Original Article

Improving Classification of Cancer and Mining Biomarkers from Gene Expression Profiles Using Hybrid Optimization Algorithms and Fuzzy Support Vector Machine

Niloofar Yousefi Moteghaed, Keivan Maghooli, Masoud Garshasbi[1]

Department of Biomedical Engineering, Islamic Azad University, Science and Research Branch, [1]Department of Medical Genetics, Faculty of Medical Sciences, Tarbiat Modares University, Tehran, Iran

Abstract

Background: Gene expression data are characteristically high dimensional with a small sample size in contrast to the feature size and variability inherent in biological processes that contribute to difficulties in analysis. Selection of highly discriminative features decreases the computational cost and complexity of the classifier and improves its reliability for prediction of a new class of samples. **Methods:** The present study used hybrid particle swarm optimization and genetic algorithms for gene selection and a fuzzy support vector machine (SVM) as the classifier. Fuzzy logic is used to infer the importance of each sample in the training phase and decrease the outlier sensitivity of the system to increase the ability to generalize the classifier. A decision-tree algorithm was applied to the most frequent genes to develop a set of rules for each type of cancer. This improved the abilities of the algorithm by finding the best parameters for the classifier during the training phase without the need for trial-and-error by the user. The proposed approach was tested on four benchmark gene expression profiles. **Results:** Good results have been demonstrated for the proposed algorithm. The classification accuracy for leukemia data is 100%, for colon cancer is 96.67% and for breast cancer is 98%. The results show that the best kernel used in training the SVM classifier is the radial basis function. **Conclusions:** The experimental results show that the proposed algorithm can decrease the dimensionality of the dataset, determine the most informative gene subset, and improve classification accuracy using the optimal parameters of the classifier with no user interface.

Keywords: *Cancer classification, fuzzy support vector machine, gene expression, genetic algorithm, particle swarm optimization algorithm*

Introduction

The DNA microarray technology allows monitoring of thousands of genes simultaneously in a single experiment. The use of this technology to monitor changes in expression levels of genes among samples can help physicians efficiently and accurately diagnose disease, classify tumors and cancer types, and propose effective treatment procedure. Gene expression is a dynamic process that provides valuable knowledge about biological networks and cellular states. The expression level of each gene indicates the activation and transcription of that gene in cell states.

The gene expression pattern of a cell or a tissue determines the structure and function of that cell or tissue. On a microarray chip, the number of genes are exceeding more than a thousand , in contrast of small number of samples. Thus, the curse of dimensionality, noisiness, and stochastic nature of this data are major problems that arise in microarray data analysis and lead to many data mining and machine learning challenges.[1-4] Determination of a small subset of relevant genes in a given dataset as a solution for high-dimensional problem can improve the classification accuracy.[3,4] Furthermore, the problem of stability can be tackled using other biological databases and bioinformatics tools such as protein–protein interaction and pathway databases.[4,5]

Several methods have been proposed for informative gene selection and classification. The Taguchi-genetic algorithm (GA) and Taguchi-particle swarm optimization (PSO) use correlation-based feature selection and are hybrid methods where *k*-NN serves as a classifier[6-8] and Shen *et al*.[9] used a modified PSO and a support vector machine (SVM). Li *et al*.[10] and Hernandeze *et al*.[11] developed a hybrid GA and SVM model. Tong and

How to cite this article: Yousefi Moteghaed N, Maghooli K, Garshasbi M. Improving classification of cancer and mining biomarkers from gene expression profiles using hybrid optimization algorithms and fuzzy support vector machine. J Med Sign Sens 2018;8:1-11.

Address for correspondence:
Dr. Keivan Maghooli, Department of Biomedical Engineering, Islamic Azad University, Science and Research Branch, Tehran, Iran. E-mail: k_maghooli@srbiau.ac.ir

Website: www.jmss.mui.ac.ir

Schierz developed a hybrid GA and a neural network classifier[12] and Li *et al*. and Yang *et al*.[13,14] used *k*-NN to apply to microarray data. Chuang *et al*.[15] proposed improved PSO and used the *k*-NN method for tumor classification. Shen *et al*.[16] developed a hybrid PSO and Tabu search with LDA classification for cancer classification.

Martinez *et al*.[17] proposed an algorithm based on swarm intelligence feature selection. Lee and Leu[18] used a GA with dynamic parameter settings, a Chi-square test for homogeneity and SVM for cancer classification. Alba *et al*.[19] combined a PSO and a GA individually with a SVM to find small samples of informative genes. Zhenyu *et al*.[20] proposed a multiple kernel SVM-based data mining and knowledge discovery system. Wang and Simon[21] used single genes to create classification models such as *k*-NN, SVM, and the random forest models. Shah and Kusiak[22] developed an integrated algorithm involving a GA and correlation-based heuristics for data preprocessing and a decision tree and SVM to make predictions. Chuang *et al*.[23] and Mao *et al*.[24] applied fuzzy SVMs to gene expression profiles to classify multiple cancer types. Ng and Chan[25] combined an information-theoretic approach with sequential forward floating searches and a decision tree. Yeh *et al*.[26] applied a GA and decision tree to build a model of selected genes. In,[27] hybrid PSO and GA algorithms are used as a feature selection method and also, in,[28] a novel-weighted SVM based on PSO are used for gene selection and tumor classification are applied on gene expression data. Chu and Wang[29] used principal component analysis, a class separability measure, the Fisher's ratio and *t*-test for gene selection and a voting scheme for multigroup classification using a binary SVM.

The present study used a hybrid GA and PSO algorithm as the feature selection method. The fitness function of each gene subset was determined using the fuzzy support vector machine (FSVM) classifier. The use of fuzzy logic in the SVM training phase decreased the effect of redundant noisy data by determining the importance of each sample in the training stage. The *t*-test method was initially used to preprocess the original gene expression data and the proposed hybrid method was then applied to select the most important subsets of genes using 10-fold cross validation. The 10-fold cross-validation accuracy of each gene subset was the evaluation criteria. One purpose of this study was to increase the classification accuracy by selecting the best parameters for a classifier using the proposed hybrid PSO/GA/FSVM algorithm without need for user trial and error. The use of a suitable combination of optimization algorithms for feature selection and selection of the proper model for the classifier improve classification results to allow accurate prediction of blind test samples.

Materials and Methods

The proposed method was evaluated using four public microarray datasets. There are several types of blood cancer and it is important to distinguish between them. The first dataset comprised 72 samples of acute lymphoblastic leukemia (ALL) and mixed lineage leukemia (MLL) cancer types with 12582 genes by Armstrong Scott.[30] The second dataset comprised 72 samples of ALL and acute myeloid leukemia (AML) cancer types with 7129 genes by Golub *et al*.[31] The third dataset generated by Alon *et al*.[32] contains the expression of 2000 genes in 62 samples for normal and colon tumor tissues. The last dataset comprised 49 samples with 7129 genes by West.[33] Table 1 provides the details of the datasets.

Genetic algorithm and particle swarm optimization

A GA is a computational optimization method that searches all parts of a solution space using different groups of feature subsets to find the best answer. The initial population is generated randomly, and then, all chromosomes are evaluated using a fitness function. The GA operators are selection, crossover, and mutation. The crossover operator creates new population by combining two chromosomes, depending on the selection operator. The crossover operator in a GA can eliminate fragmentation and genetic variation in the population. Mutation is another operator that creates a variety of solutions. The process continues to the last generation in which the best fitness is satisfied. PSO, like GA, is an algorithm inspired by the social behavior of birds in a flock.[34] This algorithm was developed by Eberhart and Kennedy.[35] In PSO, each particle moves in the search space at a velocity that is adjusted using its own memory and its neighbors' experiences. The fitness values are obtained using a fitness function.

Support vector machine

In machine learning and data mining tasks, SVMs are supervised learning algorithms associated with learning models that are used for classification and regression analysis problems. The current standard incarnation was proposed by Cortes and Vapnik.[36] SVM is specifically designed for two-class analysis problems.

Table 1: Datasets which used for testing the efficiency of proposed method

Data sets	Tissue	Sample	Number of class	Number of samples for each class	Classes	Number of Genes[#]
Amstrong, 2002	Blood	72	2	24, 48	ALL, MLL	12,582
Golub, 2002	Leukemia	72	2	47, 25	ALL, AML	7129
Alon U, 1999	Colon	62	2	22, 40	N, tumor	2000
West, 2001	Breast	49	2	25, 24	ER+ ER−	7129

ALL – Acute lymphoblastic leukemia; AML – Acute myeloid leukemia; MLL – Mixed lineage leukemia; ER – Estrogen receptor; N – Normal

Let data set $D\{(x_i, y_i) | x_i \in \mathbb{R}^p, y_i \in \{1,-1\}\}\ i=1,...,n\}$ $I = 1$, n, be, where X_i is the set of training samples and y_i are the associated labels. Each y_i can take one of two values (+1 or −1) depending the class.[36,37] In the linear case, classification of new data can be done by using the following formula:

$$max.\ Q(\alpha) = \sum_{i=1}^{N}\alpha_i - \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_j y_i y_j \quad (1)$$

$$s.t: \sum_{i=1}^{N}\alpha_i y_i = 0,\ 0 \le \alpha_i \le C\ for\ i = 1,\ldots,N$$

Where C is the soft margin constant parameter with an upper bound in the Lagrange multipliers. For the nonlinear case, SVM transforms the input data into higher dimensional feature space using a kernel function, so it can be solved as a separable case. With the use of a kernel function, the optimization problem becomes:

$$f(z) = sgn\left(\sum \alpha_i y_i k(x_i, z) + b\right) \quad (2)$$

$$max.\ Q(\alpha) = \sum_{i=1}^{N}\alpha_i - \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_j y_i y_j k(x_i, x_j) \quad (3)$$

$$s.t: \sum_{i=1}^{N}\alpha_i y_i = 0\ ,\ 0 \le \alpha_i \le C\ for\ i = 1,\ldots,N$$

Where C is the soft margin constant parameter with an upper bound in the Lagrange multipliers. The most familiar kernel functions are:

Linear kernel function $K(X_i, X_j) = X_i^T X_j$ polynomial kernel function $K(X_i, X_j) = (1 + X_i^T X_j)^p$ (p: degree), Gaussian kernel function $K(X_i, X_j) = \exp(-\frac{\| X_i - X_j \|^2}{2\sigma^2})$, (σ: Standard deviation) and sigmoid kernel function $K(X_i, X_j) = \tanh(\beta_0 X_i^T X_j + \beta_1)$ (β_0: Slope and β_1: Intercept constant).

The Gaussian kernel is one of the most useful functions and the common SVM kernel can be used in different kinds of problems. Each kernel function has its own parameters and the related parameters must be properly set to increase classification accuracy.[38]

Proposed algorithm

The proposed algorithm is a combination of the GA and PSO algorithms. The goal is to combine the properties of both algorithms by integration of GA operators into the PSO algorithm. The main difference between GA and PSO is that there are no crossovers and mutation operators in PSO; thus, it is more likely to be caught in a local minimum. The best particle in PSO can be remembered and so that it has an effect on other particles. This property increases convergence.[39-42] The hybrid PSO/GA requires the following 11 steps.

Step 1

Step 1 is the preparation of data by filtering and normalization. Most genes in databases are not useful and do not have the desired patterns for analysis of microarray data. These genes must be removed because: (a) their expression value is very low; (b) they show little change in expression value in whole samples; (c) they have low standard deviations and do not substantially change around the mean expression value and; (d) they have low information entropy. The *t*-test can then be used to examine the data to select the top-ranked genes and apply them as an input to the hybrid PSO/GA system.

Step 2

The initial values of each parameter used in the algorithm are set as shown in Table 2.

Step 3

Step 3 is to create the initial population. At first, a population with N chromosomes is randomly generated. Primary binary initialization is applied so that (1) denotes the existence of a feature in the training system and (0) denotes the absence of that feature. The lengths of the particles or chromosomes are determined by adding the number of features selected based on a statistical method (segment 1) and the 17 additional genes used to determine the optimal parameters of a classifier in the hybrid algorithm.

Table 3 shows the details of the subparts (segment 2 through segment 6). Subparts 1 and 2 contain 2 bits of chromosome that determines the type of kernel function as linear, polynomial, a radial basis function (RBF), or sigmoid. The third subpart (5 bits) represents values of C (penalty factor), which lie between 0.1 and 100000. The fourth subpart (6 bits) determines the RBF kernel parameter, which is between 0.001 and 0.128. The fifth segment (2 bits) represents the value of polynomial kernel parameter (d), which can be 1, 2, 3, or 4. The sixth segment (2 bits) represents the value of the sigmoid kernel parameter, which can be 1, 2, 3, or 4.[43]

Step 4

In this step, the fitness values for all particles are calculated to determine the functionality of each particle; this is called validation of particles. The data are divided into training and evaluation parts using 10-fold cross-validation as input

Table 2: Parameters in particle swarm optimization genetic algorithm

PSO/GA parameters	ALL, MLL	ALL, AML	Colon	Breast
Population	20	20	15	15
Individual length	77	77	67	67
Number of features	60	60	50	50
Number of iteration	10	10	10	10
Inertia weight (w)	0.72	0.72	0.72	0.72
Acceleration constants	1.49	1.49	1.49	1.49
Crossing rate	0.9	0.9	0.9	0.9
Mutation rate	0.1	0.1	0.1	0.1

ALL – Acute lymphoblastic leukemia; AML – Acute myeloid leukemia; MLL – Mixed lineage leukemia; PSO – Particle swarm optimization; GA – Genetic algorithm

Table 3: A sample chromosome of particle swarm optimization genetic algorithm/fuzzy support vector machine population

Segment 1	Segment 2	Segment 3	Segment 4	Segment 5	Segment 6
Features	Kernel function type	Value of C parameter	RBF kernel parameter	Polynomial kernel parameter	Sigmoid kernel parameter
Number of features in Bit	2 bits	6 bits	5 bits	2 bits	2 bits

RBF – Radial basis function

for the cost function. This step is carried out for every particle to determine it as either a training or testing particle based on the selected features that exist in that particle.

The importance of each sample in the SVM training phase is examined. Standard SVM assumes that the training samples occur in pairs, such as (x_i, y_i) and $y_i \in (-1, +1)$ next, the importance of each sample is considered in each pair as (x_i, y_i, s_i) where s_i denotes the level of importance of each sample. The membership degree of sample X is assigned rather than its class, which can be achieved by a slight alteration of the main formula as:

$$\begin{aligned} &\textit{minimize} && \left[\frac{1}{2}w^T.w + C\sum_{i=1}^{N}s_i\varepsilon_i\right], \\ &\textit{subject to} && y_i\left(w^T x_i + b\right) \geq 1-\varepsilon_i, \\ & && \varepsilon_i \geq 0\ ,\ \textit{for}\ i = 1,2,\ldots N\ \text{and} \\ &\textit{maximize} && \left[-\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_j y_i y_j \vec{x}_i^T\vec{x}_j + \sum_{i=1}^{N}a_i\right], \\ &\textit{subject to} && 0 \leq a_i \leq s_i C, \\ & && \sum_{i=1}^{N}a_i y_j = 0 \end{aligned} \quad (4)$$

The difference between standard SVM and FSVM is the upper limit of the Lagrange multipliers; a_i in FSVM equals s_iC, while this value in SVM equals C. Next, the membership is computed for s_i of each sample x_i rather than the class.

Lin and Wang[44] obtained the value of s_i from the ratio of the distance of the sample from the center of the class to the distance of the farthest sample in same class from the center of the class. This method is sensitive to outliers and is not suitable for this kind of problem. The proposed method computes the weight and importance of each sample as:

$$s_i = \exp\left(-\frac{1}{2}(x_i-\mu)^T\sum^{-1}(x_i-\mu)\right) + \varepsilon \quad (5)$$

Where ε is a small value equal to 0.001 and μ, $\sum^{-1}$ are the mean vector and covariance matrix of the sample class, respectively. For simplicity and to decrease computation, the covariance matrix is assumed to be a diagonal matrix. Using this method and entering the extent of each sample in the training phase decrease the effect of outliers by multiplication of each sample weight in the sample error.[45]

Step 5

Update the best particle as $x^{g,best}$ and the best personal memories of each particle $x^{i,best}$ with the velocity and position of the particles as:

$$\begin{aligned} v_j^i[t+1] = {} & wv_j^i[t] + C_1r_1\left(x_j^{i,best}[t] - x_j^i[t]\right) + \\ & C_2r_2\left(x_j^{gbest}[t] - x_j^i[t]\right) \end{aligned} \quad (6)$$

In a binary algorithm, velocity is defined as a change in the means of probability and the velocity is explained by the probability of being in position 1.[46] Velocity is considered to be between 0 and 1, which explains the probability of being in position 1. The velocity is calculated using Eq. 8 and by mapping the values of 0 and 1 by limiting the sigmoid function. The final position of particle (i) is determined as:

$$x_j^i[t] = \begin{cases} 1 & \sigma < s\left(v_j^i[t]\right) \\ 0 & \textit{otherwise} \end{cases} \quad (7)$$

$$\textit{where}\ s(z) = \frac{1}{1+e^{-v_j^i}}$$

σ is a random number with uniform distribution in the range of 0 and 1.

To increase the velocity of divergence of the system, the limitation of velocity in the system must be considered based on maximum and minimum velocity. The roulette wheel approach has been used for selection in the proposed method. After the steps for parent selection, the steps executes the genetic operators commence. Single point, double point, and uniform crossover by random probability are used to benefit these crossover methods simultaneously.

Step 6

Again evaluate the amount of the fitness function.

Step 7

Combine the offspring and sort them based on the fitness value. Then, select the best parents using the elitism method and a defined population size.

Step 8

Go back to step 5 and repeat the steps until the termination condition is reached. The termination condition is the number of generations.

Step 9

When there is no further progress, the best features with the best parameters for the classifier have been selected. These

features and parameters can be applied to a blind test with no interference in the training and validation phases.

Step 10

Determine the occurrence frequency of each feature in the whole process. On average, biomarkers that have been repeated more than 6 times in the best locations are reported.

Step 11

The rules can be found using the best features extracted by the decision-tree algorithm. Figure 1 is a flowchart of the process. This flowchart summarizes how the system works and the relationships between the feature selection method and the classifier.

Results

The accuracy, sensitivity, precision, and specificity values were evaluated by applying the proposed algorithm to four public data sets. These values are statistical indicators for evaluation of binary classification. The goal is to find the best possible combination and compare this modified algorithm with others methods. Tables 4 and 5 show the result of application of the algorithm to the databases. Proportional to the number of samples in each database, 5–60 genes were selected and the hybrid algorithm was applied to them. All algorithms were implemented in MATLAB and LIBSVM software.

This section introduces the biomarkers obtained using the hybrid algorithm. The results indicate the good performance of algorithm for finding small subsets of features with high accuracy by decreasing the effect of outliers and noisy data and finding good similarity between these biomarkers and the biomarkers introduced by others in the literature.

Discussion and Analysis of Results

To investigate the accuracy of the proposed PSO/GA/FSVM hybrid algorithm, the results were examined in greater detail. Figure 2 shows the most frequent genes identified while running the algorithm with 10-fold cross-validation to determine which genes occurred more frequently in each database. Figure 2a shows the results for leukemia cancer types (ALL, AML), where 25 biomarkers were selected by

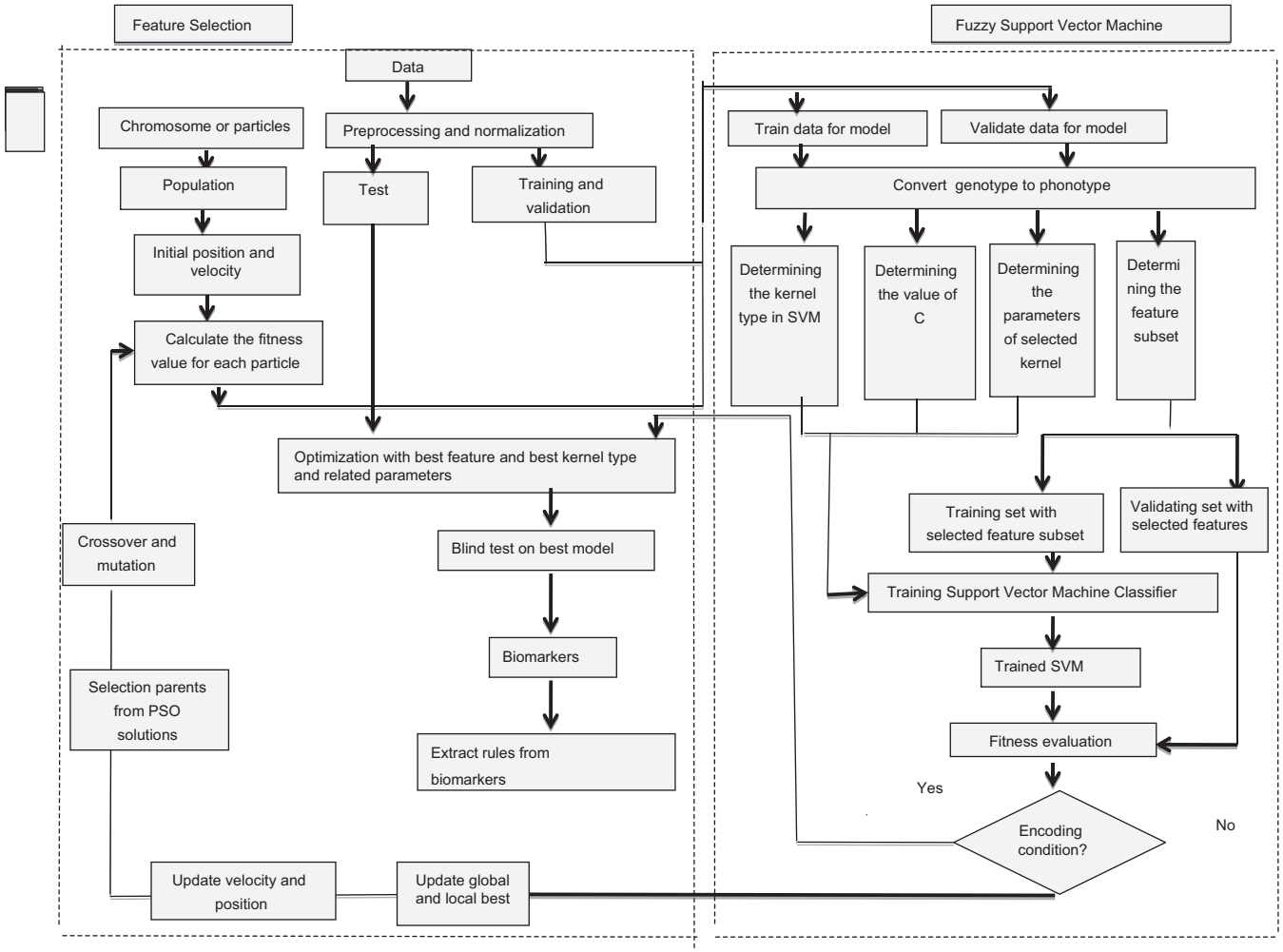

Figure 1: Hybrid algorithm flowchart (particle swarm optimization/genetic algorithm/fuzzy support vector machine)

Table 4: The results of applying hybrid (particle swarm optimization/genetic algorithm) to support vector machine classifier

Data sets	*t*-test/SVM					Best parameters		
	Methods	Accuracy	Sensitivity	Specificity	Precision	C	Kernel	Parameters
ALL/MLL	GA	4.52±98.57	100	97.50±7.91	97.50±7.91	100	RBF	0.0034
	PSO	100	100	100	100	100,000	RBF	0.0089
	PSO/GA	98.5714±4.51	100	98±6.324	96.67±10.54	100	RBF	0.012
ALL/AML	GA	6.900±95.714	90±16.10	100	100	1000	RBF	0.004
	PSO	6.900±95.714	100	95±8.05	85±24.15	100	Poly	Degree=3
	PSO/GA	98.5714±4.51	100	98±6.324	96.67±10.54	100	RBF	0.002
Colon	GA	91.6667±14.16	94±13.50	80±42.16	96.33±7.77	100	RBF	0.0098
	PSO	90±14.05	97.50±7.91	75±42.49	91.33±14.42	10	RBF	0.0139
	PSO/GA	95±8.05	96±8.34	90±31.67	98±5.27	10,000	RBF	0.0029
Breast	GA	9.664±94	85±24.15	100	100	10,000	RBF	0.0048
	PSO	13.499±94	100	90±2.250	91.67±18	10,000	Poly	Degree=4
	PSO/GA	98±6.324	100	95±15.81	97.50±7.91	1000	RBF	0.003

SVM – Support vector machine; ALL – Acute lymphoblastic leukemia; AML – Acute myeloid leukemia; MLL – Mixed lineage leukemia; PSO – Particle swarm optimization; GA – Genetic algorithm; RBF – Radial basis function

Table 5: The results of applying hybrid (particle swarm optimization/genetic algorithm) to fuzzy support vector machine classifier

Datasets	Methods	*t*-test/FSVM				Best parameters		
		Accuracy	Sensitivity	Specificity	Precision	C	Kernel	Parameters
ALL/MLL	GA	4.52±98.57	97.50±7.91	100	100	100,000	RBF	0.005
	PSO	98.57±4.52	100	90±31.62	98.57±4.52	10,000	RBF	0.0122
	PSO/GA	100	100	100	100	1000	Poly	Degree=4
ALL/AML	GA	98.57±4.52	100	98±6.34	96.67±10.54	1000	Poly	Degree=3
	PSO	95.71±6.90	90±16.10	100	100	10,000	RBF	0.0128
	PSO/GA	100	100	100	100	1000	RBF	0.0054
Colon	GA	95±8.05	100	90±16.10	92.50±12.08	100	RBF	0.0011
	PSO	95±8.05	8.43±96	90±31.62	98.33±5.27	1000	RBF	0.0024
	PSO/GA	96.67±7.03	97.50±7.91	95±15.81	98±6.324	100	RBF	0.0091
Breast	GA	98±6.324	100	96.67±10.54	96.67±10.54	10,000	Poly	Degree=2
	PSO	98±6.324	96.67±10.54	100	100	1000	Poly	Degree=2
	PSO/GA	98±6.324	100	97.50±7.91	95±15.81	10	RBF	0.002

ALL – Acute lymphoblastic leukemia; AML – Acute myeloid leukemia; MLL – Mixed lineage leukemia; PSO – Particle swarm optimization; GA – Genetic algorithm; RBF – Radial basis function; FSVM – Fuzzy support vector machine

the proposed hybrid algorithm. The most frequent genes selected comprised 19 genes for cancer types ALL and MLL in Figure 2b, 14 genes for colon cancer in Figure 2c and 18 genes for breast cancer in Figure 2d. All these genes repeated more than 6 times in the 10 runs of the algorithm.

A heat map was used to examine the biomarkers as a graphical representation of the changes in the behavior of the genes in the dataset. For example, it is desirable for the behavior of genes in cancer samples to be similar to one another and different from healthy samples. One group of genes may exhibit with low expression in normal samples and another group may exhibit high expression in normal samples. These genes can interact to aid in the accurate separation of cancer samples from normal samples.

Figure 3 shows heatmaps of two types of leukemia [Figure 3a and b], colon [Figure 3c] and breast cancer [Figure 3d]. The red denotes values above the mean, black denotes the mean, and green denotes values below the mean of a gene across all columns. The decision-tree algorithm was applied to biomarkers obtained using the proposed hybrid approach to find rules in common to them. Several criteria are specified to determine features or traits, including information gain, gain ratio, and the Gini index. The C5.0 decision-tree algorithm by SPSS Clementine 12 software[47] was employed.(4-SPSS clementine is a software package used for logical batched and non-batched statistical analysis which was acquired by IBM in 2009). Table 6 shows the rules discovered using the hybrid PSOGA/FSVM. Three rules with 93% accuracy were found using 10-fold cross-validation on the blood cancer types (ALL, MLL). Classification was performed using the u29175 and X95735 genes. Gene X95735 has high expression in

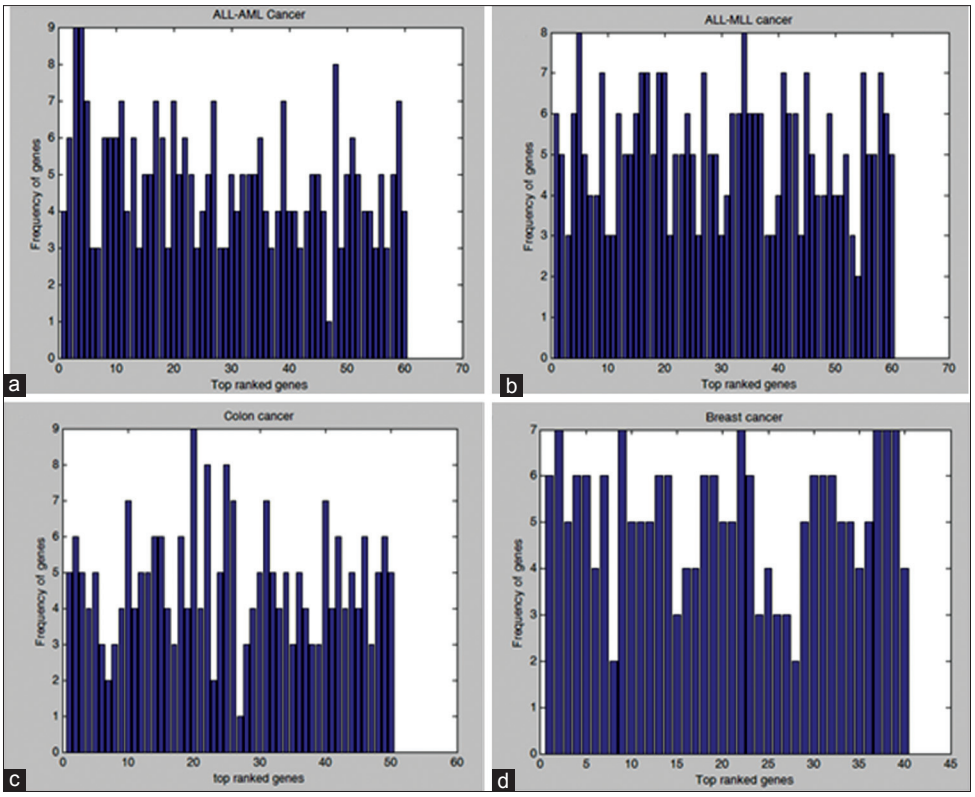

Figure 2: Occurrence frequency of genes by hybrid particle swarm optimization/genetic algorithm/fuzzy support vector machine algorithm with 10-fold cross validation. (a) Acute lymphoblastic leukemia, acute myeloid leukemia (b) acute lymphoblastic leukemia, mixed lineage leukemia (c) colon cancer (d) breast cancer

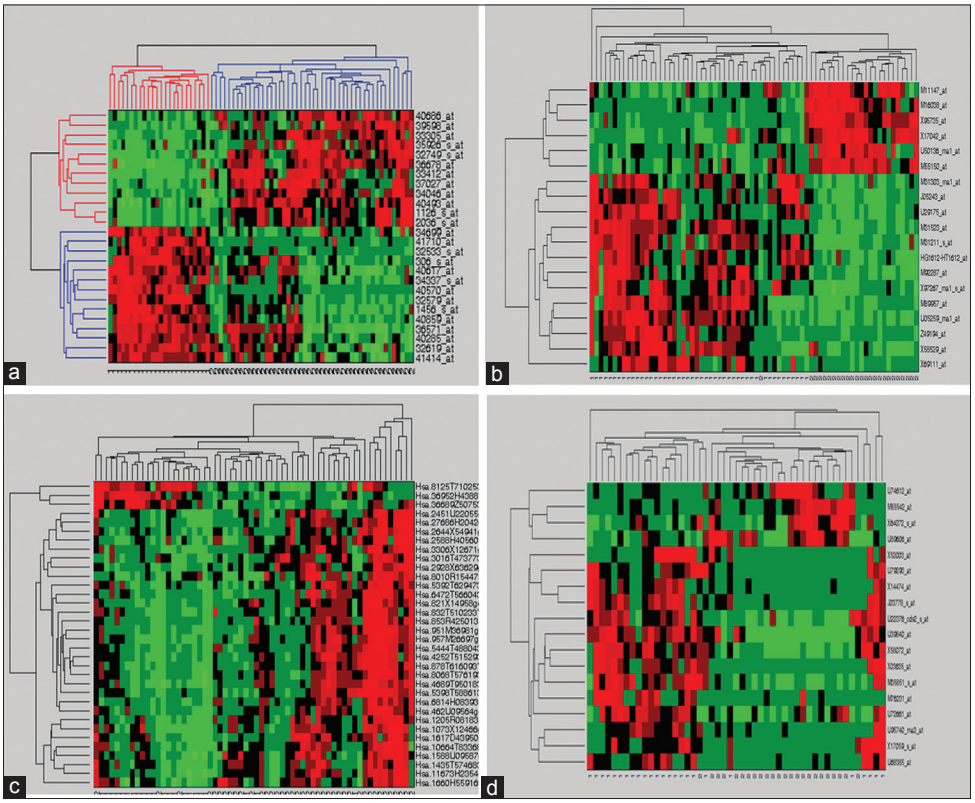

Figure 3: Heatmaps on 4 cancer data show the differences behavior of genes in 2 classes of data. (a-d) the result for leukemia cancer in types acute lymphoblastic leukemia and mixed lineage leukemia, acute lymphoblastic leukemia and acute myeloid leukemia, colon, and breast cancer data, respectively

Table 6: Extracted rules by decision tree on 4 cancer database

Databases	Rules with decision tree
Leukemia cancer rules	If gene u29175_at > - 0.551 then ALL
	If gene X95735_at ≤ - 0.587 then ALL
	If gene X95735_at > - 0.587 and gene u29175 at ≤ - 0.551 then AML
Blood cancer rules	If gene 32533_s_at > - 0.249 then ALL
	If gene 40570_at > - 0.461 then ALL
	If gene 32533_s at ≤ - 0.249 and gene 40570 at ≤ - 0.461 then MLL
Breast cancer rules	If X03635_at > - 0.950 and X64072 s at ≤ - 0.293 then ER+
	If X03635_at ≤ - 0.950 then ER−
	If X64072_s_at > - 0.293 then ER−
Colon cancer rules	If T62947 ≤ - 0.874 then normal
	If M76378 > - 0.303 then normal
	If T62947 > - 0.874 and M76378 ≤ - 0.303 then tumor

ALL – Acute lymphoblastic leukemia; AML – Acute myeloid leukemia; MLL – Mixed lineage leukemia; ER – Estrogen receptor

Table 7: Summarizes results and comparison with literatures

Datasets methods	Accuracy (%)			
	ALL/MLL	ALL/AML	Colon	Breast
Hernandez Montiel *et al*., 2011[11]	-	98.61	98.38	-
Li *et al*., 2008[27]	-	95.1	88.7	93.40
Shen *et al*., 2008[16]	-	95.81	90.31	93.50
Shen *et al*., 2007[9]	-	-	90.43	-
Abdi *et al*., 2012[28]	-	100	93	-
Moteghaed *et al*., 2015[48]	-	100	96.67	96
Presented PSO/GA/SVM	98.57	98.57	95	98
Presented PSO/GA/FSVM	100	100	96.6667	98

ALL – Acute lymphoblastic leukemia; AML – Acute myeloid leukemia; MLL – Mixed lineage leukemia; PSO – Particle swarm optimization; GA – Genetic algorithm; FSVM – Fuzzy support vector machine

AML samples; gene u29175 has low expression in this cancer type (AML). The table also shows the rules for the other databases for blood cancer types ALL and MLL, breast cancer, and colon cancer. The classification accuracy for the cancer data and for highly ranked genes was 93%, 89%, and 80%, respectively.

Comparisons were made between the proposed algorithm and other algorithms. Table 7 shows the results of the comparison based on classification accuracy.

The paper[48] by the same authors employed a multilayer perceptron (MLP) for the classification. However, the running procedure of algorithm takes more time than the FSVM and SVM classifier. One of the advantages of the FSVM classifier is its high speed in running procedure. In MLP classifier, all the samples have the same weight in training phase; but in second paper, we use FSVM as the classifier and the importance of each sample in training phase.

The extracted biomarkers from the proposed algorithm and those reported in other studies which were present in Tables 8 and 9 were also compared. For blood cancer types ALL and MLL, the proposed algorithm found 24 biomarkers; 11 were the same as biomarkers from Armstrong. The biomarkers that were common for blood cancer were 36678, 34699, 33305, 32579, 41710, 32533, 33412, 32749, 37027, 2036, and 40570. For the ALL and AML cancers, the algorithm found 19 biomarkers, 11 of which were the same as those presented by Golub *et al*. These common biomarkers were X17042, U50136, X95735, M55150, M92287, U29175, M31211, M16038, U05259, M31303, and M31523.

For breast and colon cancer, 7 out of 18 biomarkers were in common with the results presented by West for breast cancer and 3 out of 14 were in common for colon cancer with the results presented by Alon *et al*. The biomarkers in common for breast cancer were M35851, X52003, X58072, X14474, U95740, U68385, and U22376. The biomarkers in common for colon cancer were T57619, T58861, and X55715.

Conclusions

The results of the present study provide a comprehensive comparison of the proposed algorithm and those from previously published sources. The proposed algorithm is a hybrid of PSO and GA with FSVM classifier. This classifier has the ability to enter the importance of each sample into training of the system for further prediction without the need for trial-and-error to determine classifier parameters. Good results have been demonstrated for the proposed

Table 8: Discovered biomarkers for leukemia and blood cancer (acute lymphoblastic leukemia, acute myeloid leukemia, mixed lineage leukemia)

Blood cancer (ALL/MLL)		Blood cancer (ALL/AML)	
Gene number	**Description**	**Gene ID**	**Description**
40,285	Hs. 58927 *H. sapiens* nuclear VCP-like protein	X58529	Immunoglobulin heavy constant gamma 1
36,678	Hs. 75725 human mRNA for KIAA0120 gene	X69111	Inhibitor of DNA binding 3, dominant negative helix
40,570	Hs. 170133 *H. sapiens* forkhead protein	Z49194	POU class 2 associating factor 1
2036	Hs. 169610 human cell adhesion molecule (CD44)	X17042	
33,305	Hs. 183583 human monocyte, M93056	X97267	Protein tyrosine phosphatase, receptor type, C-associated
36,571	Hs. 75248 *H. sapiens* Top IIb mRNA	U50136	Leukotriene C4 synthase
39,598	Hs. 2679 human mRNA for gap junction protein	M11147	Similar to ferritin, light polypeptide; ferritin, light
32,579	Hs. 78202 human transcriptional activator (BRG1)	X95735	Zyxin
306	Hs. 251064 human nonhistone chromosomal	M55150	Fumarylacetoacetate hydrolase (fumarylacetoacetase)
34,337	Hs. 31016 *H. sapiens* mRNA for M96A	M92287	Cyclin D3
41,710	Hs. 12969 *H. sapiens* mRNA full length insert	U29175	SWI/SNF related, matrix associated, actin dependent
32,533	Hs. 74669 *H. sapiens* VAMP5 mRNA	M31211	Myosin, light chain 6B, alkali
32,749	Hs. 195464 *H. sapiens* mRNA; c DNA	M16038	V-yes-1 Yamaguchi sarcoma viral related oncogene
33,412	Hs. 227751 vicpro2.D07.r H. sapiens c DNA, 5’	U05259	CD79a molecule, immunoglobulin-associated alpha
1126	Hs. 169610 human hyaluronate receptor (CD44)	M31303	Stathmin 1
1456	Hs. 155530 human interferon-gamma induced	M31523	Transcription factor 3 (E2A immunoglobulin enhancer)
40,617	Hs. 157426 *H. sapiens* chromosome 16 BAC	M89957	CD79b molecule, immunoglobulin-associated beta
40,493	Hs. 169610 human hyaluronate receptor (CD44)	J05243	Spectrin, alpha, nonerythrocytic 1 (alpha-fodrin)
40,859	Hs. 173684 tq27a01.x1 *H. sapiens* c DNA, 3’	X58529	Immunoglobulin heavy constant gamma 1 (G1m marker)
37,027	Hs. 76549 human novel protein AHNAK mRNA		
35,926	Hs. 204040 AF009007-leukocyte immunoglobulin		
32,619	Hs. 81073 yv22a08.r1 *H. sapiens* c DNA, 5’ end		
40,686	Hs. 83321 ws06b05.x1 *H. sapiens* c DNA, 3’		
34,046	Hs. 40342 human DNA sequence		
41,414	Hs. 106730 novel human gene		
34,699	Hs. 265561 *H. sapiens* mRNA		

ALL – Acute lymphoblastic leukemia; AML – Acute myeloid leukemia; MLL – Mixed lineage leukemia; *H. sapiens* – *Homo sapiens*

Table 9: Discovered biomarkers for colon and breast cancer by particle swarm optimization/genetic algorithm/fuzzy support vector machine

Breast cancer		Colon cancer	
Gene ID	**Description**	**Gene ID**	**Description**
M35851	Androgen receptor		
X52003	Trefoil factor 1	T62947	60S ribosomal protein L24
M55542	Guanylate-binding protein 1, interferon-inducible	T95018	120032 40S ribosomal protein S18
J03778	Microtubule-associated protein tau	T70062	Human nuclear factor NF45 mRNA, complete cds
X17059	N-acetyltransferase 1 (arylamine N-acetyltransferase)	T57619	40S ribosomal protein S6 (Nicotiana tabacum)
X58072	GATA-binding protein 3	T58861	60S ribosomal protein L30E
U72661	Ninjurin 1	X55715	Human Hums3 mRNA for 40S ribosomal protein s3
X14474	Microtubule-associated protein tau	T57468	Fibrillarin (human)

Contd...

Table 9: Contd...

Breast cancer		Colon cancer	
Gene ID	**Description**	**Gene ID**	**Description**
U39840	Forkhead box A1	X63629	*H. sapiens* mRNA for P cadherin
M76231	Sepiapterin reductase	M76378	Human CRP gene
U89606	Pyridoxal (pyridoxine, vitamin B6) kinase	H55758	Alpha enolase (human)
U74612	Forkhead box M1	R08183	Q04984 10 KD heat shock protein
U95740	KIAA0430	H55916	Peptidyl-prolyl cis-trans isomerase
U95740	Chromosome 16 open reading frame 45	U22055	Human 100 kDa coactivator mRNA, complete cds
U68385	Meis homeobox 3 pseudogene 1	T86749	Human (clone PSK-J3) cyclin-dependent protein
X64072	Integrin, beta 2		
U22376	V-myb myeloblastosis viral oncogene homolog (avian)		
X03635	ER 1		

ER – Estrogen receptor; CRP – Cysteine-rich protein; *H. sapiens* – *Homo sapiens*

algorithm. The classification accuracy for leukemia data is 100%, for colon cancer is 96.67%, and for breast cancer is 98%. These results are better than the others works because the algorithm can determine the training parameters and small feature subsets in the databases perfectly with no user interface. The results show that the best kernel used in training the SVM classifier is the RBF.

Financial support and sponsorship

None.

Conflicts of interest

There are no conflicts of interest.

References

1. Schena M, Shalon D, Davis RW, Brown PO. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. Science 1995;270:467-70.
2. Schena M, Shalon D, Heller R, Chai A, Brown PO, Davis RW, *et al.* Parallel human genome analysis: Microarray-based expression monitoring of 1000 genes. Proc Natl Acad Sci U S A 1996;93:10614-9.
3. Dehnavi AM, Sehhati MR, Rabbani H. Hybrid method for prediction of metastasis in breast cancer patients using gene expression signals. J Med Signals Sens 2013;3:79-86.
4. Sehhati MR, Dehnavi AM, Rabbani H, Javanmard SH. Using protein interaction database and support vector machines to improve gene signatures for prediction of breast cancer recurrence. J Med Signals Sens 2013;3:87-93.
5. Sehhati M, Mehridehnavi A, Rabbani H, Pourhossein M. Stable gene signature selection for prediction of breast cancer recurrence using joint mutual information. IEEE/ACM Trans Comput Biol Bioinform 2015;12:1440-8.
6. Chuang LY, Yang CH, Wu KC, Yang CH. A hybrid feature selection method for DNA microarray data. Comput Biol Med 2011;41:228-37.
7. Chuang LY, Yang CS, Wu KC, Yang CH. Gene selection and classification using Taguchi chaotic binary particle swarm optimization. Expert Syst Appl 2011;38:13367-77.
8. Shen Q, Mei Z, Ye BX. Simultaneous genes and training samples selection by modified particle swarm optimization for gene expression data classification. Comput Biol Med 2009;39:646-9.
9. Shen Q, Shi WM, Kong W, Ye BX. A combination of modified particle swarm optimization algorithm and support vector machine for gene selection and tumor classification. Talanta 2007;71:1679-83.
10. Li L, Jiang W, Li X, Moser KL, Guo Z, Du L, *et al.* A robust hybrid between genetic algorithm and support vector machine for extracting an optimal feature gene subset. Genomics 2005;85:16-23.
11. Hernandez Montiel LA, Huerta EB, Caporal RM. A multiple-filter-GA-SVM method for dimension reduction and classification of DNA-microarray data. Rev Mex Ing Biomed 2011;32:32-9.
12. Tong DL, Schierz AC. Hybrid genetic algorithm-neural network: Feature extraction for unpreprocessed microarray data. Artif Intell Med 2011;53:47-56.
13. Li L, Weinberg CR, Darden TA, Pedersen LG. Gene selection for sample classification based on gene expression data: Study of sensitivity to choice of parameters of the GA/KNN method. Bioinformatics 2001;17:1131-42.
14. Yang CH, Chuang LY, Yang CH. A hybrid filter/wrapper method for feature selection of microarray data. J Med Biol Eng 2009;30:23-8.
15. Chuang LY, Chang HW, Tu CJ, Yang CH. Improved binary PSO for feature selection using gene expression data. Comput Biol Chem 2008;32:29-37.
16. Shen Q, Shi WM, Kong W. Hybrid particle swarm optimization and tabu search approach for selecting genes for tumor classification using gene expression data. Comput Biol Chem 2008;32:52-9.
17. Martinez E, Alvarez MM, Trevino V. Compact cancer biomarkers discovery using a swarm intelligence feature selection algorithm. Comput Biol Chem 2010;34:244-50.
18. Lee CP, Leu Y. A novel hybrid feature selection method for microarray data analysis. Appl Soft Comput 2011;11:208-13.
19. Alba E, Garcıa-Nieto J, Jourdan L, Talbi EG. Gene selection in cancer classification using PSO/SVM and GA/SVM Hybrid Algorithms. In: IEEE 2007 Congress on Evolutionary Computation. Singapore: IEEE; 2007. p. 284-90.
20. Zhenyu C, Jianping L, Liwei W, Weixuan Xu, Yong SH. Multiple-kernel SVM based multiple-task oriented data mining system for gene expression data analysis. Expert Syst Appl 2011;38:12151-9.
21. Wang X, Simon R. Microarray-based cancer prediction using single genes. BMC Bioinformatics 2011;12:391.
22. Shah S, Kusiak A. Cancer gene search with data-mining and genetic algorithms. Comput Biol Med 2007;37:251-61.
23. Chuang LY, Yang CH, Jin LC. Classification of multiple cancer

types using fuzzy support vector machines and outlier detection methods. Biomed Eng Appl Basis Commun 2005;17:300-8.
24. Mao Y, Zhou X, Pi D, Sun Y, Wong ST. Multiclass cancer classification by using fuzzy support vector machine and binary decision tree with gene selection. J Biomed Biotechnol 2005;2005:160-71.
25. Ng M, Chan L. Informative gene discovery for cancer classification from microarray expression data. In: IEEE 2005 Workshop on Machine Learning for Signal Processing. Mystic, CT: IEEE; 2005. p. 393-8.
26. Yeh JY, Wu TS, Wu MC, Chang DM. Applying data mining techniques for cancer classification from gene expression data. In: IEEE 2007 International Conference on Convergence Information Technology. Gyeongju: IEEE; 2008. p. 703-8.
27. Li S, Wu X, Tan M. Gene selection using hybrid particle swarm optimization and genetic algorithm. Soft Comput 2008;12:1039-48.
28. 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;2012:320698.
29. Chu F, Wang L. Applications of support vector machines to cancer classification with microarray data. Int J Neural Syst 2005;15:475-84.
30. Armstrong SA, Staunton JE, Silverman LB, Pieters R, den Boer ML, Minden MD, *et al.* MLL translocations specify a distinct gene expression profile that distinguishes a unique leukemia. Nat Genet 2002;30:41-7.
31. Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, Mesirov JP, *et al.* Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. Science 1999;286:531-7.
32. Alon U, Barkai N, Notterman DA, Gish K, Ybarra S, Mack D, *et al.* Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. Proc Natl Acad Sci U S A 1999;96:6745-50.
33. West M, Blanchette C, Dressman H, Huang E, Ishida S, Spang R, *et al.* Predicting the clinical status of human breast cancer by using gene expression profiles. Proc Natl Acad Sci U S A 2001;98:11462-7.
34. Boyd R, Richerson PJ. Culture and the Evolutionary Process. USA: The University of Chicago; 1985.
35. Eberhart R, Kennedy J. A new optimizer using particle swarm theory. In: IEEE 1995 Proceeding of 6th International Symposium on Micro Machine and Human Science. Nagoya: IEEE; 1995. p. 39-43.
36. Cortes C, Vapnik V. Support vector networks. Mach Learn 1995;20:273-9.
37. Burges CJ. A tutorial on support vector machines for pattern recognition. Data Min Knowl Discov 1998;2:121-67.
38. Han J, Kamber M, Pei J. Data Mining, Concept and Techniques. 3rd ed. Waltham, MA, USA: Elsevier Morgan Kaufmann; 2011.
39. Kao YT, Zahara E. A hybrid genetic algorithm and particle swarm optimization for multimodal functions. Appl Soft Comput 2008;8:849-57.
40. Du S, Li W, Cao K. A learning algorithm of artificial neural network based on GA–PSO. In: IEEE 2006 Proceedings of the 6th World Congress on Intelligent Control and Automation. Dalian, China: IEEE; 2006. p. 3633-7.
41. Juang CF. A hybrid of genetic algorithm and particle swarm optimization for recurrent network design. In: IEEE Transactions on Systems, Man and Cybernetics Part B (34). 2004 IEEE International Joint Conference on Neural Networks; 2004. p. 997-1006.
42. Robinson J, Sinton S, Yahya RS. Particle swarm, genetic algorithm, and their hybrids. In: IEEE 2002 Antennas and Propagation Society International Symposium. San Antonio: IEEE; 2002. p. 314-7.
43. Avci E. Selecting of the optimal feature subset and kernel parameters in digital modulation classification by using hybrid genetic algorithm–support vector machines: HGASVM. Expert Syst Appl 2009;36:1391-402.
44. Lin CF, Wang SD. Fuzzy support vector machines. IEEE Trans Neural Netw 2002;13:464-71.
45. Kaboodiyan J, Moradi MH. Yek Machine Bordare Poshtibane Fyzzy Jadid ba Fuzzy Sazi dar do Marhaleh. Proc 12th Conference of Electrical Engineering. Ferdosi Mashhad University, Iran; 2004.
46. Kennedy J, Eberhart RC. A discrete binary version of the particle swarm algorithm. In: IEEE 1997 International Conference on Computational Cybernetics and Simulation. Orlando, FL: IEEE; 1997. p. 4104-9.
47. Quinlan JR. C4.5: Programs for Machine Learning. San Francisco, CA: Morgan Kaufmann Publishers Inc.; 1993.
48. Moteghaed NY, Maghooli K, Pirhadi S, Garshasbi M. Biomarker discovery based on hybrid optimization algorithm and artificial neural networks on microarray data for cancer classification. J Med Signals Sens 2015;5:88-96.