Biomarker Discovery Based on Hybrid Optimization Algorithm and Artificial Neural Networks on Microarray Data for Cancer Classification

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

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

Submission: 25-07-2014 Accepted: 03-02-2015

ABSTRACT

The improvement of high-through-put gene profiling based microarrays technology has provided monitoring the expression value of thousands of genes simultaneously. Detailed examination of changes in expression levels of genes can help physicians to have efficient diagnosing, classification of tumors and cancer's types as well as effective treatments. Finding genes that can classify the group of cancers correctly based on hybrid optimization algorithms is the main purpose of this paper. In this paper, a hybrid particle swarm optimization and genetic algorithm method are used for gene selection and also artificial neural network (ANN) is adopted as the classifier. In this work, we have improved the ability of the algorithm for the classification problem by finding small group of biomarkers and also best parameters of the classifier. The proposed approach is tested on three benchmark gene expression data sets: Blood (acute myeloid leukemia, acute lymphoblastic leukemia), colon and breast datasets. We used 10-fold cross-validation to achieve accuracy and also decision tree algorithm to find the relation between the biomarkers for biological point of view. To test the ability of the trained ANN models to categorize the cancers, we analyzed additional blinded samples that were not previously used for the training procedure. Experimental results show that the proposed method can reduce the dimension of the data set and confirm the most informative gene subset and improve classification accuracy with best parameters based on datasets.

Key words: *Artificial neural network, cancer classification, gene expression, genetic algorithm, particle swarm optimization algorithm*

INTRODUCTION

The DNA microarray technology has provided monitoring of thousands of genes simultaneously in a single experiment. However, gene expression data have some characteristics which cause difficulty in analyzing data with many classifiers such as high-dimension - often exceeds more than ten of thousands - in contrast of small-sample size - often a few hundred samples and high-noise nature of data. Hence, the main challenge is to find a small subset of relevant genes to improve classification accuracy with robustness. Using this technology and check-outs the changes in expression levels of genes between samples, can help physicians to have efficient diagnosing as well as effective treatments (Schena, 1996),[1] (Schena *et al*. 1995),[2] Study (Dong Ling Tong, 2011),[3] developed a hybrid genetic algorithm (GA) - neural network model for feature selection on unpreprocessed microarray data. The fitness value GA is based on an accuracy of standard feed-forward artificial neural network (ANN). The main point of the genetic algorithm-neural network algorithm is to select highly informative genes by the calculation of the both GA fitness function and the ANN weights simultaneously In (Li-Yeh Chuang, 2011),[4] Taguchi-GA and correlation-based feature selection used as a hybrid methods, and the K-nearest neighbor (K-NN) served as a classifier and also in paper (Li-Yeh Chuang *et al*., 2011)[5] another study based on Taguchi binary particle swarm optimization (PSO) conducted by the same authors. In paper (Bing Liu, 2004),[6] a combinational feature selection method with ensemble neural networks was used for classification.

Rank sum test, principal components analysis (PCA), clustering, and *t*-test are used to extract and select features. In this work, bootstrap technique is used to resample data, and also cooperative and competitive neural networks are tested on data and create the output. In paper

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

(Shen Qi, 2007),[7] the combination of modified discrete PSO and support vector machines (SVM) for tumor classification is applied to select genes with the ability of high accuracy classification. In paper (Li-Yeh Chuang, 2008),[8] improved binary PSO is in order to feature selection, and the K-NN method serves as a classifier for gene expression data classification problems. In (Shen Qi, 2008),[9] a hybrid PSO and tabu search with linear discriminant analysis (LDA) classification were developed for gene selection and cancer classification. Paper (Emmanuel Martineza, 2010)[10] proposed an algorithm based on swarm intelligence feature selection method in which, the initialization and update of only a subset of particles are happened in the swarm. The most frequent genes are evaluated by the GA/SVM again to obtain the most final relevant gene subset. In (Leping li, 2001)[11] GA and the K-NN are combined to identify most frequent gene for cancer classification. In (Yang, 2009),[12] a hybrid method based on information gain and GAs are proposed for gene selection in microarray data sets. The K-NN method with leave-one-out cross validation served as a classifier for evaluating the fitness function of this hybrids algorithm. In study (Jenny Önskog, 2011),[13] classification performance of five normalization methods and three gene selection methods as *t*-test, relief, paired distance, and eight machine learning methods as a decision tree with Gini index and information gain criterion, SVM classifier with different kernels and also neural network are compare with each other.

In paper (Xiaosheng Wang, 2011),[14] use single genes to create classification models and identified the most powerful genes for class discrimination. By these kinds of classifiers, include diagonal LDA, K-NN, support vector machine and random forest. Then they constructed simple rules for cancer prediction by these single genes. In (Shital Shah, 2007),[15] an integrated algorithm involves a GA and correlation-based heuristics for data preprocessing and decision tree, and SVM algorithms are used for making predictions. Paper (Jinn-Yi Yeh, 2007),[16] applies GAs with an initial solution provided by *t*-statistics for selecting a subset of genes and the decision tree is used as a classifier to build model on top of these selected genes. In study (Chu, 2005),[17] feature selection methods, such as PCA, class separability measure, Fisher ratio, and *t*-test are used for gene selection. And a voting scheme is then applied to do multi-group classification by binary SVM. In study (Makoto Takahashi a, 2010),[18] an unpaired *t*-tests with one of the supervised classifiers, ANNs was applied to schizophrenia gene expression data sets. Study (Khan javed, June 2001),[19] applied a method for classifying cancers using ANNs on small, round blue cell tumors as a model. *T*-test and PCA are used to reduction dimensionality of data sets. In (Nikhil R Pal, 2007),[20] a multilayer networks with online gene selection ability and relational fuzzy clustering was used to identify a small set of biomarkers for accurate classification.

In our paper, we use the hybrid of GA and PSO algorithm as a feature selection method and the fitness of each gene subset (chromosome) is determined by ANN classifier's accuracy. The 10-fold cross validation classification accuracy on the gene subset in the training and evaluation samples is evaluation criteria. The group of gene subset with the highest 10-CV classification accuracy is considered as the optimal gene subset. After we have selected the most frequent genes, we can use them for discrimination of blind test data to see the response of evaluation hybrid system on these kinds of data. At last, we use a decision tree classifier to see the relation between founded biomarkers and rule extraction. This point should be considered that one of our purposes is increasing accuracy of classification problems by selecting the best parameters of the classifier without using any trial and errors of users. The classifier parameters in the training and testing phase. Hence, using a suitable combination of optimization algorithms for feature selection and also selecting proper classifier can improve the classification results.

MATERIALS AND METHODS

In this section, we introduce the gene expression datasets which were used in this paper and also propose the modified hybrid algorithm. Three datasets are used to test our proposed algorithm. The first data include 72 samples in two type of classes as acute lymphoblastic leukemia (ALL) and acute myeloid leukemia (AML). The original size of genes in this dataset is 7129. These two categories of cancer are quite similar at the microscopic level and have a same behavior over the years. This dataset is generated by Golub in (Golub T, 1999).[21] The second are generated by (Alon U, 1999)[22] for colon cancer categories. These data have 22 samples for normal class and 40 samples for tumor class. The size of genes in this dataset is 2000. The last data include 49 samples in two class of breast cancers 25 samples are placed in estrogen receptor (ER+) class, and 24 of them are placed in ER − class. The original size of features or genes in these data is 7129 and was generated by (West, 2001).[23] Table 1 shows the summary of the data which are used in this paper.

In the following discussion, we introduce the proposed algorithm which is used on gene expression profiles. GA and PSO are two optimization algorithms which have

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

Dataset	Tissue	Sample	Number of class	Sample per class	Classes	Number of genes
Golub-2002	Leukemia	72	2	47, 25	ALL, AML	7129
Alon-1999	Colon	62	2	22, 40	Normal, tumor	2000
West-2001	Breast	49	2	25, 24	ER+, ER−	7129

ALL – Acute lymphoblastic leukemia; AML – Acute myeloid leukemia; ER – Estrogen receptor many advantages in these kinds of problems. They are computational optimization method that search all part of the solution space with a different kind of solution or a group of feature subsets to find the best answer in each iteration. In GA, the searching process only needs to determine the value of the objective function at different points and also, no additional information like differentiation of function is needed. The most important operators in GA are crossover and mutation that create variety solutions. PSO algorithm was developed by Kennedy and Eberhart in 1995 (Eberhart R, 1995).[24] In PSO, each particle moves in the search space with a velocity adjusted by its own memory and its neighbors to find the best solutions. The main difference is that there are no crossover and mutation operators in PSO. Hence, it is more likely to be caught in a local minimum.

But the best particles in PSO can be remembered which affect the other particles. Hence, this property of the algorithm can lead to faster convergence.

In contrast to PSO algorithm, chromosomes in GA algorithm share the information between each other.

ANN is an information processing system that got its idea from human brain. It performs data processing by providing small processor that are parallel interconnected with each other to form a network to solve a problem. Neural networks are used to implement complex functions in various fields, including pattern recognition, identification, classification, speech and image processing, and control systems. After tuning or training the neural network, each particular input has a particular response. A neural network consists of components as layers and weights. Network behavior is related to the connections between its members. In general, the neural network has three layers of neurons such as an input layer, hidden layers, and output layer.

The input layer receives raw data and feature vectors. Performance of the hidden layers is determined by inputs and weighted vectors between input and hidden layers. Weights between input and hidden units have to be determined when a hidden unit is been active. Performance of the output layers depends on the weights between the hidden and output units. In multi-layer perceptron networks or feed-forward networks, each layer may be determined by it's parameter matrices and the network can form by a combination of nonlinear operators. The goal is finding and estimation of the mapping function between input and output spaces. Estimation of suitable network is based on a minimization of the error between the desired output and network's output. In each layer, activation functions can be nonlinear in both layers and also can be different from each other. In these networks, there are two types of weight matrices, such as an intermediate layer or hidden layers and output layer weight matrix. These matrixes' sizes depend on the number of neurons in hidden layers and output layer's neurons. So how the network works is as follows

$$u\left(n\right)=W^{h}\times x\left(n\right),\ h\left(n\right)=\varnothing\ (u(n))$$

$$v\left(n\right)=W^{y}\times h\left(n\right),\ \ y\left(n\right)=\varphi\ (v(n))$$

In summary, we can write:

$$y\ (n)=\varphi\left(W^{y}\times\varnothing\left(W^{h}\times x\left(n\right)\right)\right)$$

Training neural networks mean selecting the best model of network by the best parameters such as weights, number of neurons based on the cost function. The task of pattern classification in ANN is to assign an input pattern as gene expression profile represented by feature vector to one of the introduced classes such as normal or cancer. After providing the best network based on feature vectors and parameters, our model can be able to predict the class of new data based on training.

Proposed Algorithm

General description of the GA and PSO is presented in the previous part. Now, in this section, we give a detailed description of our proposed algorithm. We can implement both of these algorithms in hybrid form to benefit the useful advantages of both of them and covered their problems. In this paper, ANN is used as a classifier and fitness function of hybrid PSO/GA algorithm (Kao, 2008),[25] (Du, 2006),[26] (Juang, 2004),[27] (Robinson, 2002).[28]

At first of the implementation, we have to preparation of data such as, filtering and normalization stage.

The integration of data includes whole genomes of human, so most of the genes in the database are not useful and irrelevant for classification problems. These genes are considered as noisy data and can produce difficulty in classification problems. We have eliminate genes that (1) Their expression value is very low, (2) have little change in expression value in hole samples, (3) genes that have a low standard deviation and have no impressive changes around the mean of expression value, (4) genes that have low information entropy. Next, we have select top ranked genes by, *t*-test method and apply them as an input of hybrid PSO/GA system. We also try to divide data into two parts.

In the following, 10% of data must belong randomly as a blind test data, and also remaining 90% of data can be entering to training and evaluation phase of the algorithm by 10-fold cross validation. The value of parameters such as, size of population, length of chromosome and particles, rate of mutation and crossover in GA, inertia coefficient (W), training factors (learning factors), and maximum velocity is mentioned in Table 2.

In addition of creating initial position (X_{id}), initial velocity (V_{id}) of every particle should be determined randomly in the population. This stage is related to making the initial population, at first the population with N chromosome create randomly. The length of particles or chromosomes can be explained as, adding number of features which has been selected based on statistical method and 11 additional genes which have been used for determination of optimum parameters of classifier by hybrid algorithm. Primary random and binary initialization are taken place first, in such a way that 1 shows the existence of the feature in training system and 0 is meaning of not existing of that feature. Now, each chromosome is a word of bits in two main parts. First part is equal to feature dimensions size (segment 1), and the second part is used for determining and designing classifier parameters. The second part contains three sub-parts, which can be seen with details in Table 3. The second segment of a chromosome (one bit of chromosome) determines the number of layers in the network. The third and fourth segments show the number of neuron in each hidden layers. We assign 5 bit for each layer which converted to the decimal number during training. Table 3 shows a sample of chromosome in the population. The fitness values for all particles have to be calculated in order to determine functionality of each particle, which is so-called validation of particles.

The velocity and position of the particles have been updated based on equation below:

$$v_j^i\left[t+1\right]=wv_j^i\left[t\right]+C_1r_1\left(x_j^{i,best}\left[t\right]-x_j^i\left[t\right]\right)+C_2r_2\left(x_j^{g,best}\left[t\right]-x_j^i\left[t\right]\right)$$

Table 2: Parameters in PSOGA

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

ALL – Acute lymphoblastic leukemia; AML – Acute myeloid leukemia; PSOGA – Particle swarm optimization and genetic algorithm

Table 3: A sample chromosome of PSOGA/ANN population

Segment 1	Segment 2	Segment 3	Segment 4
Features	Number of hidden layer	Number of neurons in first layer	Number of neurons in second layer
110101011....10	1 or 0	10...01	11...01
Number of features bit	1 bits	5 bits	5 bits

PSOGA – Particle swarm optimization and genetic algorithm; ANN – Artificial neural network

The best particle as $x^{g,\ best}$ and the best personal memories of each particle as $x^{i,\ best}$ is updating. In this paper, a binary PSO algorithm is used by the authors (Kennedy, 1997).[29]

It is important to note that in genetic operators, there is no discussion in speed changes or the best memory of offspring; hence, we have determined the best memory of offspring based on the best memory of parents which have the best fitness value. After this step, this is the time for running GA, from the solutions which are presented by the PSO, the crossover and mutation operators are applied on selected parents. In this paper, we have used roulette-wheel as a selection method. Roulette-wheel is a technique which selects parents based on the fitness value on each of them. Since the algorithm is binary, we use bit inversion (set zero to one and vice versa) method for the mutation operator. In this paper, we are using three crossover methods such as a single point, double point, and uniform crossover by a random probability to be able to use all benefits and advantages of these crossover methods simultaneously. At the end of the progress, the best features with the best parameters of classifier are selected, so we have applied these features and parameters to blind test that has no interference in the training and validation phase at all. Determine the occurrence frequency of each feature in the whole process. On average, biomarkers that have been repeated >6 times in the best locations are reported. Finally, the decision tree's rules can be found from the best-extracted features. The whole work is presented in the following flowchart in Figure 1. This flowchart shows a summary of how the system works and the relation between feature selection method and classifier.

RESULT

By applying the proposed algorithm to 3 cancer databases, the amounts of accuracy, sensitivity, precision, specificity were computed. These values are statistical indicators for the evaluation of a binary classification. Our goal is to find the best possible combination and comparison of this modified algorithm with the others methods. Table 4 shows the result of the applying algorithm to databases. Proportional to the number of samples in each database, we select top ranked genes (50–60), and then we apply them to a hybrid algorithm. Following a discussion, we introduce the biomarkers which obtained by a hybrid algorithm, then extract rules which are achieved by the decision tree from the biomarkers. The results indicate the good performance of our proposed algorithm in finding small subset of features with high accuracy. Furthermore, the results show the good similarities between our biomarkers and the biomarkers that have been introduced in others literature. From the results, we can understand that the hybrids algorithm with this classifier have a better result rather than the result which obtained from individual genetic and PSO algorithms. Furthermore, we can improve the accuracy of classification by determining its parameters automatically during the

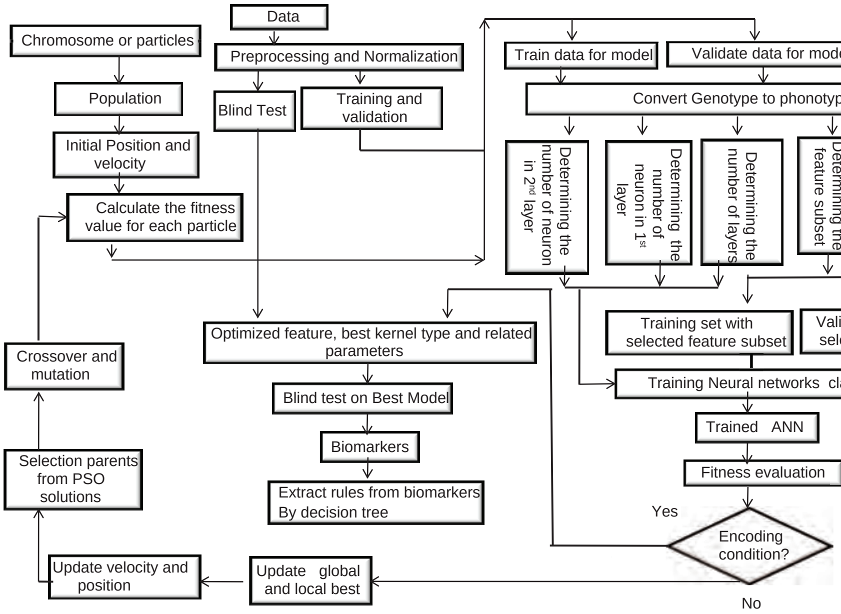
Data
Chromosome or particles
Preprocessing and Normalization
Population
Blind Test
Training and validation
Initial Position and velocity
Calculate the fitness value for each particle
Train data for model
Validate data for model
Convert Genotype to phonotype
Determining the number of neuron in 2nd layer
Determining the number of neuron in 1st layer
Determining the number of layers
Determining the feature subset
Training set with selected feature subset
Validating set with selected features
Training Neural networks classifier
Trained ANN
Fitness evaluation
Crossover and mutation
Optimized feature, best kernel type and related parameters
Blind test on Best Model
Biomarkers
Extract rules from biomarkers By decision tree
Selection parents from PSO solutions
Yes
Encoding condition?
No
Update velocity and position
Update global and local best

Figure 1: Hybrid algorithm flowchart (particle swarm optimization/genetic algorithm/artificial neural network)

Table 4: The result of applying hybrid algorithm (PSO/GA) to ANN classifier with *t*-test preprocessing on cancer databases

Datasets	Methods	PSOGA/ANN				Parameters		
		Accuracy	Sensitivity	Specificity	Precision	Number of layers	Neurons size in first layer	Neurons size in second layer
ALL/AML	GA	94.29	100	90	90	1	11	-
	PSO	94.29	90	100	100	2	6	12
	PSO/GA	100	100	100	100	1	4	-
Colon	GA	90	87.50	95	98	1	10	-
	PSO	93.33	98	70	94.67	2	5	9
	PSO/GA	96.67	96	100	100	2	3	5
Breast	GA	92	100	90	80	2	8	12
	PSO	96	96.67	95	97.50	2	8	8
	PSO/GA	96	100	95	90	1	10	-

PSO – Particle swarm optimization; GA – Genetic algorithm; ANN – Artificial neural network; ALL – Acute lymphoblastic leukemia; AML – Acute myeloid leukemia

feature selection stage with small suitable feature of subsets.

DISCUSSION

Because of more emphasizing on presented PSO/GA/ANN hybrid algorithm, we do further checks with more details on these results. Figure 2 shows the most frequent genes of running algorithm with 10 fold cross validation. In leukemia cancer type (AML, AML), 17 biomarkers are selected with our hybrid algorithm, 20 genes in colon cancer and 12 genes in breast cancer, are selected as the most frequent genes, respectively. All these genes have been repeated >6 times out of 10 times of running the algorithm. These genes are introduced on details in Table 5.

For more details, we use a heat map showing on discovered biomarkers. The point which is important is that we can view a graphical representation of the changes in genes behavior in cancer data by displaying heat map. It is desirable that the behavior of genes in cancer samples is similar but different from healthy samples. For example, a group of genes have low expression in normal samples in contrast, another group of genes has high expression in the normal sample. Hence, the thing which is important is that these genes can interact with each other to separate cancer samples from normal samples correctly.

In Figure 3, images show the heat maps of leukemia cancer in two types, colon and breast cancer, respectively. In these heat maps, red color represents values above the mean, black represents the mean, and green represents values below the mean of a gene across all columns samples.

M13241, U84487, L22524, L22524, U74612, D76435 have high expression value in ER − and the others biomarkers have low expression in these groups.

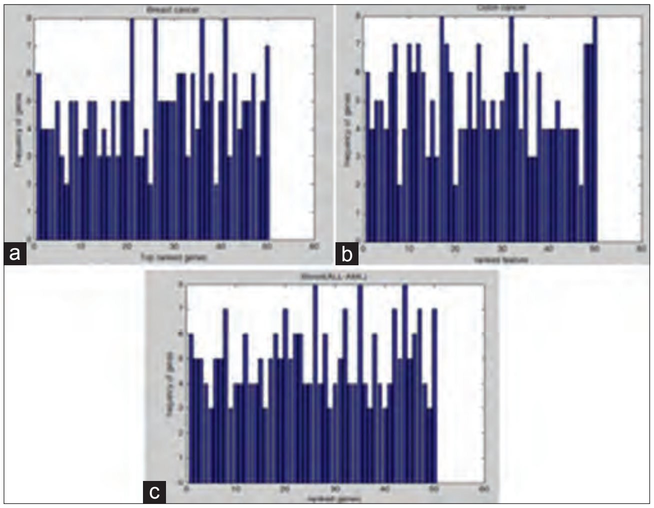

Figure 2: Occurrence frequency of genes by hybrid particle swarm optimization/genetic algorithm/artificial neural network algorithm with 10-fold cross validation. Figures from left to right are: (a) For breast cancer (b) colon cancer and (c) blood cancer type acute lymphoblastic leukemia and acute myeloid leukemia

D49950, M55150, M32304, M16038, M62762, X61587 have high expression in AML groups and M31303, M65214, D86967, D63880, X59417, S50223 X97267, X66401, U07139, L07633, M31211 have low expression in these groups. These genes can discriminant AML and ALL groups clearly. M76378, H43887 have low expression value in colon cancer, but the others founded biomarkers have high expression value in cancer sample.

At last, we apply decision tree algorithm on biomarkers which obtained by the introduced hybrid approach in Table 5 to use for finding rules between them. We use C5.0, which is one of the decision tree algorithms by SPSS clementine 12 software. In this work, we find 3 rules with 91% accuracy using 10 fold-cross validations for blood cancer type ALL, AML. In this type of blood cancer, classification is performed using two genes, M31211 and X61587. With consideration of these two genes, we found that gene X61587 has high expression in AML samples; in contrast, gene M31211 has low expression in this cancer type AML. In fact, using these two genes that have different behaviors in two class of sample, and discover rules with decision tree, can help us to have proper classification. The rest of the Table 6 shows the rules for the breast cancer and colon cancer. The obtained accuracy for these cancer data and on high ranked genes in occurrence frequency are, 91%, 89%, and 83%, respectively [Table 7]. We use a decision tree classifier to for biological point of view in our works.

In the following, we can have some comparison on proposed algorithm with the others works. The first comparison is based on an accuracy of classification which is shown in Table 6. Then we perform a comparison between the biomarkers presented in this article and the references articles. The tables show that our algorithm can achieve to high accuracy for the classification problem than others. Also in addition of high accuracy rather than other works, the procedures of modeling in references papers are test with different parameters by users, but our algorithm was running without any user's interference and any trial and errors.

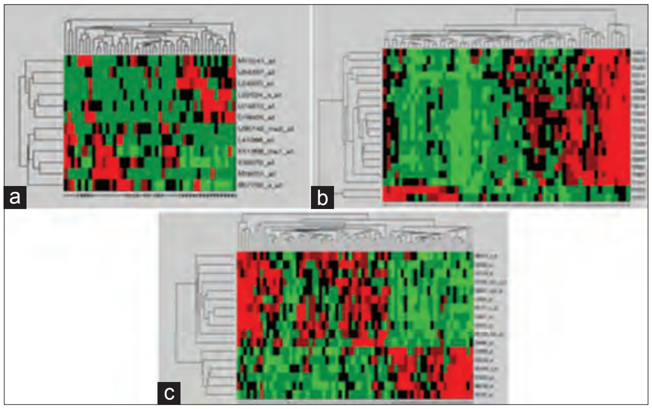

Figure 3: Heat maps view on three cancer data show the difference behavior of genes in two classes of data. (a-c) The result for breast and colon cancer data and leukemia cancer in types acute lymphoblastic leukemia and acute myeloid leukemia, respectively

In ALL and AML, the algorithm finds 17 biomarkers that 7 of them are common with reference paper (Golub, 1999).[21] Biomarkers which are in common with reference paper on leukemia (ALL, AML) are M62762, M31303, M31211, M16038, X59417, S50223, M55150. In rest two type of cancer, in breast cancer 4 biomarkers out of 12 biomarkers are the same with (West, 2001)[23] and also 4 genes out of 20 in colon cancer with (Alon, 1999).[22] Biomarkers which are in common with reference paper on breast cancer are X58072, U95740, L24203, and S37730. Biomarkers which are in common with reference paper on colon cancer are T57619, T48804, X55715, T61609.

CONCLUSION

In this paper, we have used hybrid combination of PSO and GA algorithm with ANN without any trial and user interface in determining the classifier's parameters such as number of layers and number of neurons in each layer. We give some comparison on proposed algorithm with the others. The main comparison is based on accuracy of classification that is, shown in Table 5. Following a discussion and regarding to result, it can be understood that we obtained a good result with this algorithm. The accuracy of 100% is achieved for blood cancer types 96.67% and 96% is achieved for colon and breast cancer data, respectively. This result is better than the individual use of PSO and GA algorithm and also the ability of algorithm in determining the training parameters and small feature subsets in databases perfectly with no user interface is another point of work which is proper.

Table 5: Discovered biomarkers for all groups by PSO/GA/ANN

Gene ID	Description
Colon cancer	
M76378	Human cysteine-rich protein gene, exons 5 and 6
U09587	Human glycyl-tRNA synthetase mRNA
X54941	Homo sapiens CKsHs1 mRNA for Cks1 protein homologue
T56604	Tubulin beta chain (*Haliotis discus*)
T57619	40S ribosomal protein S6 (*Nicotiana tabacum*)
U30825	Human splicing factor SRp30c mRNA
R08183	Q04984 10 kD heat shock protein, mitochondrial
T70062	Human nuclear factor NF45 mRNA
T61609	Laminin receptor (human)
H43887	Complement factor D precursor (Homo sapiens)
T86749	Human (clone PSK-J3) cyclin-dependent protein kinase mRNA
H08393	Collagen alpha 2 (XI) chain (Homo sapiens)
M26697	Human nucleolar protein (B23) mRNA
U09564	Human serine kinase mRNA
T86473	NDP kinase A (human)
T48804	40S ribosomal protein S24 (human)
T51529	Meis homeobox 3 pseudogene 1
X55715	Human Hums3 mRNA for 40S ribosomal protein s3
M36981	Human putative NDP kinase (nm23-H2S) mRNA, complete
R15447	Calnexin precursor (Homo sapiens)
Breast cancer	
X58072	GATA binding protein 3
L22524	Matrix metallopeptidase 7 (matrilysin, uterine)
L24203	Tripartite motif-containing 29
M76231	Sepiapterin reductase (7,8-dihydrobiopterin: NADP+ oxidoreductase)
U74612	Forkhead box M1
U95740	KIAA0430
D76435	Zic family member 1
U84487	Chemokine (C-X3-C motif) ligand 1
M13241	v-myc myelocytomatosis viral related oncogene, neuroblastoma derived (avian)
X51956	Enolase 2 (gamma, neuronal)
L41066	Nuclear factor of activated T-cells, cytoplasmic, calcineurin-dependent 4
S37730	Insulin-like growth factor binding protein 2, 36 kDa
Blood cancer (ALL-AML)	
M62762	ATPase, H+transporting, lysosomal 16 kDa, V0 subunit C
M16038	v-yes-1 Yamaguchi sarcoma viral-related oncogene homolog
M32304	TIMP metallopeptidase inhibitor 2
M31211	Myosin, light chain 6B, alkali, smooth muscle and nonmuscle
M31303	Stathmin 1
M65214	Transcription factor 3 (E2A immunoglobulin enhancer-binding factors E12/E47)
D49950	Interleukin 18 (interferon-gamma-inducing factor)
D86967	ER degradation enhancer, mannosidase alpha-like 1
D63880	Non-SMC condensin I complex, subunit D2
X59417	Proteasome (prosome, macropain) subunit, alpha type, 6
S50223	Zinc finger protein 22 (KOX 15)
X97267	Protein tyrosine phosphatase, receptor type, C-associated protein
M55150	Fumarylacetoacetate hydrolase (fumarylacetoacetase)
X66401	Transporter 2, ATP-binding cassette, sub-family B (MDR/TAP)
U07139	Calcium channel, voltage-dependent, beta 3 subunit
X61587	RAS homolog gene family, member G (Rho G)
L07633	Proteasome (prosome, macropain) activator subunit 1 (PA28 alpha)

PSO – Particle swarm optimization; GA – Genetic algorithm; ANN – Artificial neural network; ALL – Acute lymphoblastic leukemia; AML – Acute myeloid leukemia; SMC – Structural maintenance of chromosomes; NDP – Nucleoside diphosphate; NADP – Nicotinamide adenine dinucleotide phosphate; RAS – ???

Table 6: Summarizes results and comparison with literatures

Methods	Accuracy (%)		
Datasets	ALL/AML	Colon	Breast
Li S, 2008[30]	95.1	88.7	93.4
Shen Qi, 2008[9]	95.81	90.31	93.5
Shen Qi, 2007[7]	-	90.43	-
Mohammad Javad Abdi, 2012[31]	100	93	-
Presented PSO/GA/ANN	100	96.67	96

PSO – Particle swarm optimization; GA – Genetic algorithm; ANN – Artificial neural network; ALL – Acute lymphoblastic leukemia; AML – Acute myeloid leukemia

Table 7: Extracted rules by decision tree on 4 cancer database

Databases	Rules with decision tree
Leukemia cancer rules	If gene M31211_at >−0.451 then ALL If gene X61587_at ≤−0.587 then ALL If gene X61587_at >−0. 587 and gene M31211_at ≤−0. 451 then AML
Breast cancer rules	If X03635_at >−0.850 and M13241_at ≤−0.593 then ER+ If X03635_at ≤−0.850 then ER− If M13241_at >−0.593 then ER−
Colon cancer rules	If T48804≤−0.886 then normal If M76378>−0.303 then normal If T48804>−0.886 and M76378≤−0.303 then tumor

ALL – Acute lymphoblastic leukemia; ER – Estrogen-receptor; AML – Acute myeloid leukemia

REFERENCES

1. Schena M, Shalon D, Heller R, Chai A, Brown PO, Davis RW. Parallel human genome analysis: Microarray-based expression monitoring of 1000 genes. Proc Natl Acad Sci U S A 1996;93:10614-9.
2. Schena M, Shalon D, Davis RW, Brown PO. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. Science 1995;270:467-70.
3. Tong DL, Schierz AC. Hybrid genetic algorithm-neural network: Feature extraction for unpreprocessed microarray data. Artif Intell Med 2011;53:47-56.
4. Li-Yeh Chuang, Cheng-San Yang , Kuo-Chuan Wu, Cheng-Hong Yang; Gene selection and classification using Taguchi chaotic binary particle swarm optimization. Expert Syst Appl 2011;38:13367-77.
5. Chuang LY1, Yang CH, Wu KC, Yang CH. A hybrid feature selection method for DNA microarray data. Comput Biol Med 2011;41:228-37.
6. Liu B, Cui Q, Jiang T, Ma S. A combinational feature selection and ensemble neural network method for classification of gene expression data. BMC Bioinformatics 2004;5:136.
7. Shen Q, Wei-Min S, Kong W. A combination of modified particle swarm optimization algorithm and support vector machine for gene selection and tumor classification. Talanta 2007;71:1679-83.
8. Chuang LY, Chang HW, Tu CJ, Yang CH. Improved binary PSO for feature selection using gene expression data. Comput Biol Chem

2008;32:29-38.

9. 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.
10. Martineza E, Alvarezb MM, Trevino V. Compact cancer biomarkers discovery using a swarm intelligence feature selection algorithm. J Comput Biol Chem 2010;34:244-50.
11. 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.
12. Yang CH. A hybrid filter/wrapper method for feature selection of microarray data. J Med Biol Eng 2009;30:23-8.
13. Önskog J, Freyhult E, Landfors M, Rydén P, Hvidsten TR. Classification of microarrays; synergistic effects between normalization, gene selection and machine learning. BMC Bioinformatics 2011;12:390.
14. Wang X, Simon R. Microarray-based cancer prediction using single genes. BMC Bioinformatics 2011;12:391.
15. Shah S, Kusiak A. Cancer gene search with data-mining and genetic algorithms. Comput Biol Med 2007;37:251-61.
16. Yeh JY, Wu TS, Wu MC. Applying data mining techniques for cancer classification from gene expression data. International Conference on Convergence Information Technology [s.l.]: IEEE Computer Society; 2007.
17. Chu F, Wang L. Applications of support vector machines to cancer classification with microarray data. Int J Neural Syst 2005;15:475-84.
18. Takahashi M , Hayashi H, Watanabe Y, Sawamura K, Fukui N, Watanabe J, Kitajima T, Yamanouchi Y, Iwata N, Mizukami K, Hori T, Shimoda K and *et al*. Diagnostic classification of schizophrenia by neural network analysis of blood-based gene expression signatures. Schizophr Res 2010;119:210-8.
19. Khan J, Wei JS, Ringnér 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.
20. Nikhil R Pal , Kripamoy Aguan , Animesh Sharma, Shun-ichi Amari Discovering biomarkers from gene expression data for predicting cancer subgroups using neural networks and relational fuzzy clustering. BMC Bioinformatics 2007;8:5.
21. 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.
22. 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.
23. 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.
24. Eberhart R, Kennedy J. A New Optimizer Using Particle Swarm Theory. Proceeding of 6th International Symposium on Micro Machine and Human Science; 1995. p. 39-43.
25. Kao YT, Zahara E. A hybrid genetic algorithm and particle swarm optimization for multimodal functions. Appl Soft Comput 2008;8:849-57.
26. Du SW, Cao LK. A Learning Algorithm of Artificial Neural Network Based on GA–PSO Proceedings of the 6th World Congress on Intelligent Control and Automation – Dalian, China: [s.n.]; 2006. p. 3633-7.
27. Juang CF. A hybrid of genetic algorithm and particle swarm optimization for recurrent network design. IEEE Trans Syst Man Cybern B Cybern 2004;34:997-1006.
28. Robinson JS, Sinton RS. Yahya Particle Swarm, Genetic Algorithm, and their Hybrids: Optimization of a Profiled Corrugated Horn Antenna. San Antonio: IEEE Antennas and Propagation Society International Symposium [s.n.]; 2002. p. 314-7.
29. Kennedy J, Eberhart R. A Discrete Binary Version of the Particle Swarm Algorithm. IEEE Service Center – Piscataway: In: Proceeding of IEEE International Conference on Neural Networks; 1997. p. 4104-9.
30. Li S, Wu X, Tan M. Gene selection using hybrid particle swarm optimization and genetic algorithm. Soft Comput 2008;12:1039-48.
31. Abdi MJ, Hosseini SM. A novel weighted support vector machine based on particle swarm optimization for gene selection and tumor classification. Comput Math Methods Med 2012.

How to cite this article: 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 Sign Sence 2015;5:88-96.

Source of Support: Nil, **Conflict of Interest:** None declared

BIOGRAPHIES

Niloofar Yousefi Moteghaed received B.Sc. and M.Sc degree in biomedical engineering from the Science and research branch of Islamic Azad university University, Tehran in 2010 and 2012 respectively . Now she is Ph.D student in biomedical engineering in Shahid Beheshti University of Medical Sciences and Health since 2012 . She works in the area of bioinformatics, data mining and medical image processing.

E-mail: : nilofar.yousefi@gmail.com

Keivan Maghooli has received his B.Sc. in electronic engineering from the Shahid Beheshti University, Tehran, Iran, M.Sc. in biomedical engineering from the Tarbiat Modaress University, Tehran, Iran, and Ph.D. in biomedical engineering from the Research and Science branch, Azad University, Tehran, Iran, majoring in Data Mining, Signal Processing and Artificial Intelligence. He has been with the Biomedical Faculty at Research and Science branch, Azad University, Tehran, Iran, since 2000,where he is currently an Assistance of Professor and Head of Bioelectric department.

E-mail: K_maghooli@srbiau.ac.ir

Shiva Pirhadi received B.Sc. and M.Sc degree in biomedical engineering from the Science and Research branch of Islamic Azad university University, Tehran, Iran in 2010 and 2012 respectively . Now she is Ph.D student in biomedical engineering in Science and Research branch of Azad University since 2012 . She works in the area of bioinformatics, data mining and medical image processing.

E-mail: shv_prhd@yahoo.com

Masoud Garshasbi has received his B.Sc. in Biology from the Ferdowsi University, Mashhad, Iran (2001), and his M.Sc. in Human Genetics from the University of Social Welfare and Rehabilitation (USWR), Tehran, Iran (2003).He obtained his Ph.D. (2009) and Post doctoral (2011) in Human Molecular Genetics from Max Planck Institute, Berlin, Germany by working on the genes involved in Mental retardation. At 2011 he joined as an assistant professor to the Department of Medical Genetics, Faculty of Medical Sciences, Tarbiat Modares, Tehran, Iran. He is also founder and head of Medical Genetic Department at the DNA laboratory, Tehran, Iran.

E-mail: masoud.garshasbi@gmail.com