., *et al.*: 'Cancer classification using a novel gene selection approach by means of shuffling based on data clustering with optimization', *Appl. Soft Comput.*, 2015, **35**, pp. 43–51
11 Cao, J., Zhang, L., Wang, B.J., *et al.*: 'A fast gene selection method for multi-cancer classification using multiple support vector data description', *J. Biomed. Inf.*, 2015, **53**, pp. 381–389
12 Díaz-Uriarte, R., De Andres, S.A.: 'Gene selection and classification of microarray data using random forest', *BMC Bioinf.*, 2006, **7**, (1), p. 3
13 Santana, L.E.A.S., Canuto, A.M.P.: 'Filter-based optimization techniques for selection of feature subsets in ensemble systems', *Expert Syst. Appl.*, 2014, **41**, (4), pp. 1622–1631
14 Soufan, O., Kleftogiannis, D., Kalnis, P.: 'DWFS: a wrapper feature selection tool based on a parallel genetic algorithm', *PloS One*, 2015, **10**, (2), p. e0117988
15 Bermejo, P., de la Ossa, L., Gámez, J.A.: 'Fast wrapper feature subset selection in high-dimensional datasets by means of filter re-ranking', *Knowl. Based Syst.*, 2012, **25**, (1), pp. 35–44
16 Hsu, H.H., Hsieh, C.W., Lu, M.D.: 'Hybrid feature selection by combining filters and wrappers', *Expert Syst. Appl.*, 2011, **38**, (7), pp. 8144–8150
17 Hu, Z., Bao, Y., Xiong, T., *et al.*: 'Hybrid filter-wrapper feature selection for short-term load forecasting', *Eng. Appl. Artif. Intell.*, 2015, **40**, pp. 17–27
18 Bonilla-Huerta, E., Duval, B., Hernández, J.C.H., *et al.*: 'Hybrid filter-wrapper with a specialized random multi-parent crossover operator for gene selection and classification problems'. Bio-Inspired Computing and Applications, Springer, Berlin, Heidelberg, 2012, pp. 453–461
19 Sivagaminathan, R.K., Ramakrishnan, S.: 'A hybrid approach for feature subset selection using neural networks and ant colony optimization', *Expert Syst. Appl.*, 2007, **33**, (1), pp. 49–60
20 Jin, C., Jin, S.W., Qin, L.N.: 'Attribute selection method based on a hybrid BPNN and PSO algorithm', *Appl. Soft Comput.*, 2012, **12**, (8), pp. 2147–2155
21 Khan, S., Baig, A.R., Shahzad, W.: 'A novel ant colony optimization based single path hierarchical classification algorithm for predicting gene ontology', *Appl. Soft Comput.*, 2014, **16**, pp. 34–49
22 López-Ibánez, M., Stützle, T.: 'The automatic design of multiobjective ant colony optimization algorithms', *IEEE Trans. Evol. Comput.*, 2012, **16**, (6), pp. 861–875
23 Gandomi, A.H., Yun, G.J., Yang, X.S., *et al.*: 'Chaos-enhanced accelerated particle swarm optimization', *Commun. Nonlinear Sci. Numer. Simul.*, 2013, **18**, (2), pp. 327–340
24 Goksal, F.P., Karaoglan, I., Altiparmak, F.: 'A hybrid discrete particle swarm optimization for vehicle routing problem with simultaneous pickup and delivery', *Comput. Ind. Eng.*, 2013, **65**, (1), pp. 39–53
25 Chen, T., Hong, Z., Deng, F.A., *et al.*: 'A novel selective ensemble classification of microarray data based on teaching-learning-based optimization', *Int. J. Multimedia Ubiquit. Eng.*, 2015, **10**, (6), pp. 203–218
26 Rathore, S., Hussain, M., Iftikhar, M.A., *et al.*: 'Ensemble classification of colon biopsy images based on information rich hybrid features', *Comput. Biol. Med.*, 2014, **47**, pp. 76–92
27 Fraz, M.M., Remagnino, P., Hoppe, A., *et al.*: 'An ensemble classification-based approach applied to retinal blood vessel segmentation', *IEEE Trans. Biomed. Eng.*, 2012, **59**, (9), pp. 2538–2548
28 Varshney, K.R., Prenger, R.J., Marlatt, T.L., *et al.*: 'Practical ensemble classification error bounds for different operating points', *IEEE Trans. Knowl. Data Eng.*, 2013, **25**, (11), pp. 2590–2601
29 Yu, G., Domeniconi, C., Rangwala, H., *et al.*: 'Transductive multi-label ensemble classification for protein function prediction'. Proc. 18th ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining, KDD'12, ACM, New York, NY, August 2012, pp. 1077–1085
30 Kennedy, J., Eberhart, R.C., Shi, Y.: 'Swarm Intell' (Morgan Kaufmann, San Francisco, 2001)
31 Chen, K.H., Wang, K.J., Tsai, M.L., *et al.*: 'Gene selection for cancer identification: a decision tree model empowered by particle swarm optimization algorithm', *BMC Bioinf.*, 2014, **15**, (1), p. 49
32 Zhang, Y., Wang, S., Ji, G., *et al.*: 'A comprehensive survey on particle swarm optimization algorithm and its applications', *Math. Probl. Eng.*, 2015, **501**, p. 931256
33 Fraz, M.M., Rudnicka, A.R., Owen, C.G., *et al.*: 'Delineation of blood vessels in pediatric retinal images using decision trees-based ensemble classification', *Int. J. Comput. Ass. Rad. Surgery*, 2014, **9**, (5), pp. 795–811
34 Estoup, A., Lombaert, E., Marin, J.M., *et al.*: 'Estimation of demo-genetic model probabilities with approximate Bayesian computation using linear discriminant analysis on summary statistics', *Mol. Ecol. Resour.*, 2012, **12**, (5), pp. 846–855
35 Muja, M., Lowe, D.G.: 'Scalable nearest neighbor algorithms for high dimensional data', *IEEE Trans. Pattern Anal. Mach. Intell.*, 2014, **36**, (11), pp. 2227–2240
36 Chen, H.L., Yang, B., Liu, J., *et al.*: 'A support vector machine classifier with rough set-based feature selection for breast cancer diagnosis', *Expert Syst. Appl.*, 2011, **38**, (7), pp. 9014–9022
37 Ho, M., Leung, K.N.: 'Dynamic bias-current boosting technique for ultralow-power low-dropout regulator in biomedical applications', *IEEE Trans. Circuits Syst. II: Express Briefs*, 2011, **58**, (3), pp. 174–178
38 Kärkkäinen, T.: 'On cross-validation for MLP model evaluation'. Structural, Syntactic, and Statistical Pattern Recognition, Springer, Berlin, Heidelberg, 2014, pp. 291–300
39 'Leukemia', http://www.broadinstitute.org/cgi-bin/cancer/publications/pub_paper.cgi?mode=view&paper_id=43
40 'Colon tumor', http://genomics-pubs.princeton.edu/oncology/
41 'SRBCT', http://research.nhgri.nih.gov/microarray/Supplement/

42 'Lymphoma', http://genome-www.stanford.edu/lymphoma
43 Zou, H., Hastie, T.: 'Regularization and variable selection via the elastic net', *J. Roy. Statist. Soc., Ser. B: Statist. Methodol.*, 2005, **67**, (2), pp. 301–3320
44 Mohamad, M.S., Omatu, S., Deris, S., *et al.*: 'An enhancement of binary particle swarm optimization for gene selection in classifying cancer classes', *Algorithms Mol. Biol.*, 2013, **8**, (1), p. 15
45 Alshamlan, H.M., Badr, G.H., Alohali, Y.A., *et al.*: 'Genetic bee colony (GBC) algorithm: a new gene selection method for microarray cancer classification', *Comput. Biol. Chem.*, 2015, **56**, pp. 49–60
46 Amine, A., Aboutajdine, D.: 'A new gene selection approach based on minimum redundancy-maximum relevance (MRMR) and genetic algorithm (GA)'. Proc. Int. Conf. Computer Syst. Appl. (AICCSA2009), IEEE, Rabat, Morocco, May 2009, pp. 69–75
47 Cho, S.B., Won, H.H.: 'Cancer classification using ensemble of neural networks with multiple significant gene subsets', *Appl. Intell.*, 2007, **26**, (3), pp. 243–250
48 Peng, Y.H.: 'A novel ensemble machine learning for robust microarray data classification', *Comput. Biol. Med.*, 2006, **36**, (6), pp. 553–537
49 Wang, S.L., Chen, H.W., Li, F.R., *et al.*: 'Gene selection with rough sets for the molecular diagnosing of tumor based on support vector machines'. Int. Computer Symp., Taiwan, 2006, pp. 1368–1373
50 Khan, J., Wei, J.S., Ringner, M., *et al.*: 'Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks', *Nature Med.*, 2001, **7**, pp. 673–679
51 Deutsch, J.: 'Evolutionary algorithm for fining optimal gene sets in microarray prediction', *Bioinf.*, 2003, **19**, (1), pp. 45–52
52 Zhang, S.W., Huang, D.S., Wang, S.L.: 'A method of tumor classification based on wavelet packet transforms and neighborhood rough set', *Comput. Biol. Med.*, 2010, **40**, (4), pp. 430–437
53 Lee, C.P., Leu, Y.: 'A novel hybrid feature selection method for microarray data analysis', *Appl. Soft Comput.*, 2011, **11**, (1), pp. 208–213