Genomics 107 (2016) 231–238

Contents lists available at ScienceDirect

Genomics

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

A hybrid gene selection approach for microarray data classification using cellular learning automata and ant colony optimization

Fatemeh Vafaee Sharbaf [a], Sara Mosafer [a], Mohammad Hossein Moattar [b,*]

[a] Department of Computer Engineering, Imam Reza International University, Mashhad, Iran
[b] Department of Software Engineering, Mashhad Branch, Islamic Azad University, Mashhad, Iran

ARTICLE INFO

Article history:
Received 2 January 2016
Received in revised form 20 April 2016
Accepted 1 May 2016
Available online 3 May 2016

Keywords:
Gene selection
Microarray data
Cellular learning automata
Ant colony optimization
K-nearest neighbor
Naïve Bayes

ABSTRACT

This paper proposes an approach for gene selection in microarray data. The proposed approach consists of a primary filter approach using Fisher criterion which reduces the initial genes and hence the search space and time complexity. Then, a wrapper approach which is based on cellular learning automata (CLA) optimized with ant colony method (ACO) is used to find the set of features which improve the classification accuracy. CLA is applied due to its capability to learn and model complicated relationships. The selected features from the last phase are evaluated using ROC curve and the most effective while smallest feature subset is determined. The classifiers which are evaluated in the proposed framework are K-nearest neighbor; support vector machine and naïve Bayes. The proposed approach is evaluated on 4 microarray datasets. The evaluations confirm that the proposed approach can find the smallest subset of genes while approaching the maximum accuracy.

1. Introduction

Analyzing the gene expression level, one can gather valuable information regarding the mutual influence of genes in a genetic network [1].In the databases used for gene expression analysis, the number of samples is few but the dimension is too high. These two factors make classification and data analysis challenging. However, all genes do not participate in the occurrence of cancer. Using all genes to discriminate and classify cancer may lead to incorrect decisions. Using feature selection techniques to identify the effective subset of features is an important issue in the problem of gene expression analysis. The main goal of feature selection is to identify a minimum subset of features that increase the decision accuracy.

Traditional feature selection approaches are divided into four categories namely filter, wrapper, embedded and hybrid approaches. In filter approaches each feature is evaluated individually [2]. These approaches can be easily applied to high dimensional datasets; their complexity is low and the approaches are classifier independent. For this purpose, measures such as t-test [3], information gain [4], minimum Redundancy Maximum Relevance (mRMR) [5] and Euclidean distance [6] are the most popular. In this type of feature selection approaches, the features which have the best statistical score are selected. In filter feature selection approaches, the performance of the classifier and interdependency of the features play no role, therefore it is not surprising that the performance of the classifier would be low or redundant features may be found in the selected feature set [7].

In wrapper approaches, classifier performance is used as the measures for feature evaluation. Wrapper approaches are categorized as deterministic and stochastic approaches. Sequential forward selection (SFS) and Sequential backward elimination (SBE) are categorize as deterministic and optimization based approaches such as randomized hill climbing [8], Ant colony [9] and genetic algorithms [10] are stochastic approaches. Although the classifier performance is high for this approaches but the search space complexity is very high for the problems with thousands of feature and this leads to higher time complexity.

Embedded approaches take advantage of the model properties to analysis the problem and select the most important features [11]. Approaches such as decision tree and neural network fall in this group of methods, however these approaches are also of high computational complexity. Guyon et. al. [12] introduced one of the most widely applied embedded techniques based on support vector machine and Recursive Feature Elimination (SVM-RFE) for gene selection and cancer classification. Also, Maldonado et. al. [13] proposed an embedded approach by introducing a penalty factor in the dual formulation of SVM.

None of the above mentioned approaches are able to overcome all the problems solely. Therefore ensemble approaches are proposed in the literature [14,15]. In these approaches, feature selection is done using a hybrid model and the results are integrated. Mundra et al. hybridize two of the most popular feature selection approaches, namely SVM-RFE and mRMR [16]. Shreem et. al. [17] proposed RM-GA approach

* Corresponding author.
E-mail addresses: vafaeeshaarbaf@gmail.com (F. Vafaee Sharbaf), sa.mosafer90@yahoo.com (S. Mosafer), moattar@mshdiau.ac.ir (M.H. Moattar).

http://dx.doi.org/10.1016/j.ygeno.2016.05.001
0888-7543/

which was a hybrid of ReliefF, mRMR and genetic algorithm (GA). Chuang et al. [18] proposed a hybrid approach named CFS-TGA which was the hybrid of correlation based feature selection (CFS) and Taguchi-Genetic Algorithm (TGA) and used KNN as the classifier. Lee and Liu [19] proposed an approach called Genetic Algorithm Dynamic Parameter (GADP) for producing every possible subset of genes and rank the genes using their occurrence frequency. Also, Yassi and Moattar [20] proposed a feature selection approach for microarray data which combined both ranking methods and wrapper approaches to satisfy the data scarcity problem.

In this paper, we have proposed an ensemble approach to select the smallest subset of features to have the best possible classifier performance. This approach consists of two phases. The first phase uses a filter and the second phase is based on a wrapper approach. In the first phase, the features are ranked using Fisher criterion. The use of the filter approach is intended to lower the search space complexity. Then, the best features from the previous stage are fed to the wrapper approach which is based on the hybrid cellular learning automata and ant colony optimization. The rest of this paper is organized as follows. Section 2 introduces the main materials of the proposed approach including cellular automata and ant colony optimization. Section 3 explains the proposed methodology. The evaluation datasets are described in Section 4. Section 5 summaries the experimental results and discussions. Finally conclusions and guide for feature works are offered in Section 6.

2. Materials and methods

2.1. Cellular learning automata

Cellular learning automata (CLA) are system modeling approaches which consists of simple basic parts. In CLA, the behavior of every part is modified based on the behaviors of its neighbors and its personal previous experiments. The simple parts of this model can show complicated functionalities via interactions with each other. A CLA is a cellular automaton in which every cell (or a group of cells) is equipped with learning capability.

Local rule, φ controls the cellular automata and determines if a selected action should be punished or rewarded. The rewards and punishes leads to the structural update of the cellular learning automata to achieve a specific objective. A cellular learning automaton is denoted by a penury $<\Lambda, A, \Omega, \varphi, L>$. $\Lambda = \{\lambda_1, \lambda_2, \ldots, \lambda_n\}$ denotes the set of cells in the cellular learning automata which constructs a Cartesian network. $A = \{a_1, a_2, \ldots, a_k\}$ is the set of allowed actions of a CLA in a cell. $A^t(\lambda_i)$ denotes the executed action in time t and cell λ_i and φ is the rule with governs the cellular learning automata. Ω is the neighboring cells and L is the set of learning cells. Depending on the application, the neighboring cells are determined using different approaches (i.e. Von Neumaan, Smith, Moore and Cole neighborhood) [21]. Learning automata is capable of simulating complicated systems using simple interactions of cells, and hence is appropriate for solving NP-complete problems.

2.2. Ant colony optimization

Ant colony optimization is a meta-heuristic algorithm inspired from the explorative behavior of ants. In spite of being blind and weakly intelligent, the ants can find the shortest path from home to the food and vice versa. Biologists found out that this is because of the pheromone trails that they use to communicate and exchange routing data among each other. These trails lead the ants to the shortest possible paths. Ants choose the routes, based on a probability which is proportional to the amount of the pheromones remained on the paths. The stronger the pheromone trail, the fittest the path. This algorithm has some compelling features such as: positive feedback, distributed computation, and a constructive greedy heuristic, which have attracted the researchers [22]. Positive feedback brings about a faster speed to find good solutions. Besides that, distributed computation stops the algorithm from premature and early convergence. And finally, the greedy heuristic helps in finding acceptable solutions in early stages of the search. These are the characteristics which have made the Ant Colony Algorithm robust, versatile and controllable.

3. Proposed approach

The proposed method consists of three main stages including: feature ranking using Fisher criterion, optimum feature subset selection using the hybrid method of cellular learning automata and ant colony, and final feature determination using Receiver Operating Characteristics (ROC) curve. Fig. 1 depicts a view of the proposed methodology.

3.1. Feature ranking using Fisher criterion

In this stage, in order to eliminate the weak features, we utilize a ranking method. With regards to the fact that, in recent studies the focus has been on the Fisher information measure, and this metric has proven its robustness against data scarcity [20], in this work, we used Fisher ratio to rank the features. The Fisher ratio is calculated for features using Eq. (1).

$$FR(j) = \frac{\left(\mu_{j1} - \mu_{j2}\right)^2}{{\sigma_{j1}}^2 - {\sigma_{j2}}^2} \tag{1}$$

Where, μ_{jc} is the sample mean of feature j in class c and ${\sigma^2}_{jc}$ is variance of feature j in c. The N features possessing the highest Fisher value are sent to the next stage.

3.2. Cellular learning automata-ant colony optimization feature selection (CLACOFS)

In this stage we analyze a variety of feature subsets. The N best features in the ranking phase are the input, and a subset with the smallest number of features and high discrimination would be the output. To do this, we consider the problem space as a two dimensional grid of cells. The number of cells is the least power of 2 which is greater than N. The neighborhood is considered to be of Moore type, which implies that each cell will have eight neighbors. Likewise, the cells on the left, right, up and down boundaries are considered to be neighbors.

Each cell can have one of the three states of asleep, awake, and dead. At first, all cells are awake. We consider the environment in the cellular automata to be of type Q. In this case, the feedback of the environment to a cell can have three forms of good, average, and bad. The cell would be rewarded or penalized proportional to the environment's feedback.

We assigned an ant to each cell. We used the Fisher values as heuristic information (initial predictions of feature's performance), and their average as the initial amount of pheromones. In each living cell, the ant uses the probability rule in Eq. (2) to choose features. The number of features each ant is authorized to choose is calculated randomly. The performance of the classifier is determined by the features each ant chooses and is used to update the local pheromone. The environment also analyzes each cell and proportionally rewards or penalizes it based on the performance of the classifier; which changes the cell's energy. Dropping the cell's energy level below the threshold causes it to go asleep and if these conditions remain steadily and sequentially for some iteration the cell dies out.

$$P_i^k(t) = \begin{cases} \dfrac{[\tau_i(t)]^a [\eta_i(t)]^\beta}{\sum_{u \varepsilon j^k} [t_u(t)]^a [\eta_u(t)]^\beta} & \text{if } i \varepsilon j^k \\ 0 & \text{Otherwise} \end{cases} \tag{2}$$

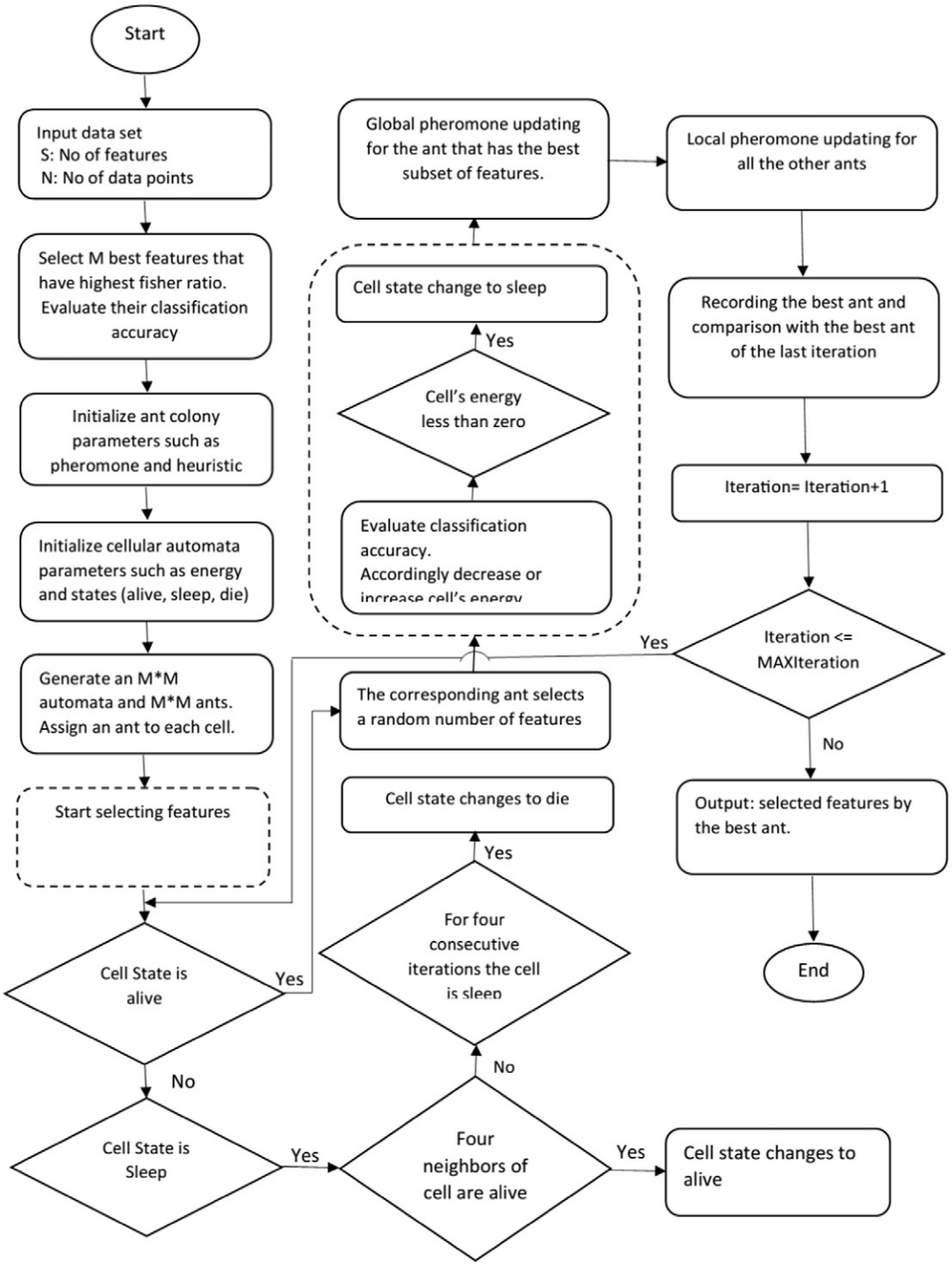

Fig. 1. The flowchart of the proposed approach.

$P_i^k(t)$ The probability of selecting i[th] feature by k[th] ant in time step t.
τ_i The pheromone amount of i[th] feature.
η_i The heuristic information of i[th] feature.
α The relational importance of pheromone.
β The relational importance of heuristic information.

If the cell is at the asleep state, we analyze its neighbor cells. If M of its neighbors were awake, the cell itself becomes awake. After determining the new state for the whole cells, the pheromone is updated globally. To do this, first the best ant must be determined. The criterion to choose the best ant is the classifier's performance which is in direct relation with the features that are chosen by that ant. If we gain the same classifier performance for some ants, we choose the one with minimum number of features. If once again we see some ants under the same circumstances, we use the cell's energy as a criterion to choose the best cell. After determining the best ant, the global pheromone is updated according to Eq. (3).

$$\tau_i(t+1) = (t+1) = (1-\rho).\tau_i(t) + \sum_{k=1}^{m} \Delta\tau_i^k(t) + \Delta\tau_i^g(t) \quad (3)$$

P Global pheromone evaporation rate.
M Number of ants (cells).

Table 1
Evaluation datasets.

Dataset	Sample size	# genes	# Class
ALL-AML leukemia	72	7129	2
Prostate tumor	136	12,600	2
MLL-leukemia	72	12,582	3
ALL-AML-4	72	7129	4

Table 2
Parameters of Classifier.

Classifier	Parameter
SVM	Kernel function = polynomial; Order of the polynomial kernel = 3; Method used to find the separating hyper plane = SMO;
K-nearest neighbor	Distance = Standardized Euclidean distance; number of nearest neighbors = 3 Or 4;
Naïve Bayes	Distribution = Gaussian; prior probabilities for the classes = empirical;

$\Delta\tau_i^g(t)$ The change amount the best ant has had on the pheromone vector as Eq. (4).

$$\Delta\tau_i^k(t) = \begin{cases} \varphi.S^k(t) + \dfrac{(1-\varphi).\left(n-|S^k(t)|\right)}{n} & \text{if } i \varepsilon S^k(t) \\ 0 & \text{Otherwise} \end{cases} \quad (4)$$

$\Delta\tau_i^k(t)$ Pheromone of feature i of the kth ant at time t.
Φ Local evaporation coefficient.
$S^k(t)$ Accuracy of the cell's classifier.
|. |: The number of chosen features by the kth ant.
N The total number of features.

In case the system prematurely converges, ants will choose the same features. After updating the global pheromone, the information of the current and the previous iterations are compared, the ant with best performance is saved as the algorithm's final solution up to this iteration. The algorithm continues running until the number of features in the final solution of the algorithm goes below x and the accuracy of the classifier exceeds y, or it reaches the maximum number of iterations noted as T. If in some sequential iteration, the information of the current iteration's best ant remains steady, the information of whole cells is reset. A summary of the algorithm is depicted in Algorithm 1.

Table 3
Average classification accuracy for each feature ranking method.

Data Set	Classifier	T test	Information Gain	Fisher	Z score
Prostate tumor	SVM	86.12	73.00	**88.00**	85.75
	KNN	96.25	73.75	**98.12**	97.37
	NB	**97.00**	95.87	**97.00**	**97.00**
ALL-AML leukemia	SVM	74.12	43.5	**76.37**	58.37
	KNN	**91.00**	76.37	89.50	63.00
	NB	88.37	60.25	**96.25**	58.00
MLL-leukemia	KNN	**75.25**	63.75	**75.25**	71.37
	NB	73.12	**76.25**	64.12	68.75
ALL-AML-4	KNN	67.37	67.12	**67.62**	63.37
	NB	67.75	**69.37**	68.12	63.12

The bold entries denote the highest accuracy among the feature selection approaches for the mentioned data set and classifier.

Algorithm 1. The pseudo code of the proposed CLACOFS

While fulfill termination condition do

1. Set parameters (Each ant is assigned to a cell).
2. For alive cells do
  a. R features selected by each ant (R is random for each ant)
  b. Determine the efficiency classifier
  c. If the classifier accuracy is more than threshold T1
    Increase the energy of the cell (energy = energy + R1)
    else if the classifier accuracy is between threshold T1 and T2
    Increase the energy of the cell (energy = energy + R2)
    else
    Decrease the energy of the cell (energy = energy-P)
    If cell energy is less than T, put cell in sleep state
  d. Update local pheromone
  e. Select best ant
  f. Update global pheromone

3. For sleep cells do
  a. If cell has N neighbor in alive state Put cell in alive state
  b. If cell is in asleep in successive step put it in die state

3.3. Final feature determination using ROC curve

Each time we run the algorithm a subset of features are chosen. Among these subsets, the subsets with the minimum cardinality and

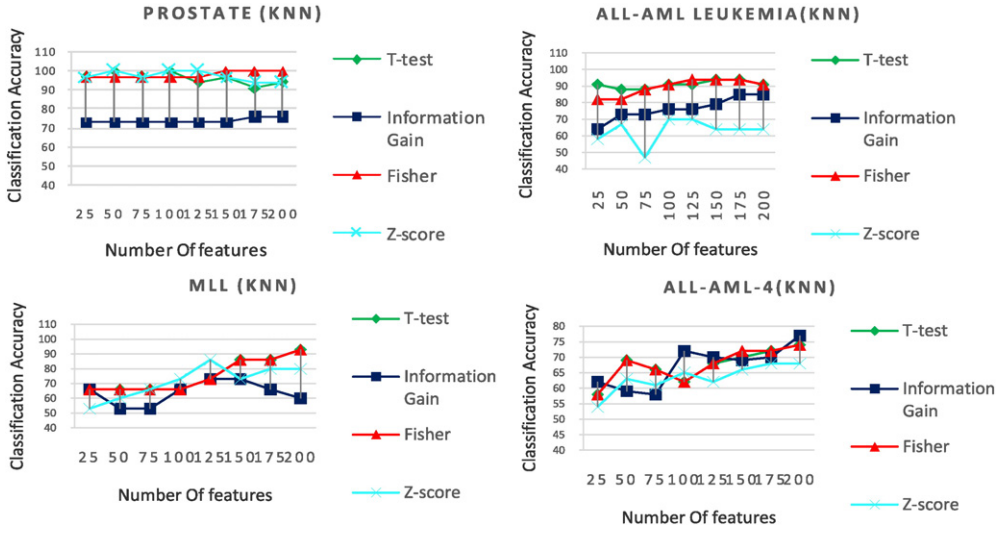

Fig. 2. Classification accuracy of KNN with different feature ranking methods on evaluation datasets.

Table 4
The classification accuracy at the end of the proposed CLACOFS step.

Data set	Classifier	Classification using all features		Classification using selected features in first phase		Classification using selected features in second phase	
		No of features	Accuracy %	No of features	Accuracy %	Average No of selected features	Accuracy %
ALL-AML leukemia	SVM	7129	58.8235	100	74.65	2.55	95.95
				150	73.45	2.65	94.30
	KNN		70.5882	100	86.05	3.60	95.95
				150	86.20	3.35	95.20
	NB		94.1176	100	96.10	5.25	97.60
				150	96.40	5.35	97.00
Prostate tumor	SVM	12,600	73.5294	100	83.50	14.05	98.35
				150	88.15	15.40	99.25
	KNN		97.0588	100	96.55	9.45	99.40
				150	98.20	14.50	99.85
	NB		26.4706	100	97.05	6.50	99.10
				150	97.05	9.20	99.40
MLL	KNN	12,582	80.00	200	93.33	18.70	97.55
				250	93.33	24.75	94.05
	NB		100.00	200	93.33	12.40	98.95
				250	100	14.90	99.30
ALL-AML-4	KNN	7129	73.60	100	62.49	15.77	80.99
				150	70.82	20.65	80.51
	NB		91.66	100	73.60	12.23	86.30
				150	69.44	13.08	86.38

maximum accuracy are chosen. Now, to determine the final features the ROC curve is used. ROC curve shows the sensitivity versus specificity. Sensitivity shows the ratio between the correctly classified positive samples and the true positive samples.

$$\text{Sensitivity} = \frac{\text{Correctly Classified Possitive Samples}}{\text{True Positive Samples}} \quad (5)$$

Specificity shows the ratio between the correctly classified negative samples and the true negative samples.

$$\text{Specificity} = \frac{\text{Correctly Classified Negative Samples}}{\text{True Negative Samples}} \quad (6)$$

We partitioned the dataset samples into two groups of train and test using 10-fold cross validation technique. For each best feature set, we plotted the ROC curve and calculated the Area under Curve (AUC). The feature with the highest AUC is introduced as the final feature set.

4. Evaluation datasets

To evaluate the proposed algorithm, we utilized four datasets. A description of these datasets is given in Table 1.

ALL-AML Leukemia [23]: This dataset has two classes named AML, and ALL. Each sample contains 7129 genes. In the training set, there are 27 samples in ALL class and 11 samples in AML class. Also, in test set 20 samples belong to ALL class, and 14 samples are from AML class.

Prostate [24]: In this two-class dataset, the training set contains 52 prostate cancer samples and 50 healthy samples, and also the test set consists of 25 cancer samples and 9 healthy samples. Each sample has 12,600 genes.

MLL-Leukemia [25]: This dataset is consisted of three classes named ALL, MLL, and AML. Each sample is structured by 12,582 genes. The training set in this dataset consists of 57 samples. The number of samples in ALL, MLL, and AML classes are 20, 17, and 20, respectively. Similarly, there are 15 samples in the test set and ALL, MLL, and AML classes contain 4, 3, and 8 samples, respectively.

ALL-AML-4 [23]: In this dataset there exist four classes called B-cell, T-cell, BM, and PB, each class consisting 38, 9, 21, and 4 samples. Each sample is described by 7129 features. Furthermore, to create the test samples we used the 4-Fold technique.

5. Experiment and result

In the experimental study, first, we have examined the impact of different ranking methods on feature selection. Second, we have chosen the best method for the first phase of the proposed algorithm. Third, we have used the CLA-ACO model to introduce subsets of superior features. Eventually, for this purpose, we have examined different classifiers such as SVM, KNN and Naïve Bayes (NB) (worth mentioning that SVM is not applied for more than two class problems). Table 2 represents the parameters of the proposed approach.

In Fig. 2, the results of feature selection by different ranking methods such as T-test, information gain, fisher and Z-score are shown for 4 datasets. For summary, this figure only depicts the classification accuracy of KNN classifier. The horizontal axis is the number of selected features and the vertical axis shows the classification accuracy.

To evaluate the feature selection results using each ranking method, the average results are depicted in Table 3. Comparing the results, it is apparent that the fisher ranking approach has the highest performance.

Table 5
AUC value of top 5 features.

Data set	Number (gene symbol/accession)	AUC		
		SVM	NB	KNN
Prostate tumor	6185 (37639_at)	0.8427	0.7069	0.9167
	9937 (40607_at)	0.6875	0.6042	0.6011
	9267 (38740_at)	0.7886	0.6042	0.6829
	6462 (38634_at)	0.6581	0.6042	0.8272
	4690 (32076_at)	0.8038	0.6042	0.6481
ALL-AML leukemia	1834 (M23197_at)	1.0000	0.9875	0.9875
	2354 (M92287_at)	0.6883	0.8860	0.9156
	6041 (L09209_s_at)	0.9839	0.9715	0.9364
	6855 (M31523_at)	0.6000	0.9680	0.9456
	2642(U05259_rna1_at)	0.6557	0.8588	0.8491
MLL-leukemia	10,998(1718_at)	NA	0.8941	0.8728
	2436(35485_at)	NA	0.8681	0.6919
	640(32016_at)	NA	0.8312	0.7506
	10,515(33117_r_at)	NA	0.7729	0.5587
	12,291(385_at)	NA	0.9117	0.8092
ALL-AML-4	3469 (U59878_at)	NA	0.9867	0.9487
	4366 (X61587_at)	NA	0.9355	0.8556
	2121 (M63138_at)	NA	0.7396	0.9535
	4514 (X71973_at)	NA	0.9016	0.7605
	1834 (M23197_at)	NA	0.7853	0.7911

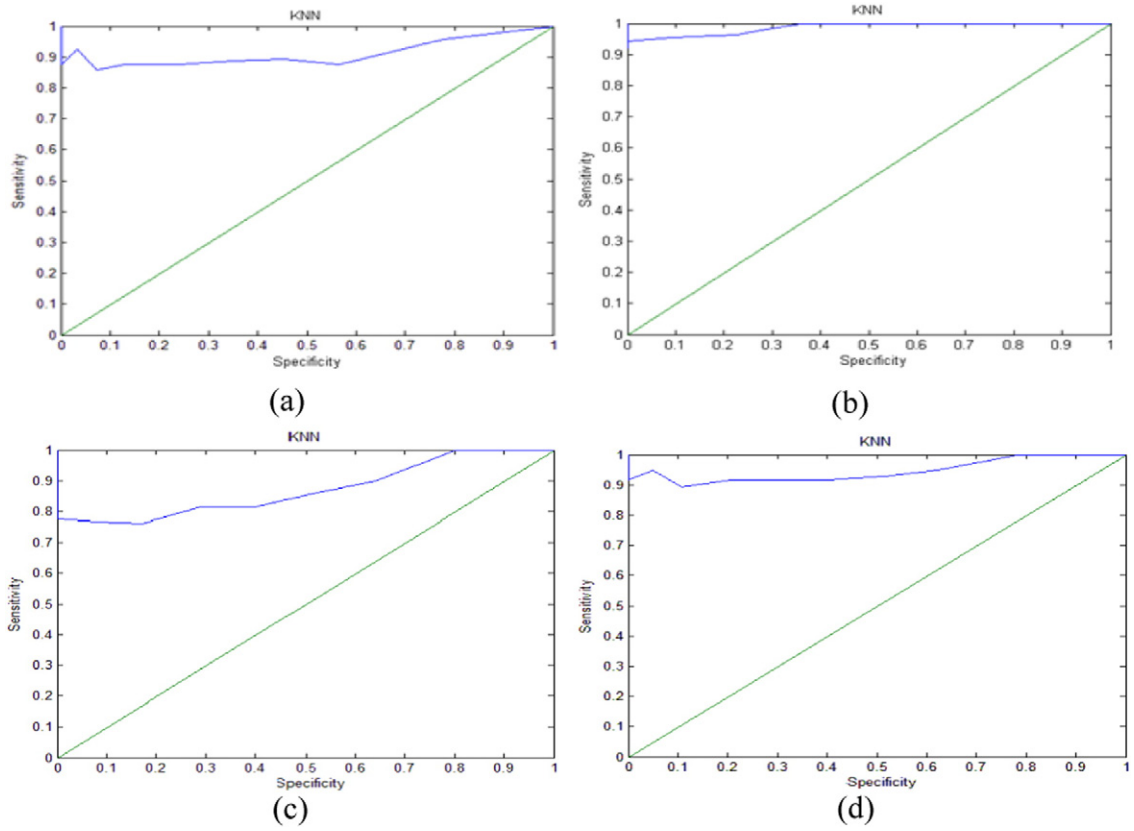

Fig. 3. ROC curve for the best selected gene and KNN classifier a) 37639_at gene for Prostate Tumor b) 1834 gene for ALL-AML Leukemia c) 10,998 gene for MLL-Leukemia and 3469gene for ALL-AML-4. Blue curve depicts the proposed approach while green curve is the diameter.

The results prove that fisher criterion is, on average, the best ranking approach for microarray.

In Table 4, the results of classification at the end of first and second stage are shown. In this table, we have first used all features for the classification of samples. Then, the samples were classified by N selected features in the ranking phase. Comparing the classification accuracy in these cases shows that the ranking phase improves the classification performance. In the second phase, the proposed CLACOFS is performed 20 times and the average results are depicted in Table 4. In this phase, by features that are far less than the selected features in the first phase, classification accuracy has been increased.

By comparing the results achieved from the different classifiers, it is shown that the Naïve Bayes classifier outperforms the other two classifiers; and the KNN classifier is better than the SVM classifier. It is due to the fact that in Naïve Bayes classifier, the features are considered independent and this consideration increases the performance of classifier. We have to remind that the SVM classifier was not applied for more than 2 class problems such as MLL and ALL-AML-4 dataset.

In Table 5, the top five genes of each database are depicted. The more frequently chosen genes, in multiple execution of the proposed algorithm, were considered superior. The ROC curve was drawn for each superior gene. For plotting the ROC, we first use the 10-fold cross validation in order to divide the database samples into train and test groups. Then, for each fold, the Sensitivity and Specificity are calculated and the ROC curve is drawn. Finally, by comparing the area under curves (AUC), the final feature is determined. The gene which has the greatest AUC will be introduced as the final feature. This gene has been shown in Table 5 with distinct color. NA in Table 5 denotes that SVM was not evaluated for MLL and ALL-AML-4 datasets as stated before.

The top 5 genes mentioned in Table 2 are comparable with the-state-of-the-art findings for each disease. For example, in [32] genes with accession number of (M23197_at), (L09209_s_at) and (M63138_at) are selected among the top 20 genes for ALL-AML Leukemia diagnosis. Also, [31] denotes that the combination of (U05259_rna1_at) and (M92287_at) are effective for Leukemia diagnosis and can achieve the accuracy of 94.12%. This article also suggests (M23197_at) and (L09209_s_at) as informative gene for Leukemia diagnosis.

As referred to in [29] which is a US patent, 1718_at is one of the most informative genes for the purpose of leukemia diagnosis. Also, this

Table 6
Comparison of the methods on the prostate tumor dataset.

Method (feature selection + classification)	Number of selected genes	Classification accuracy %
Signal to noise ratios + KNN [24]	4	77
SFS + Bayes classifier [7]	4	97
PMOGA + SVM [26]	89	100
PMOGA + Naïve Bayes [26]	89	100
Proposed feature selection + SVM	2	100
Proposed feature selection + KNN	2	100
Proposed feature selection + Naïve Bayes	2	100

Table 7
Comparison of the methods on the ALL-AML leukemia dataset.

Method (feature selection + classification)	Number of selected genes	Classification accuracy %
PMOGA + SVM [26]	89	97
PMOGA + NB [26]	89	94
CLARANS + Naïve Bayes [27]	44	97.22
AODEsr [28]	5	95.8
AODEsr [28]	10	100
Proposed feature selection + SVM	1	100
Proposed feature selection + KNN	2	100
Proposed feature selection + Naïve Bayes	1	100

feature is included in other texts such as [30,31] as one of the biomarkers which can signify the disease. On the other hand, NCBI profiles show that 1718_at can be expressed for leukemia diagnosis. The same discussion can be done for the top 5 selected genes for prostate tumor. For instance, NCBI profiles show that (37639_at) is highly associated with the growth and progression of cancers, particularly prostate cancer. Fig. 3 show the ROC curve for best feature in each dataset.

In the following, SFS [7], PMOGA [26], CLARANS [27], and AODEsr [28] approaches are compared with our proposed scheme. In SFS, which is a wrapper method, the accuracy of classifier plays a decisive role in determining the subset of effective features. The PMOGA applies a new multi-purpose evolutionary approach and uses two evaluation functions which are based on the concepts of mathematics (probability theory and Rough set theory). The CLARANS method uses the clustering algorithm based on random search. In this method, the clustering and dimension reduction are done based on gene ontology (GO). In AODEsr approach entropy is used for gene selection. AODEsr applies a decision approach called averaged-on dependence estimator with subsumption resolution (AODEsr) to solve cancer recognition problem. Also, the authors of [28] applied Entropy Minimization Discretization (EMD) method which is popular in discretization of high-dimensional data.

The proposed approach is compared with the above mentioned approaches for two of the four datasets namely, Prostate Tumor and ALL-AML Leukemia. The results of evaluations are depicted in Tables 6 and 7.

As shown below the proposed framework has achieved the accuracy of 100% on both datasets while from the other approaches only PMOGA and AODEsr has gained similar performance on Prostate Tumor and ALL-AML Leukemia, respectively. However, compared to the best performed approaches, still the proposed approach has found the least number of features (i.e. 1 or 2) to achieve this gain. This fact shows that the proposed approach is both efficient in reaching the best accuracy and finding the smallest subset of features which is due the high capability of CLA to model the complex problem of high dimensional data.

6. Discussion

This paper proposed a feature selection approach to find the most informative genes for cancer classification and diagnosis which applies cellular automata (CA) to model the interaction between genes and ant colony optimization (ACO) to learn the rules and structure of CA. The main objective of the proposed approach is to find the smallest subset of biomarkers which can signify the disease efficiently. The contribution of CA is that it can effectively model the interactions in complex systems and is appropriate for the problem of feature selection from high dimensional data. Besides being effective in modeling complicated relations, CA provides the possibility of parallel processing. Also, ACO is chosen due to its advantages such as higher convergence rate and computational effectiveness. Also, using ROC curve in order to select the final subset, will guarantee the method's stability. By stability, we mean the ability of the approach to be extended to datasets with much less samples. The proposed approach is evaluated on 4 popular microarray datasets and compared with some of the most recent approaches.

The results show that although selecting a minimal subset of features, the selected genes have high influence on separating different classes. The experiments denote that the proposed approach has achieved high accuracy rate even with 1 or 2 highly informative genes. Also, as compared with the-state-of-the-art, the selected genes are previously found meaningful in the biology texts. In future works, other evolutionary algorithms such as artificial bee colony that have been used for feature selection can be evaluated together with CA. Also, due to the good discrimination power of the selected genes, it is worth evaluating them through biological researches and experiments.

Appendix A. Supplementary data

Supplementary data to this article can be found online at http://dx.doi.org/10.1016/j.ygeno.2016.05.001.

References

[1] S.-Y. Ho, C.-H. Hsieh, H.-M. Chen, H.-L. Huang, "Interpretable gene expression classifier with an accurate and compact fuzzy rule base for microarray data analysis", Bio Systems 85 (2006) 165–176.
[2] R. Ruiz, J.C. Riquelme, J.S. Aguilar-Ruize, M. Garcia-Torres, Fast feature selection aimed at high-dimensional data via hybrid sequential ranked searches, Expert Syst. Appl. 9 (2012) 11094–11102.
[3] P. Jafari, F. Azuaje, An assessment of recently published gene expression data analyses: reporting experimental design and statistical factors, BMC Med. Inform. Decis. Mak 6 (2006) 27.
[4] Z. Wang, "Neuro-Fuzzy Modeling for Microarray Cancer Gene Expression Data,", Proceedings of the Second International Symposium on Evolving Fuzzy Systems 2005, pp. 241–246.
[5] C. Ding, H. Peng, "Minimum Redundancy Feature Selection from Microarray Gene Expression Data,", Proceedings of the IEEE Conference on Computational Systems Bioinformatics 2003, pp. 523–528.
[6] S.B. Cho, H.H. Won, "Machine Learning in DNA Microarray Analysis for Cancer Classification,", Proceedings of the First Asia-Pacific Bioinformatics Conference on Bioinformatics 2003, pp. 189–198.
[7] S. Miyano, S. Imoto, A. Sharma, A top-r feature selection algorithm for microarray gene expression data, IEEE/ACM Trans. Comp. Biol. Bioinf. 9 (3) (2012) 754–764.
[8] D. Skalak, "Prototype and Feature Selection by Sampling and Random Mutation Hill Climbing Algorithms,", Proceedings of the Eleventh International Conference on Machine Learning 1994, pp. 293–301.
[9] Y.M. Chiang, S.Y. Lin, "The Application of Ant Colony Optimization for Gene Selection in Microarray-based Cancer Classification,", Proceedings of the Seventh International Conference on Machine Learning and Cybernetics 2008, pp. 12–15.
[10] M. Karzynski, A. Mateos, J. Dopazo, Using a genetic algorithm and a perceptron for feature selection and supervised class learning in DNA microarray data, Artif. Intell. Rev. 20 (1) (2003) 39–51.
[11] J. Canul-Reich, L. Hall, D. Goldgof, J. Korecki, S. Eschrich, Iterative feature perturbation as a gene selector for microarray data, Int. J. Pattern Recognit. Artif. Intell. 26 (05) (2012) 111–135.
[12] I. Guyon, J. Weston, S. Barnhill, V. Vapnik, Gene selection for cancer classification using support vector machines, Mach. Learn. 46 (1–3) (2002) 389–422.
[13] S. Maldonado, R. Weber, J. Basak, "Simultaneous feature selection and classification using kernel-penalized support vector machines,", Inf. Sci. 181 (1) (2011) 115–128.
[14] Y. Ye, Q. Wu, J. Zhexue, M. Huang, X. Ng, X. Li, "Stratified sampling for feature subspace selection in random forests for high dimensional data,", Pattern Recogn. 46 (3) (2013) 769–787.
[15] Q. Song, J. Ni, G. Wang, "A fast clustering-based feature subset selection algorithm for high-dimensional data,", IEEE Trans. Knowl. Data Eng. 25 (1) (2013) 1–14.
[16] P. Mundra, J. Rajapakse, "SVM-RFE with mRMR filter for gene selection,", IEEE Trans. Nano. Biosci. 9 (1) (2010) 31–37.
[17] S. Shreem, S. Abdullah, M. Nazri, M. Alzaqebah, Hybridizing ReliefF, MRMR filters and GA wrapper approaches for gene selection, J. Theor. Appl. Inf. Technol 46 (2) (2012) 1034–1039.
[18] L. Chuang, C. Yang, K. Wu, C. Yang, A hybrid feature selection method for DNA microarray data, Comput. Biol. Med. 41 (4) (2011) 228–237.
[19] C. Lee, Y. Leu, A novel hybrid feature selection method for microarray data analysis, Appl. Soft Comput. 11 (1) (2011) 208–213.
[20] M. Yassi, M.J. Moattar, Robust and stable feature selection by integrating ranking methods and wrapper technique in genetic data classification, Biochem. Biophys. Res. Commun. (2014) 850–856.
[21] M.R. Meybodi, "Expriment With Cellular Learning Automata,", Technical Report, Amirkabir University of Technology, Computer Engineeting Department, Augest 2000.
[22] M.Dorigo, T.Stuzle, Ant Colony Optimization 2004MTI Press.
[23] T.R. Golub, D.K. Slonim, P. Tamayo, C. Huard, M. Gassenbeek, J.P. Mesirov, H. Coller, M.L. Loh, J.R. Downing, M.A. Caligiuri, C.D. Bloomfield, E.S. Lander, Molecular classification of cancer: class discovery and class prediction by gene expression monitoring, Science 286 (1999) 531–537.
[24] D. Singh, P.G. Febbo, K. Ross, D.G. Jackson, J. Manola, C. Ladd, P. Tamayo, A.A. Renshaw, A.V. D'Amico, J.P. Richie, E.S. Lander, M. Loda, P.W. Kantoff, T.R. Golub, W.R. Sellers, Gene expression correlates of clinical prostate cancer behavior, Cancer Cell 1 (2002) 203–209.
[25] S.A. Armstrong, J.E. Staunton, L.B. Silverman, R. Pieters, M.L. den Boer, M.D. Minden, S.E. Sallan, E.S. Lander, T.R. Golub, S.J. Korsemeyer, MLL translocations specify a distinct gene expression profile that distinguishes a unique leukemia, Nat. Genet. 30 (2002) 41–47.
[26] S.K. Pati, A.K. Das, A. Ghosh, "Gene Selection Using Multi-objective Genetic Algorithm Integrating Cellular Automata and Rough Set Theory,"Springer International Publishing Switzerland Lecture Notes in Computer Science Volume 8298, Vol. 2013, (2013), pp 144–155.
[27] S. Mitra, S. Ghosh, Feature selection and clustering of gene expression profiles using biological knowledge, IEEE Trans. Syst. Man Cybern. Part C Appl. Rev. 42 (2012) 1590–1599.

[28] S. Lei Win, Z.Z. Htike, F. Yusof, I.A. Noorbatcha, "Cancer recognition from DNA microarray gene expression data using averaged one-dependence estimators,", IJCI 3 (2) (April 2014).
[29] J.R. Downing, et al., "Classification and Prognosis Prediction of Acute Lymphoblastic Leukemia by Gene Expression Profiling", U.S. Patent 2014/0018513 A1, Jan 29, 2004.
[30] H. Moon, et al., "Sex-specific genomic biomarkers for individualized treatment of life-threatening diseases,", Dis. Markers 35 (6) (2013) 661–667.
[31] X. Wang, O. Gotoh, Accurate molecular classification of cancer using simple rules, BMC Med. Genet. 2 (2009) 64.
[32] H.Q. Wang, D.S. Huang, A gene selection algorithm based on the gene regulation probability using maximal likelihood estimation, Biotechnol. Lett. 27 (2005) 597–603.