A hybrid gene selection algorithm for microarray cancer classification using genetic algorithm and learning automata

Article · October 2017

DOI: 10.1016/j.imu.2017.10.004

CITATIONS
0

READS
108

4 authors:

Habib MotieGhader
University of Tehran
14 PUBLICATIONS **22** CITATIONS
SEE PROFILE

Ali Najafi
BUMS
91 PUBLICATIONS **244** CITATIONS
SEE PROFILE

Balal Sadeghi
Shahid Bahonar University of Kerman
31 PUBLICATIONS **50** CITATIONS
SEE PROFILE

Ali Masoudi-Nejad
http://LBB.ut.ac.ir
134 PUBLICATIONS **1,300** CITATIONS
SEE PROFILE

Informatics in Medicine Unlocked 9 (2017) 246–254

Contents lists available at ScienceDirect

Informatics in Medicine Unlocked

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

A hybrid gene selection algorithm for microarray cancer classification using genetic algorithm and learning automata

Habib Motieghader [a], Ali Najafi [b], Balal Sadeghi [c], Ali Masoudi-Nejad [a,*]

[a] Laboratory of Systems Biology and Bioinformatics (LBB), Institute of Biochemistry and Biophysics, University of Tehran, Tehran, Iran
[b] Molecular Biology Research Center, Baqiyatallah University of Medical Sciences, Tehran, Iran
[c] Food Hygiene and Public Health Dept., Faculty of Veterinary Medicine, Shahid Bahonar University of Kerman, Kerman, Iran

ARTICLE INFO

Keywords:
Biomarker
Cancer classification
Gene selection
Genetic algorithm
Learning automata

ABSTRACT

Cancer classification is an important problem in cancer diagnosis and treatment. One of the most effective methods in cancer classification is gene selection. However, selecting a subset of genes which increases the classification accuracy is an NP-Hard problem. A variety of algorithms were proposed for gene selection in cancer classification in previous studies. In this study, a hybrid meta-heuristic algorithm, which is an integration of Genetic Algorithm and Learning Automata (GALA), is proposed for this purpose. The time complexity of GALA is $O(G.m.n^3)$ and it has acceptable accuracy and performance on some well-known cancer datasets. To evaluate the performance of GALA, six different cancer datasets including Colon, ALL_AML, SRBCT, MLL, Tumors_9 and Tumors_11 were selected. Based on the evaluation process, the GALA algorithm provided remarkable results on each dataset compared to some recently proposed algorithms.

1. Introduction

Over the past decade, high-throughput technologies such as microarray and Next-Generation Sequencing (NGS) have been used in research centers and clinics for deep understanding of molecular mechanisms and effective treatment of complex diseases. It is a well-established fact that gene expression dysregulation is one of the key features of cancerous cells. Hence, finding cancer related genes in microarray gene expression data using gene selection algorithms is a vital step in cancer systems biology [1]. Microarray data contains some disease samples, each of which includes thousands of different genes. In microarray datasets, the number of samples is fewer than that of genes, so the results of the present methods are not precise enough [1]. Hence, this property of microarray data matrix poses a challenging problem.

Optimal selection of genes from all available genes is known as an NP-hard problem [2]. In order to select the most important genes from a pool of genes, different gene selection methods can be used [3–5]. Gene selection methods can be classified into two categories: filter selection methods and wrapper selection methods [6]. Filter selection methods are independent of classifying algorithms and their selection is based on inner features [7]. On the other hand, Wrapper selection methods evaluate learning methods on the selected features [7]. Most supervised learning algorithms such as Bayesian networks, support vector machines (SVMs), artificial neural networks (ANN), Decision Tree (DT) and so on are used in combination with gene selection methods to overcome the challenge [8]. Different evolutionary algorithms such as Genetic Algorithm (GA), Ant Colony Optimization (ACO), Artificial Bee Colony Algorithm (BOA), World Competitive Contests (WCC) [9] and so forth can be applied to deal with this problem.

Genetic algorithm is a heuristic search algorithm and an optimization method inspired by natural evolution [10]. Based on the literature, there are some heuristic search algorithms based on GA which have been utilized in gene selection. For instance, a new hybrid algorithm which is a combination of GA and K-Nearest Neighbor joining (KNN) was introduced in Li et al., 2001. A genetic algorithm based on probability model (PMBGA) with a combination of KNN is another related method which

* Corresponding author. Tel.: +98 21 6695 9256; fax: +98 21 6640 4680.
E-mail address: amasoudin@ut.ac.ir (A. Masoudi-Nejad).
URL: http://lbb.ut.ac.ir/

https://doi.org/10.1016/j.imu.2017.10.004
Received 10 August 2017; Received in revised form 14 October 2017; Accepted 14 October 2017
Available online 20 October 2017

Table 1
Profile of cancer microarray datasets.

Microarray datasets	no. of Classes	No. of Samples	no. of Genes
Colon	2	62 (Tumor:40,normal:22)	2000
ALL_AML	2	72 (ALL:47, AML:11)	7129
SRBCT	4	63 (EWS:23, NHL:8, NB:12, RMS:20)	2308
MLL	3	72 (ALL:24, MLL:20, AML:28)	12582
Tumors_9	9	60 (NSCLC:9, Colon:7, Breast:8, Ovary:6, Leukemia:6, Renal: 8, Melanoma:8, Pprostate:2, CNS:6)	5727
Tumors_11	11	174 (Ovary:27, Bladder/ureter:8, Breast:26, Colorectal:23, Gastroesophagus:12, Kidney:11, Liver: 7, Prostate:26, Pancreas:6, Lung Adeno:14, Lung Squamous:14)	12534

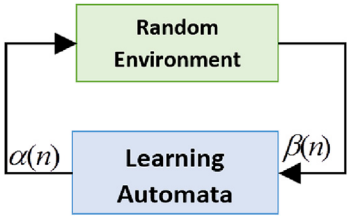

Fig. 1. Learning automata connection with environment [27].

was applied on three datasets in 2005 [11]. In another case, a Genetic algorithm was combined with SVM and Gaussian kernel function to obtain an acceptable accuracy in classification [12]. Moreover, the combination of GA, PSO, and SVM was applied on ovarian cancer dataset [13].

Yang et al. proposed two hybrid algorithms [14], namely GA-MTL

Fig. 3. An example of a random chromosome. This chromosome has four genes. Every number inside a gene corresponds to a selected gene index from the input dataset.

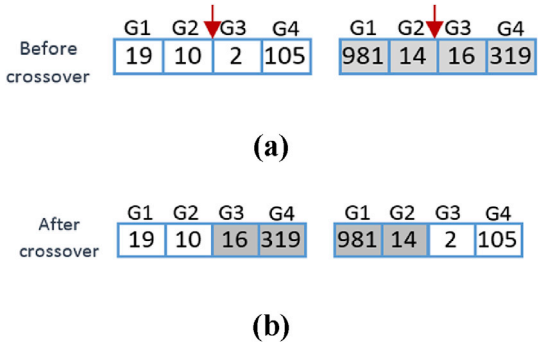

Fig. 4. An example of a crossover operator. (a) Two new chromosomes before crossover. (b) Two random chromosomes after crossover.

inspired by genetic algorithm based on multipurpose learning and E-GA-MTL based on improved GA-MTL. In Ref. [15], a novel heuristic algorithm based on genetic algorithms and artificial intelligence is applied on gene selection problem. The Combination of genetic algorithm and Information Gain are used for the gene selection problem in Ref. [16]. In another study BPSO and compact GA (CGA) algorithms were employed in a heuristic search and KNN was used for classification [17]. GA in combination with ABC algorithm which selects a subset of genes with high accuracy is one other related work [18]. In paper [19], a new recently published algorithm is used to solve the gene selection problem

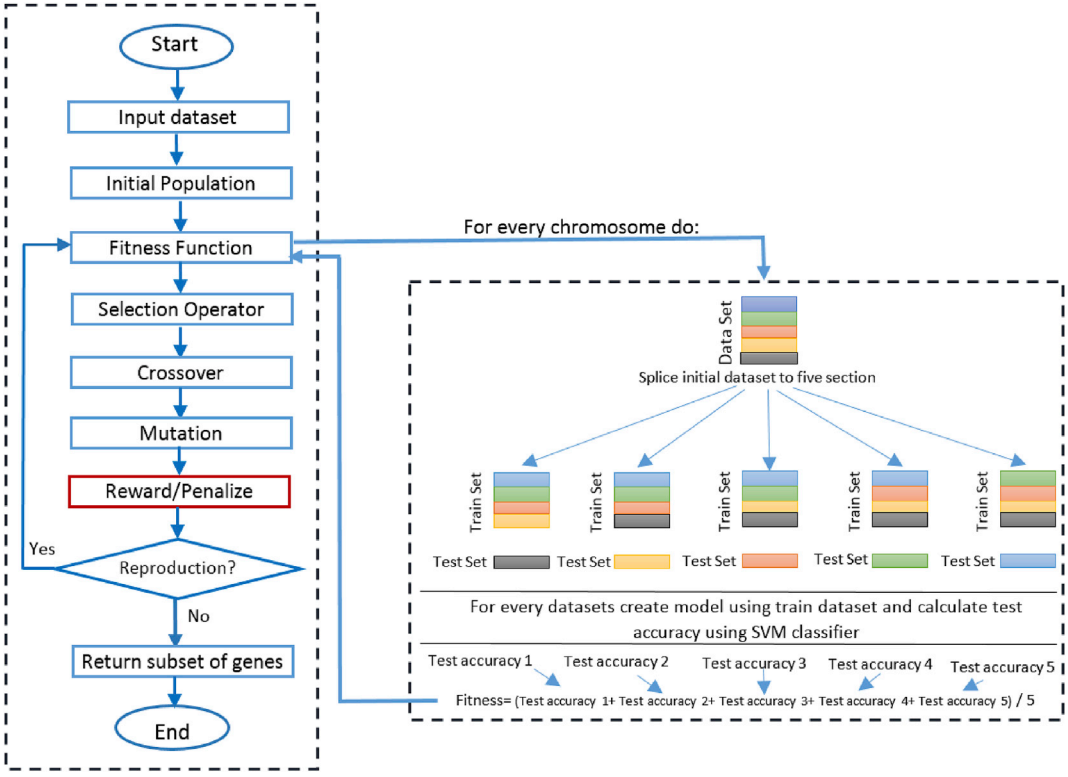

Fig. 2. The GALA flowchart. Fitness value of every chromosome calculated through SVM classifier.

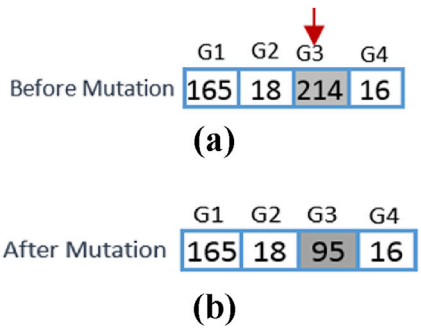

Fig. 5. An example of mutation operator. (a) Resulted chromosome before mutation operator. (b) A random chromosome after mutation operator.

by combining of genetic algorithm and Harmony Search method. In addition to the above-mentioned studies which are mostly based on genetic algorithm, additional heuristic or hybrid methods have also been proposed in the literature.

In this paper, a hybrid meta-heuristic search algorithm based on the integration of genetic algorithm and learning automata (GALA) is proposed for gene selection problem through different cancer datasets [20]. Learning automata (LA) have been proposed as learning model and it is an abstract model, which selects an action from action sets and applies it to the environment [21]. Then, the environment evaluates the selected action and produces a proper answer. In LA, an action is selected based on predefined rules for every learning iteration, and it is penalized or rewarded.

On the other hand, genetic algorithm strives to converge to an optimal answer regarding fitness function. In this algorithm, the positions of genes in chromosomes are disordered. GA chooses the best chromosome from the population. If a better position of gene in chromosome is selected, it would be possible to find the best answer among fewer generations. Through combining the merits of both GA and LA, our proposed algorithm tries to converge to an optimal answer in fewer steps. In GALA, LA operators (reward and penalty) are added to GA. Consequently, each chromosome in GA is equal to an automaton in LA, and, similarly, each gene in GA is equivalent to an action of an automaton.

In order to evaluate the proposed algorithm, GALA, we applied it on six different cancerous microarray datasets including both two and multi-class datasets. The simulation results demonstrated that GALA algorithm had the best performance in comparison to the recently used gene selection methods.

Table 2
GALA algorithm parameters.

GALA parameters	
Population size	100
Crossover rate	0.70
Mutation rate	0.30
Memory depth	3
Generation no.	100

2. Material and methods

2.1. Dataset

In this study, different cancer microarray datasets were used to evaluate GALA algorithm. We evaluated the performance of GALA algorithm and other previously proposed algorithms for gene selection problem using six binary- and multi-class microarray cancerous datasets.

Fig. 6. Reward operator on G2 gene.

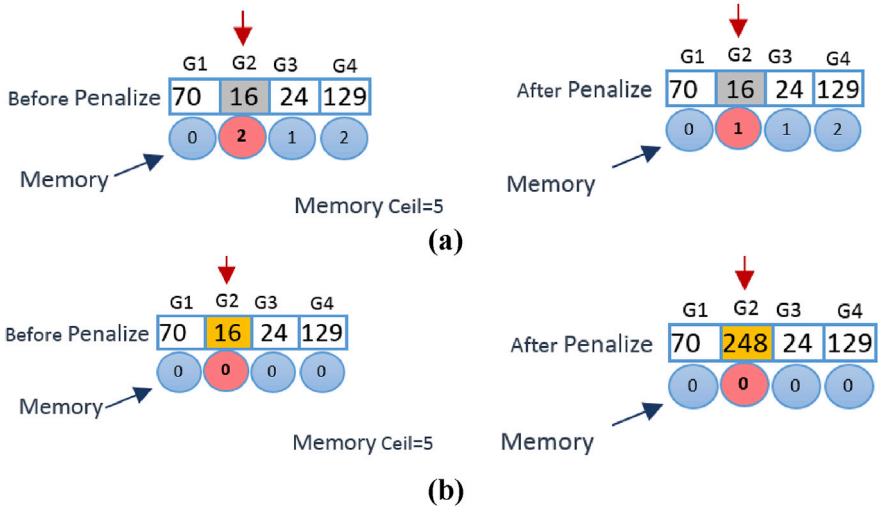

Fig. 7. Penalty operator on gene G2: (a) Memory size of gene G2 reduced one value. (b) Gene G2 value changed to a random available value.

Table 3
Experimental results of 40 independent runs, using GALA algorithm on all mentioned datasets with different genes.

Number of selected genes	Dataset name	Accuracy (%)	Sensitivity (%)	Specificity (%)	Run time avg. (sec.)
8	Colon	99.46 ± 0.68	96.06 ± 0.15	93.15 ± 0.13	734.97
9	Colon	99.81 ± 0.25	98.95 ± 0.28	94.65 ± 0.09	849.77
10	Colon	99.77 ± 0.41	99.45 ± 0.19	95.32 ± 0.18	98.33
2	ALL_AML	100 ± 0.0	98.51 ± 0.28	97.45 ± 0.16	343.49
4	SRBCT	97.35 ± 0.91	94.26 ± 0.16	96.78 ± 0.34	389.69
5	SRBCT	98.50 ± 0.51	98.11 ± 0.25	95.14 ± 0.18	503.22
6	SRBCT	99.34 ± 0.76	97.36 ± 0.35	94.15 ± 0.27	614.06
3	MLL	93.96 ± 0.99	95.43 ± 0.29	91.24 ± 0.23	638.70
8	Tumors_9	82.11 ± 0.36	86.32 ± 0.18	81.12 ± 0.29	1506.20
10	Tumors_9	86.52 ± 0.95	88.25 ± 0.29	83.46 ± 0.47	1980.30
8	Tumors_11	82.16 ± 0.57	86.14 ± 0.21	80.89 ± 0.29	2983.26
10	Tumors_11	84.38 ± 0.28	85.24 ± 0.59	83.02 ± 0.41	3784.16

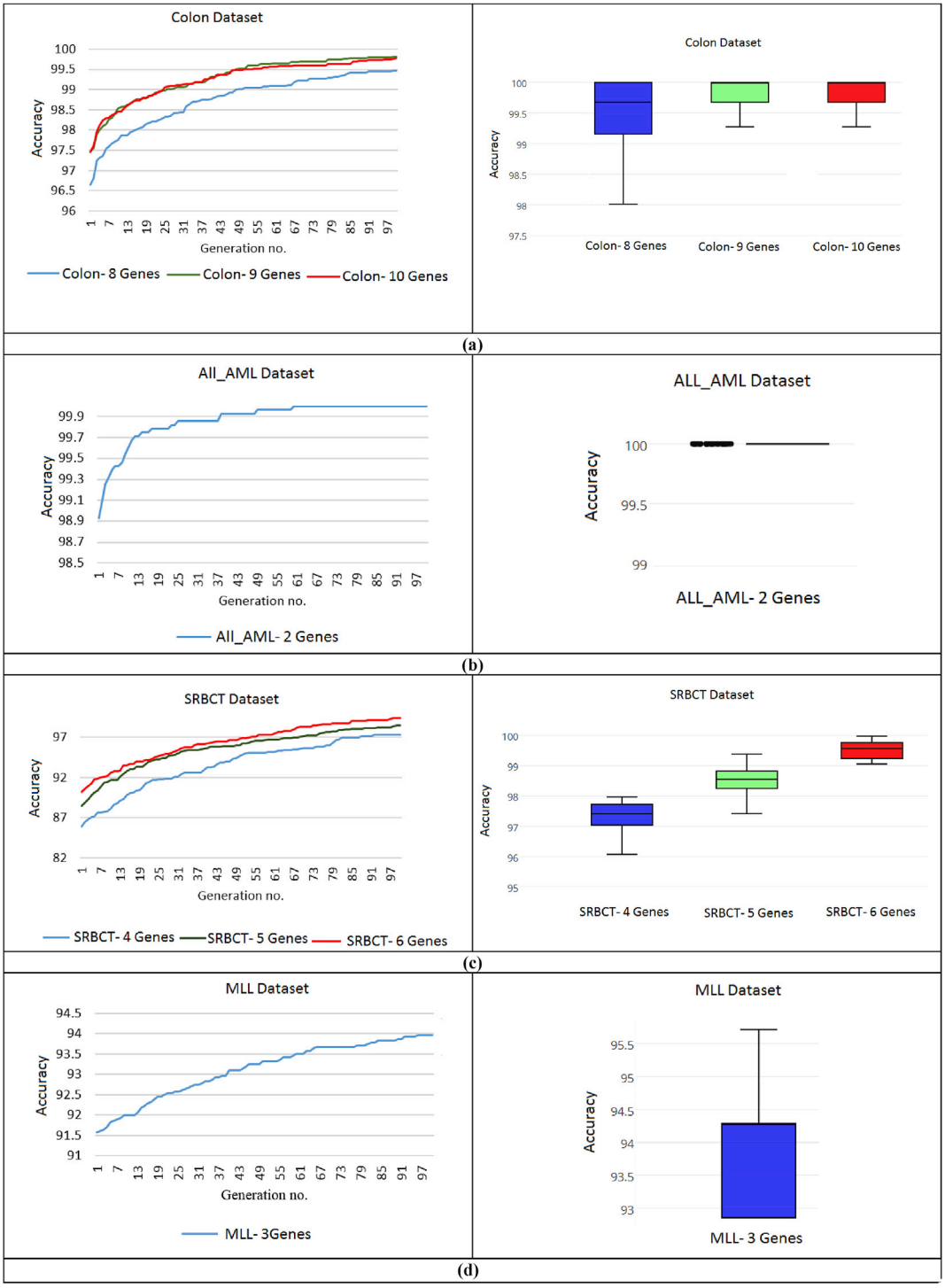

Fig. 8. Average convergence of GALA algorithm along with box plot for all mentioned datasets over 40 independent runs. **(a)** Colon dataset with 8,9 and 10 genes, **(b)** ALL_AML dataset with 2 genes, **(c)** SRBCT dataset with 4,5 and 6 genes, **(d)** MLL dataset with 3 genes, **(e)** Tumors_9 with 8 and 10 genes and **(f)** Tumors_11 with 8 and 10 genes.

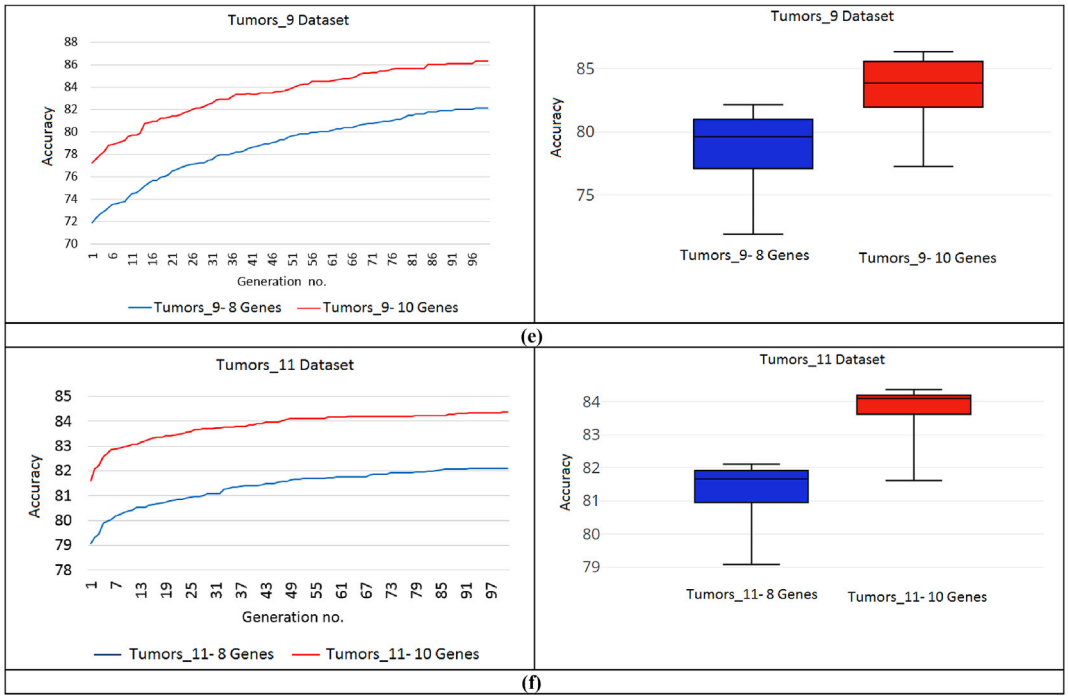

Fig. 8 (*continued*).

Table 4
Comparison of the methods on the Colon dataset.

Experiments	Algorithms	Gene no.	Accuracy (%)
This paper	GALA	8	99.46
		9	99.81
		10	99.77
[34]	GA + SVM	15	93.6
[22]	SMMDA	8	92.19
[35]	mAnt	8	91.5
[36]	mRMR-PSO	10	90.32
[18]	GBC	10	98.38
[37]	mRMR-ABC	15	96.77
[38]	PSO	20	85.48
[39]	GA	12	93.55
[33]	AAElastic	28	96.40
[18]	mRMR-PSO	78	93.55
[18]	mRMR-GA	83	95.61

Table 5
Comparison of the methods on the ALL_AML dataset.

Experiments	Algorithms	Gene no.	Accuracy (%)
This paper	GALA	2	100
[34]	GA + SVM	4	100
[6]	IBPSO	2	100
[23]	PSO + KNN	3	97.0588
[41]	SVM-RFE	4	97.06
[22]	SMMDA	8	96.88
[40]	PLS–RFE	16	SVM (100)
		12	KNN (100)
[42]	Stepwise	3	88.14
	Pure TS	5	94.24
	Pure PSO	7	94.19
	HPSOTS	7	95.81
[43]	PLSVIP	9	100
	PLSIEG	8	100
[44]	MSFCM + WNN	10	98.61
[45]	Probabilistic mechanism	–	SVM [45] 97.38
[46]	RNBC	–	RNBC [46] RNBC 94.29
[47]	GSA	10	100

The analyzed binary-class microarray datasets were Colon[1] [22] and ALL_AML[2] [23], and the multi-class microarray datasets were SRBCT[3] [23], MLL[4] [23], Tumors_9[5] [24] and Tumors_11[5] [24].

The Colon dataset contained two classes of normal and tumor data. The ALL_AML encompassed two classes of cancers, which are acute lymphoblastic leukemia (ALL) and acute myeloid leukemia (AML). The SRBCT dataset consisted of four classes of cancers. These four classes were ewing sarcoma (EWS), non-Hodgkin lymphoma (NHL), neuroblastoma (NB), and rhabdomyosarcoma (RMS). The MLL dataset consisted of three classes of lymphoblastic leukemia (ALL), myeloid lymphoid leukemia (MLL), and acute myeloid leukemia (AML). The Tumors_9 and Tumors_11 datasets contain 9 and 11 classes of different cancers data, respectively. The properties of all stated datasets are shown in Table 1.

2.2. Genetic algorithm

Genetic algorithm was invented by john Holland in 1960 [25]. This algorithm is an optimization algorithm, which is inspired by evolution process in nature [25]. The first step of GA is chromosome constitution. Every chromosome generates a single answer for the problem. New answers are produced after applying crossovers, mutations, and selection operations. Fitness function evaluates the merits of chromosomes. Finding the most deserving chromosome with maximum fitness function value in generations is the goal of this algorithm. Many circumstances such as the number of initial population, the number of generations,

[1] http://cilab.ujn.edu.cn/datasets.htm.
[2] http://www.broadinstitute.org/cgi-bin/cancer/datasets.cgi.
[3] https://research.nhgri.nih.gov/microarray/Supplement/Images/supplemental_data.
[4] http://mldata.org/repository/data/viewslug/leukemia-mll/.
[5] http://www.gems-system.org/.

Table 6
Comparison of the methods on the SRBCT dataset.

Experiments	Algorithms	Gene no.	Accuracy (%)
This paper	GALA	4	97.35
		5	98.50
		6	99.34
[48]	SFS + LDA with NCC	4	100
	SFS + Bayes classifier	4	90
	SFS + NNC	4	95
[6]	IBPSO	6	100
[23]	PSO + KNN	6	100
[49]	FSMLP + NERFCM	7	100
[50]	ANN	96	100
[51]	NSC	43	100
[41]	SVM-RFE	19	100
[43]	PLSVIP	24	100
	PLSVEG	15	100
[44]	MSFCM + WNN	10	100
[52]	KLLE + LLE + PCA	20	100
[53]	AGA + KNN	14	100
[54]	PSODT	–	92.94

crossover operators, mutation operators and fitness functions determine genetic algorithm performance. For more accuracy in fitness function, fewer generations are necessary in order to reach an optimal answer.

2.3. Learning automata

Learning automata, considered as an abstract model, selects an action from action sets and then applies it to the environment which evaluates the mentioned action [26]. The environment returns the results of the action using a boosted signal [26]. After receiving the boosted signal, learning automata changes its situation and then selects the next action [27,28]. A schematic representation of LA approach is depicted in Fig. 1. An environment can be considered as $E = \{\alpha, \beta, c\}$, in which $\alpha = \{\alpha_1, \alpha_2, ..., \alpha_r\}$ is the set of inputs, $\beta = \{\beta_1, \beta_2, ..., \beta_r\}$ is the set of outputs and $c = \{c_1, c_2, ..., c_r\}$ is the set of penalty rates.

2.4. Hybrid GALA algorithm

In this section, GALA algorithm as a hybrid algorithm is introduced. GALA draws upon the plus points of both GA and LA. Every chromosome in which gene positions are absolutely random has a fitness value calculated by fitness function. In other words, GA assigns a score for a chromosome. The most deserving chromosome will be generated if we assign a score for gene locations in a chromosome. LA, in which each automata is equivalent to a chromosome and each action is equivalent to a gene, is used for this purpose. Gene places are determined by reward and penalize operators. This issue enhances convergence speed. Further details on how to combine these two methods are well explained in Ghader et al., 2010a. We adopted the same combination method in our study. The flowchart of GALA is depicted in Fig. 2 and its pseudo code is demonstrated in Algorithm 1.

Table 7
Comparison of the methods on the MLL dataset.

Experiments	Algorithms	Gene no.	Accuracy (%)
This paper	GALA	3	93.96
[48]	SFS + LDA with NCC	4	100
	SFS + Bayes classifier	4	100
	SFS + NNC	4	93
[6]	IBPSO	4	100 (LOOCV)
[46]	RNBC	–	(10CV) RNBC 87.14
[54]	PSODT	–	100(5CV)
[23]	PSO + KNN	4	100

Table 8
Comparison of our proposed algorithm with some important methods that Chen et al. reported in Ref. [24] for Tumors_9 and Tumors_11 datasets. The reported results for the proposed algorithm are the average results of 40 independent runs.

Microarray datasets	Algorithms	Gene no.	Accuracy (%)
Tumors_9	This paper	8	82.11 ± 0.36
	This paper	10	86.52 ± 0.95
	PSODT	57	74.00
	SVM		76.00
	SOM		40.00
	BPNN		34.00
	C4.5		52.00
	NB		70.00
	CART		62.00
	AIRS		48.00
Tumors_11	This paper	8	82.16 ± 0.57
	This paper	10	84.38 ± 0.28
	PSODT	125	97.52
	SVM		93.85
	SOM		78.76
	BPNN		69.38
	C4.5		90.31
	NB		80.3
	CART		84.73
	AIRS		59.61

2.4.1. The initial population

The initial population consists of a number of randomly created chromosomes. Each chromosome includes a number of genes which represent a gene index from input dataset. Fig. 3 shows a chromosome instance. In this example, the chromosome involves four genes and numbers in the genes correspond to the feature number.

2.4.2. The fitness function

We divided every dataset into five equal parts in order to calculate fitness value. Then, we selected one of the mentioned parts as a test set and the rest as training set. We repeated this action five times for every separate part. SVM was also applied as the classification model. The fitness value was equal to the average of acquired test accuracies from every repetition. The details of train and test datasets for classification and fitness values are depicted in Fig. 2.

2.4.3. Crossover operator

A crossover operator creates new chromosomes. Some parts of chromosomes are changed using this operator. The crossover procedure takes more than one parent solution and generates the same

Table 9
Index of selected genes for best run using GALA algorithm for all cancer datasets.

Dataset name	Gene no.	Selected genes indexes	Accuracy (%)
Colon	8	237, 739, 570, 1540, 1804, 1050, 478, 1307	100
Colon	9	214, 947, 1503, 1047, 1886, 187, 702, 803, 608	100
Colon	10	466, 1324, 1290, 840, 1565, 1546, 513, 132, 607, 67	100
ALL_AML	2	6660, 6553	100
SRBCT	4	1158, 1481, 2025, 1319	99.80
SRBCT	5	195, 26, 823, 1003, 1608	99.38
SRBCT	6	1066, 426, 1389, 1716, 1543, 401	99.94
MLL	3	1853, 1866, 1501	95.71
Tumors_9	8	114, 4817, 2171, 4919, 5650, 3116, 593, 1618	85.42
Tumors_9	10	5607, 3183, 4122, 176, 4858, 3912, 3502, 3377, 631, 4659,	89.15
Tumors_11	8	2654, 8962, 4152, 7854, 8550, 5698,11541, 10547	84.65
Tumors_11	10	10248, 568, 3571, 5002, 9852, 115, 698, 8856, 8900, 1674	85.23

number of child solutions from them. GALA uses a single-point crossover operator to create a new chromosome [25]. In a single-point crossover operator, two chromosomes are randomly selected and are connected together from the middle point. A sample crossover operator is shown in Fig. 4.

2.4.4. Mutation operator

Mutation operator was applied to maintain genetic variety across two generations. In mutation, the solution might, completely, be different from the previous solution. Therefore, GA can be optimized to arrive at better results through employing a mutation operator. The mutation operator type in this algorithm is order-based mutation. In this type of mutation, two genes are randomly selected, and the location of each gene is changed. This operator is illustrated in Fig. 5.

2.4.5. Penalty and reward operators

For every action of automata, a memory size was defined. All actions of memory size were zero at first. An action memory size shows its importance. This means that the action with the greatest importance gets more memory. The reward and the penalty can be calculated as follows:

1 Consider an automata.
2 Calculate fitness value for the selected automata (consider it as x).
3 Select an action randomly from the current automata and set a random value for it.
4 Calculate the fitness of changed automata (consider it as y).
5 Reward the automata if $x \geq y$ and penalize it if $x < y$.
6 Reward: Add 1 to the size of memory and do not change the action value if threshold memory is not satisfied.
7 Penalize: Subtract 1 out of memory size and do not change action value if the size of the memory is more than zero. If the size of the memory equals zero, do not change the memory size and change the action value to a random value. In Figs. 6–7, reward and penalty operations are depicted.

2.4.6. Selection operator and elitism

Selection and elitism operators transfer the number of chromosomes to the next generation. In the present study, 10% of individuals with the best fitness values were selected to be passed to the next generation, avoiding the crossover and mutation operators and the remaining individuals were selected through a roulette-wheel selection operator. In roulette-wheel selection operator [29], probability value is assigned for each chromosome whose sum amounts to 1.

2.4.7. GA termination

There are some different conditions for the termination of algorithm in Genetic Algorithm. In this paper, the generation number was declared, at first, and then the algorithm was implemented according to the number.

2.4.8. Implementation environment

We implemented GALA algorithm in MATLAB[6] environment on a system with core i5 CPU and 2G memory. Moreover, to evaluate SVM classifier, we used Lib-SVM package [30], that is accessible from (http://www.csie.ntu.edu.tw/~cjlin/libsvm). The kernel function used in SVM classifier is a radial basis function. Furthermore, 5-fold cross validation method was used for SVM validation.

Algorithm.1. GALA Algorithm.

```
Algorithm.1: GALA Algorithm()
{
N=Number of selected genes
P= Population size
G= Generation
Cx= Cross-over rate
Mu= Mutation rate
Mem=memory depth
//initial population//
Do (p) times {
  Generate random chromosome Ch.
  for i=1 to N {
    Ch. Gene (i).Memory=0
  }
}
//Optimization iterations//
Do G times
{
Copy 10% of current population to new population.   //elitism
Generate 90% of remaining population
    //crossover operator//
Do (Cx*P) times {
    i= random chromosome 1
    j= random chromosome 2
    Do Single-point crossover on chromosome (i) and chromosome (j)
     Copy two new generated chromosomes into new population
                    }
  In new population do {
      //mutation operator//
      Do (Mu*P) times {
          i= random chromosome
          j=random gene from chromosome(i)
          Mutate chromosome(i).Gene(j)
                    }
    //Reward & Penalize operator//
  For each chromosome {
  v1= fitness (chromosome)
  i =random gene from chromosome
  Modified chromosome= chromosome
  Modified chromosome. Gene (i)= random available gene number
  v2= fitness(Modified chromosome)
  If(v1>=v2) then // Reward chromosome//
   If(chromosome.gene(i).Memory<Mem)
    chromosome.gene(i).Memory= chromosome.gene(i).Memory+1
  else {  //Penalize current chromosome//
   If(chromosome.gene(i).Memory>0)
    chromosome.gene(i).Memory= chromosome.gene(i).Memory-1
    else
      chromosome=modified chromosome
          }
      }
}
   For each chromosome from new population
      chromosome.fitness_value=fitness(chromosome)
   }
Return best chromosome from last population as result
}

Function fitness (Chromosome)
{
   Evaluate 5-fold Cross Validation test accuracy using SVM classifier
   Return 5-fold Cross Validation test accuracy
}
```

2.4.9. Time complexity of GALA algorithm

To evaluate and assess GALA algorithm performance, we calculated and analyzed the time complexity of the algorithm. Before calculating the time complexity, we assumed the following parameters:

G: Generation number	***m***: Initial population size	***n***: number of dataset samples
Mu: Mutation rate	*Cx*: Crossover rate	

The time complexity of Lib-SVM is O (n^3) where n is the number of dataset samples [31,32]. Therefore, the overall time complexity of GALA algorithm is:

[6] MATLAB Release R2012b, The MathWorks, Inc., Natick, Massachusetts, United States.

$$G.\left[m.n^3 + m^2 + m.\left(Mu + Cx + n^3\right)\right]$$

Considering that fitness value of each chromosome was estimated using SVM, time complexity of the fitness value estimation is O (n^3) for all chromosomes. Hence, in this equation, the number of calculations is $O(m.n^3)$ for the estimation of fitness value of all chromosomes of a generation. After calculating fitness value for all chromosomes, they were sorted based on fitness value, cost of which is m^2. Since the cost of mutation and crossover operators is a fixed value, such values are not considered in accounting for time complexity. Time complexity of reward and penalty is of O (n^3) order because fitness value of a chromosome is calculated twice for the implementation of penalty and reward acts, before and after penalizing or rewarding. Therefore, time complexity of GALA algorithm is, in general, expressed as follows:

$$G.\left[m.n^3 + m^2 + m.n^3\right] \in O\left(G.m.n^3\right)$$

3. Experimental results

In this section, performance evaluation of GALA algorithm and other previously proposed algorithms for gene selection problem on six binary and multi-class microarray cancer datasets is described.

Table 2 shows GALA control parameters. Memory depth parameter determines memory size for actions or genes. For all genes, memory size of zero is assigned at first. Finally, generation number parameter shows the number of iterations in which GALA tries to reach to an acceptable answer. In our study, GALA was applied on datasets with different number of genes as indicated in Table 1. To evaluate our proposed algorithm, we used three measures of accuracy, sensitivity, and specificity. Also, every result obtained over 40 independent runs and their mean and standard deviations are displayed in Table 3.

Fig. 8 depicts average GALA convergence along with box plot over 40 time independent runs on the datasets described in Table 1. Fig. 8(a) shows average GALA convergence on colon dataset with 8, 9 and 10 genes, and Fig. 8(b) represents average GALA convergence on ALL_AML dataset with 2 genes, and finally Fig. 8(c) and (d) show average GALA convergence on SRBCT dataset with 4, 5 and 6 genes and MLL dataset with 3 genes, respectively. Fig. 8(e) and (f) show GALA average convergence on Tumors_9 and Tumors_11 datasets, respectively with 8 and 10 genes. The results of GALA on all mentioned datasets are given in Tables 4–8 and are compared with previously proposed algorithms.

Alshman et al. [18] combined GA and ABC (artificial bee colony) and SVM classifier and reached 98.38% of accuracy with ten genes on Colon dataset. They also used LOOCV cross validation method in order to evaluate their algorithm. Jian Yang et al. [22] applied SMMDA method on colon and ALL_AML datasets. They attained 92.19% and 96.88% of classification accuracies with 8 genes, respectively. Lee et al. [33] reported 96.40% of accuracy with 28 genes on the colon dataset. Since no result was obtained for IBPSO algorithm on colon dataset, GALA was not compared with this algorithm. In this study, we applied GALA on the colon dataset and selected 8 genes with a mean accuracy of 99.46%. Furthermore, we applied GALA with nine and ten genes with average selection accuracies of 99.81% and 99.77%, respectively, while the GA, without combining with the learning automata, has precision 98% with 15 genes. Table 4 demonstrates these results along with other previously proposed algorithms.

Using modified PSO and SVM classifiers, Saberi et al. [6], reached an accuracy rate of 100% on ALL_AML and MLL datasets with two and four genes. Ning An [40] reached 100% accuracy with 16 and 12 genes using PLS–RFE-SVM and PLS–RFE-KNN methods on ALL_AML dataset, respectively. GA and SVM classifiers obtained 100% accuracy on the ALL_AML dataset with 4 genes, whereas GALA achieved this level of accuracy with selecting only two genes on the same dataset. The findings of other algorithms on ALL_AML dataset with different gene numbers are given in Table 5.

Sharma et al. [48] selected four influential genes with perfect accuracy rate on SRBCT dataset. Moreover, they acquired multiple results with various parameters. Their best result was acquired through the combination of LDA and SFS which could select four genes with complete accuracy. GALA obtained 99.34% mean accuracy on the same dataset with six genes and 97.35% and 98.50% average accuracy with 4 and 5 genes, respectively. The results of other algorithms are available in Table 6.

Chen et al. obtained 100% accuracy with ten percent of all genes using PSO algorithm and decision tree (DT) classifiers [54]. GALA accomplished 93.96% average accuracy with three genes. The results of other algorithms are shown in Table 7.

In the above section, we compared the efficiency of the proposed algorithm with the recently introduced algorithms. These comparisons were evaluated on a datasets with less than 5 classes. To demonstrate the effectiveness of the proposed algorithm on datasets with a large number of classes, we compared the proposed algorithm along with several other important algorithms on the Tumors_9 and Tumors_11 datasets with 9 and 11 classes, respectively. Details of the results have been shown in Table 8. The results of other algorithms in this table have been derived from Chen et al. [24] article.

For all experimental datasets, the best result for every dataset was selected as the candidate result, and indices of the selected genes are reported in Table 9.

4. Conclusion

In this paper, a hybrid algorithm named GALA, derived from the combination of genetic algorithm and learning automata, was introduced for gene selection problem in cancer microarray datasets. It employed both genetic algorithm and learning automata and strived to identify genes having more power in the expressed data classification. The SVM classifier was applied as the classification model in our proposed algorithm. In most of the cases, the findings obtained through GALA from its implementation on six cancer gene expression datasets were remarkable compared to other recently introduced algorithms. Mean classification accuracies of GALA on colon dataset with 8 genes was found to be 99.46% and for ALL_AML, SRBCT and MLL datasets with 2, 4 and 3 genes were 100%, 97.35% and 93.96%, respectively and mean classification accuracies on Tumors_9 and Tumors_11 datasets with 10 genes were 86.52%, and 84.38%, respectively. Compared to other algorithms, GALA could select subset of genes possessing a higher accuracy resolution. The time complexity of GALA is $O(G.m.n^3)$ where n is the number of dataset samples, G is the generation number and m is the initial population size. This algorithm can be used in every feature selection and classification problems.

References

[1] Alba E, García-Nieto J, Jourdan L, Talbi E-G. Gene selection in cancer classification using PSO/SVM and GA/SVM hybrid algorithms. In: Evolutionary computation, 2007. CEC 2007. IEEE congress on. IEEE; 2007. p. 284–90.
[2] Narendra PM, Fukunaga K. A branch and bound algorithm for feature subset selection. Comput IEEE Trans 1977;100:917–22.
[3] Ghorai S, Mukherjee A, Sengupta S, Dutta PK. Multicategory cancer classification from gene expression data by multiclass NPPC ensemble. In: Systems in medicine and biology (ICSMB), 2010 international conference on. IEEE; 2010. p. 41–8.
[4] Guo S-B, Lyu MR, Lok T-M. Gene selection based on mutual information for the classification of multi-class cancer. Comput Intell Bioinforma 2006:454–63. Springer.
[5] Xue B, Zhang M, Browne W, Yao X. A survey on evolutionary computation approaches to feature selection. 2015.
[6] Mohamad MS, Omatu S, Deris S, Yoshioka M. A modified binary particle swarm optimization for selecting the small subset of informative genes from gene expression data. Inf Technol Biomed IEEE Trans 2011;15:813–22.
[7] Jagga Z, Gupta D. Machine learning for biomarker identification in cancer research-developments toward its clinical application. Pers Med 2015;12:371–87.
[8] Alshamlan H, Badr G, Alohali Y. A comparative study of cancer classification methods using microarray gene expression profile. In: Proceedings of the first international conference on advanced data and information engineering (DaEng-2013). Springer; 2014. p. 389–98.

[9] Masoudi-Sobhanzadeh Y, Motieghader H. World Competitive Contests (WCC) algorithm: a novel intelligent optimization algorithm for biological and non-biological problems. Inf Med Unlocked 2016;3:15–28.
[10] McCall J. Genetic algorithms for modelling and optimisation. J Comput Appl Math 2005;184:205–22.
[11] Paul TK, Iba H. Gene selection for classification of cancers using probabilistic model building genetic algorithm. BioSystems 2005;82:208–25.
[12] Yong M, Xiao-bo Z, Dao-ying P, You-xian S, Tc WS. Parameters selection in gene selection using Gaussian kernel support vector machines by genetic algorithm. J Zhejiang Univ Sci B 2005;6:961–73.
[13] Lee Z-J. An integrated algorithm for gene selection and classification applied to microarray data of ovarian cancer. Artif Intell Med 2008;42:81–93.
[14] Yang JY, Li G-Z, Meng H-H, Yang MQ, Deng Y. Improving prediction accuracy of tumor classification by reusing genes discarded during gene selection. BMC genomics 2008;9:S3.
[15] Dashtban M, Balafar M. Gene selection for microarray cancer classification using a new evolutionary method employing artificial intelligence concepts. Genomics 2017;109:91–107.
[16] Salem H, Attiya G, El-Fishawy N. Classification of human cancer diseases by gene expression profiles. Appl Soft Comput 2017;50:124–34.
[17] Chuang L-Y, Yang C-H, Li J-C, Yang C-H. A hybrid BPSO-CGA approach for gene selection and classification of microarray data. J Comput Biol 2012;19:68–82.
[18] Alshamlan HM, Badr GH, Alohali YA. Genetic bee colony (GBC) algorithm: a new gene selection method for microarray cancer classification. Comput Biol Chem 2015;56:49–60.
[19] Das K, Mishra D, Shaw K. A metaheuristic optimization framework for informative gene selection. Inf Med Unlocked 2016;4:10–20.
[20] Ghader HM, Fakhr K, Arzil SA. A hybrid method for task scheduling. In: Education technology and computer (ICETC), 2010 2nd international conference on. IEEE; 2010. pp. V1-91-V91-95.
[21] Narendra KS, Thathachar MA. Learning automata-a survey. IEEE Trans Syst Man Cybern 1974:323–34.
[22] Cui Y, Zheng C-H, Yang J, Sha W. Sparse maximum margin discriminant analysis for feature extraction and gene selection on gene expression data. Comput Biol Med 2013;43:933–41.
[23] Kar S, Sharma KD, Maitra M. Gene selection from microarray gene expression data for classification of cancer subgroups employing PSO and adaptive K-nearest neighborhood technique. Expert Syst Appl 2015;42:612–27.
[24] Chen K-H, Wang K-J, Tsai M-L, Wang K-M, Adrian AM, Cheng W-C, et al. Gene selection for cancer identification: a decision tree model empowered by particle swarm optimization algorithm. BMC Bioinforma 2014;15:49.
[25] Mitchell M. An introduction to genetic algorithms. MIT press; 1998.
[26] Narendra KS, Thathachar MA. Learning automata: an introduction. Courier Corporation; 2012.
[27] Ghader HM, Fakhr K, Javadi M, Bakhshzadeh G. Static task graph scheduling using learner Genetic Algorithm. In: Soft computing and pattern recognition (SoCPaR), 2010 international conference of. IEEE; 2010. p. 357–62.
[28] Oommen BJ, Ma DCY. Deterministic learning automata solutions to the equipartitioning problem. IEEE Trans Comput 1988:2–13.
[29] Lipowski A, Lipowska D. Roulette-wheel selection via stochastic acceptance. Phys A Stat Mech Appl 2012;391:2193–6.
[30] Chang C-C, Lin C-J. LIBSVM: a library for support vector machines. ACM Trans Intell Syst Technol (TIST) 2011;2:27.
[31] Abdiansah A, Wardoyo R. Time complexity analysis of support vector machines (SVM) in LibSVM. Int J Comput Appl 2015;128(3):1–7.
[32] Tsang IW, Kwok JT, Cheung P-M. Core vector machines: fast SVM training on very large data sets. J Mach Learn Res 2005;6:363–92.
[33] Algamal ZY, Lee MH. Regularized logistic regression with adjusted adaptive elastic net for gene selection in high dimensional cancer classification. Comput Biol Med 2015;67:136–45.
[34] Li S, Wu X, Hu X. Gene selection using genetic algorithm and support vectors machines. Soft Computing-A Fusion Found Methodol Appl 2008;12:693–8.
[35] Yu H, Gu G, Liu H, Shen J, Zhao J. A modified ant colony optimization algorithm for tumor marker gene selection. Genomics Proteomics Bioinforma 2009;7:200–8.
[36] Abdi MJ, Hosseini SM, Rezghi M. A novel weighted support vector machine based on particle swarm optimization for gene selection and tumor classification. Comput Math Methods Med 2012;2012.
[37] Alshamlan H, Badr G, Alohali Y. mRMR-ABC: a hybrid gene selection algorithm for cancer classification using microarray gene expression profiling. BioMed Res Int 2015;2015.
[38] Shen Q, Shi W-M, Kong W, Ye B-X. A combination of modified particle swarm optimization algorithm and support vector machine for gene selection and tumor classification. Talanta 2007;71:1679–83.
[39] Peng S, Xu Q, Ling XB, Peng X, Du W, Chen L. Molecular classification of cancer types from microarray data using the combination of genetic algorithms and support vector machines. FEBS Lett 2003;555:358–62.
[40] Wang A, An N, Chen G, Li L, Alterovitz G. Improving PLS–RFE based gene selection for microarray data classification. Comput Biol Med 2015;62:14–24.
[41] Fu LM, Fu-Liu CS. Evaluation of gene importance in microarray data based upon probability of selection. BMC Bioinforma 2005;6:1.
[42] Shen Q, Shi W-M, Kong W. Hybrid particle swarm optimization and tabu search approach for selecting genes for tumor classification using gene expression data. Comput Biol Chem 2008;32:53–60.
[43] Ji G, Yang Z, You W. PLS-based gene selection and identification of tumor-specific genes. Syst Man Cybern Part C Appl Rev IEEE Trans 2011;41:830–41.
[44] Zainuddin Z, Ong P. Reliable multiclass cancer classification of microarray gene expression profiles using an improved wavelet neural network. Expert Syst Appl 2011;38:13711–22.
[45] Wong T-T, Liu K-L. A probabilistic mechanism based on clustering analysis and distance measure for subset gene selection. Expert Syst Appl 2010;37:2144–9.
[46] Chandra B, Gupta M. Robust approach for estimating probabilities in Naïve–Bayes Classifier for gene expression data. Expert Syst Appl 2011;38:1293–8.
[47] Kumar PG, Victoire TAA, Renukadevi P, Devaraj D. Design of fuzzy expert system for microarray data classification using a novel Genetic Swarm Algorithm. Expert Syst Appl 2012;39:1811–21.
[48] Sharma A, Imoto S, Miyano S. A top-r feature selection algorithm for microarray gene expression data. IEEE/ACM Trans Comput Biol Bioinforma (TCBB) 2012;9:754–64.
[49] Pal NR. A fuzzy rule based approach to identify biomarkers for diagnostic classification of cancers. In: Fuzzy systems conference, 2007. FUZZ-IEEE 2007. IEEE international. IEEE; 2007. p. 1–6.
[50] Khan J, Wei JS, Ringner M, Saal LH, Ladanyi M, Westermann F, et al. Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. Nat Med 2001;7:673–9.
[51] Thibshirani R, Hastie T, Narasimhan B, Chu G. Diagnosis of multiple cancer types by shrunken centroids of gene expression. PNAS 2002;99:6567–72.
[52] Li X, Shu L. Kernel based nonlinear dimensionality reduction for microarray gene expression data analysis. Expert Syst Appl 2009;36:7644–50.
[53] Lee C-P, Lin W-S, Chen Y-M, Kuo B-J. Gene selection and sample classification on microarray data based on adaptive genetic algorithm/k-nearest neighbor method. Expert Syst Appl 2011;38:4661–7.
[54] Chen K-H, Wang K-J, Tsai M-L, Wang K-M, Adrian AM, Cheng W-C, et al. Gene selection for cancer identification: a decision tree model empowered by particle swarm optimization algorithm. BMC Bioinforma 2014;15:1.