Research Journal of Applied Sciences, Engineering and Technology 7(21): 4445-4454, 2014
DOI:10.19026/rjaset.7.821
ISSN: 2040-7459; e-ISSN: 2040-7467

Submitted: December 09, 2013 Accepted: January 11, 2014 Published: June 05, 2014

Research Article

Swarm Intelligence Approach Based on Adaptive ELM Classifier with ICGA Selection for Microarray Gene Expression and Cancer Classification

[1]T. Karthikeyan and [2]R. Balakrishnan
[1]P.S.G. College of Arts and Science,
[2]Dr.NGP Arts and Science College, Coimbatore, India

Abstract: The aim of this research study is based on efficient gene selection and classification of microarray data analysis using hybrid machine learning algorithms. The beginning of microarray technology has enabled the researchers to quickly measure the position of thousands of genes expressed in an organic/biological tissue samples in a solitary experiment. One of the important applications of this microarray technology is to classify the tissue samples using their gene expression representation, identify numerous type of cancer. Cancer is a group of diseases in which a set of cells shows uncontrolled growth, instance that interrupts upon and destroys nearby tissues and spreading to other locations in the body via lymph or blood. Cancer has becomes a one of the major important disease in current scenario. DNA microarrays turn out to be an effectual tool utilized in molecular biology and cancer diagnosis. Microarrays can be measured to establish the relative quantity of mRNAs in two or additional organic/biological tissue samples for thousands/several thousands of genes at the same time. As the superiority of this technique become exactly analysis/identifying the suitable assessment of microarray data in various open issues. In the field of medical sciences multi-category cancer classification play a major important role to classify the cancer types according to the gene expression. The need of the cancer classification has been become indispensible, because the numbers of cancer victims are increasing steadily identified by recent years. To perform this proposed a combination of Integer-Coded Genetic Algorithm (ICGA) and Artificial Bee Colony algorithm (ABC), coupled with an Adaptive Extreme Learning Machine (AELM), is used for gene selection and cancer classification. ICGA is used with ABC based AELM classifier to chose an optimal set of genes which results in an efficient hybrid algorithm that can handle sparse data and sample imbalance. The performance of the proposed approach is evaluated and the results are compared with existing methods.

Keywords: Adaptive extreme learning machine, artificial bee colony algorithm, biology and genetics, classifier design and evaluation, feature evaluation and selection

INTRODUCTION

Cancer detection and classification for diagnostic and prognostic purposes is usually based on pathological investigation of tissue section, resultant in individual interpretation of data (Eisen and Brown, 1999). The limited information gained from morphological analysis/pathological investigation is often insufficient to aid in cancer diagnosis and may result in expensive but ineffective treatment of cancer.

With the appearance and speedy development of DNA microarray technologies in previous work (Eisen and Brown, 1999; Lipshutz *et al*., 1999), classification of cancer by identification of corresponding gene expression profiles has previously concerned many efforts from a wide assortment of research communities. From this classification of cancer becomes major important to the diagnosis of diseases and treatment. Without the accurate identification of cancer types, it is rarely possible to give useful therapy and accomplish probable effects. Conventional classification methods are mainly dependent on the following works that is the morphological appearance of tumors, parameters derivative from clinical observations and extra biochemical technique. Their application is restricted by the presented uncertainties and their prediction accurateness needs to further improvement (Golub *et al*., 1999). DNA based microarray technologies present a new researchers to analysis the cancer, propose a new technique to examine the pathologies of cancer beginning a molecular angle under a methodical structure and more, to make further accurate prediction result in prognosis and treatment.

In order to precisely recognize cancer subtypes, numerous up to date studies have been carried out to recognize genes that might cause cancer (Peng *et al*., 2003; Saeys *et al*., 2007; Koller and Sahami, 1996). Advances in microarray technology and improved methods for processing and converted biological data

Corresponding Author: T. Karthikeyan, P.S.G. College of Arts and Science, Coimbatore, India

methods have augmented these studies. For illustration, the analysis of Microarray Gene Expression Data (MGED) (Piatetsky-Shapiro and Tamayo, 2003) enables molecular cancer classification through the gene selection, which might serve as markers for dissimilar types of cancers. However, selection of optimal sets of genes features is complex by the occurrence of a huge numeral of genes and the availability of very few and uneven numbers of samples per class (sparse and imbalanced). It significantly affects the classification performance. Computational molecular categorization combined with machine learning techniques may offer a more reliable and cost-effective method for identifying different types of cancers, which might lead to better treatment and prognosis for this disease. The recognition of genes profiles to outcome the better cancer classification, it might also improve the identification of each cancer type to survive and thrive. This kind of information make will provide better way for developing suitable drugs to treat precise cancers. Some of the important major issues to classification of the microarray data are: robustness of gene selection and gene ranking, considerate of issues associated to feature selection and assessment of the selected genes (Ein-Dor *et al*., 2006; Stolovitzky, 2003). Classification, possible use and diversity of feature selection techniques are discussed in Saeys *et al*. (2007).

In this study, a better gene selection and cancer classification technique is proposed for microarray data that is described by sample sparseness and imbalance. In this study, an Integer-Coded Genetic Algorithm (ICGA) (Saraswathi *et al*., 2011) is used for well-built and healthy gene selection of microarray data. Next, propose an Artificial Bee Colony algorithm (ABC) (Karaboga and Basturk, 2007) and driven Adaptive Extreme Learning Machine (AELM) (Jia and Hao, 2013), for managing the sparse/imbalanced data of classification problem that occurs in microarray data analysis.

LITERATURE REVIEW

In Alba *et al*. (2007) discussed the comparison of the optimization techniques such as PSO and Genetic Algorithm (GA) for gene selection and classification of the cancer is performed by using SVM classification for high dimensional in microarray dataset. To validate and estimate the cancer classification result in SVM classifier applies the 10-fold cross-validation. The primary work is to shows that PSOSVM is able to discover interesting genes and to present improved classification result. Improved version of Geometric PSO is evaluated for the primary moment in time in this effort with a binary representation (0 and 1) in Hamming space. PSOSVM based classification results can be compared with a new GASVM and also compared with other existing methods of gene selection. A following important work consists in the concrete finding of new and challenging outcomes on six public datasets identifying significant in the growth of a variety of cancers (leukemia, breast, colon, ovarian, prostate and lung).

In Liao *et al*. (2006) presented a novel gene selection procedure based on Wilcoxon rank sum test and classification is performed by using Support Vector Machine (SVM). Selection of the best feature is performed by using Wilcoxon rank, afterthat SVM classifier with linear kernel was performed to train and test the classification result. Leave-One-Out Cross Validation (LOOCV) classification consequences on two datasets: breast cancer and ALL/AML leukemia, show that the proposed technique is capable of get 100% success result with last reduced subset. The selected genes are listed and their expression levels are sketched, which show that the selected genes can make clear separation between positive and negative classes.

Microarrays technologies permit biologists to superior understand the relations among varied pathologic state at the DNA/gene level (Shanmugavadivu and Ravichandran, 2013). Though, the amount of data generated by Microarrays technology tools becomes challenging issues. To overcome these issues further need new methods in order to extract exact information regarding gene activity in responsive process like tumor cells propagation and metastasis activity. Recent Microarrays technology tools that examine microarray expression data have broken correlation-based approach such as clustering analysis. To evaluate the importance of gene selection for classification based on gene expression data developed a novel GA (Genetic Algorithm) /ANN (Artificial Neural Network) distributed technique. Numeral of different approaches/techniques has been developed and a comparison has been specified. The developed technique is applied to breast cancer tumor classification and it can be validated with real lifetime database. Gene Ontology based tools evaluation has been carried out for further evaluation of the breast cancer tumor classification.

Recently, Saraswathi *et al*. (2011) presented a novel approach which combines the optimization techniques such as Integer-Coded Genetic Algorithm (ICGA) and Particle Swarm Optimization (PSO) for gene/feature selection and it can be coupled with the neural-network-based Extreme Learning Machine (ELM) for cancer classification. PSO-ELM shows better classification result than the ICGA-ELM. Major part of the work is to further extension of Saras Saraswathi work to improve the cancer classification result and best optimal gene selection.

PROPOSED METHODOLOGY FOR GENE SELECTION AND CLASSIFICATIO

Proposed work which combines Integer-Coded Genetic Algorithm (ICGA) and Artificial Bee Colony

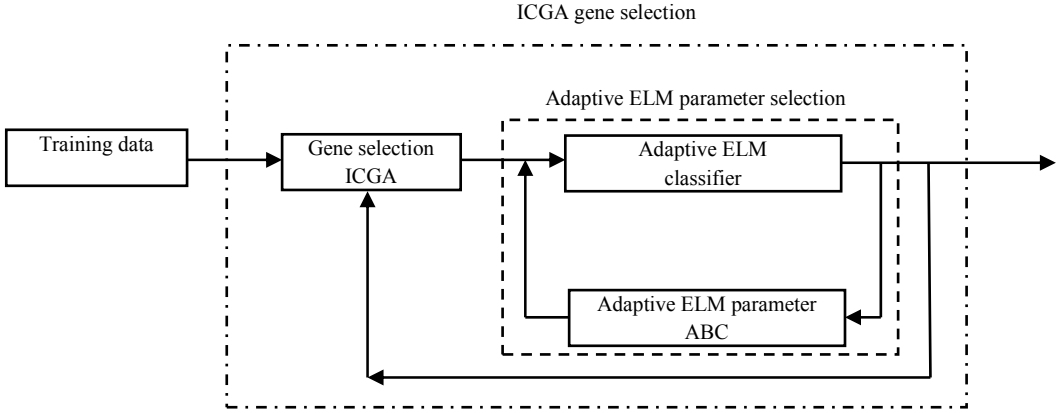


Fig. 1: Schematic diagram of the two-stage ICGA-ABC-AND_ELM multiclass classifier

(ABC), together with the Adaptive Extreme Learning Machine (AELM). ICGA and ABC are used for gene /feature selection and ALEM for cancer classification. It handles sparse data and sample imbalance dataset for classification of gene expression data.

ABC based adaptive extreme learning machine classifier for accurate classification with ICGA based gene selection: ABC based Adaptive Extreme Learning Machine (ABC based AELM) and ICGA based gene selection approach is proposed, which selects optimal feature/genes and the chosen appropriate genes are used for accurate classification of a sparse and imbalanced data set. The fundamental ELM classifier can differentiate the cancer classes between the data denoting the chosen features quickly, but the performance of ELM classifier is based on the nature of the input data distribution. For sparse and highly imbalanced data set classification, random the input weight in ELM classifier degrades the performance of classification to a huge amount (Suresh *et al*., 2010). ABC based AELM classifier is proposed in this study, where the ABC algorithm is utilized to identify the optimal input weights such that AELM classifier can differentiate the cancer classes significantly, i.e., the performance of the AELM classifier is improved. In this study, the original data are separated into training and predicting datasets. The training data results were obtained based on the input and output weight values are derived in the ABC, the cancer classification can be predicted directly through the well-known AELM.

The proposed methodology of the block diagram is shown in the Fig. 1 as follows.

The schematic diagram for the classification and optimal gene selection procedure is shown in Fig. 1. The performance of AELM classifier is mostly based on the selected input genes from ABC and ICGA methods. In order to decrease the computational aspect, an ICGA is used to select and minimize the number of genes, which can distinguish the cancer classes that is positive and negative efficiently. Based on those selected genes, AELM algorithm generates significant classifier. The proposed research work for classification and optimal gene selection procedure is shown in Fig. 1. Initially, ICGA chooses n independent genes/ features from the existing gene dataset. From this ABC will identify the optimal parameters (number of hidden nodes and input weights) such that the accuracy of the AELM multiclass classifier is improved. The best validation performance (η+) will be utilized as fitness for the ICGA evolution. The validation performance of AELM classifier (η) is used in ABC for selection of AELM parameters.

Adaptive Extreme Learning Machine model for forecasting (AELM): In general, ELM algorithm may suffer from either under-fitting or over-fitting problems. For these two problems, over-fitting is further significant when the original data is enough and the network is sufficiently difficult. ELM model with over-fitting will generally degrades the predictive performance. In order to overcome these problem Adaptive Extreme Learning Machine model (AELM) is proposed and this algorithm can decrease the chance of over fitting, improves the prediction performance of cancer classification. In this model, the perception data is used to alter the inputs of the ELM in the prediction processing making the inputs approach to the learning data. In this study, the output of the network is only one value, which is the predicted the optimal cancer classification result. Thus, we discuss our model only for one output.

Firstly, a strategy is used to initialize the input data $Q_i = [\, q_i^1, q_i^2, \ldots q_i^n]$. The strategy adopts the adaptation distance space metric which is like similar to the adaptive k nearest neighbor method presented by (Liao *et al*., 2006). The data set $Q_i = [\, q_i^1, q_i^2, \ldots q_i^n]$ is compared with the learned input patterns $X_i = [\, x_i^1, x_i^2, \ldots, x_i^n\,]$; $i = 1, 2, \ldots, N$. Estimation of closeness measure is important factor in prediction classification accuracy. Closeness is normally defined in terms of Euclidean distance space metric. But in this work closeness is measured using Minkowski distance space metrics:

$$L_M(Q_i, X_i) = (|q_i^1 - x_i^1|^d + |q_i^2 - x_i^2|^d + \cdots + |q^n{}_i - x^n{}_i|^d)^{1/d} \quad (1)$$

The above mentioned equation gives the assessment of differences among Q_i and X_i, but the differences of trends and amplitudes are not presented. In time-series forecasting, the information on trends and amplitudes is the crucial factor.

The assessment of difference among Q_i and X_i, doesn't exactly measures the closeness, introduce a adaptive metric to solve this problem and it is represented as:

$$L_A(Q_i, X_i) = min\lambda_{r,ur} f_r(\lambda_{r,ur}) \quad (2)$$

$$f_r(\lambda_r, u_r) = (|q^1{}_i - \lambda_r x^1{}_i - u_r|^2 + |q^2{}_i - \lambda_r x^2{}_i - u_r|^2 + \ldots + |q^n{}_i - \lambda_r x^n{}_r - u_r|^2)^{1/2} \quad (3)$$

For d = 2, two equations are considered:

$$\begin{cases} \frac{\partial f_r(\lambda_{r,u_r})}{\partial \lambda_r} = 0 \\ \frac{\partial f_r(\lambda_{r,u_r})}{\partial u_r} = 0 \end{cases} \quad (4)$$

When the equivalent linear system is solved, the solution of the minimization problem can be obtained systematically:

$$u_r = \frac{z_1 z_2 - z_3 z_4}{n z_2 - z_3^2}, \lambda_r = \frac{n z_4 - z_1 \frac{z}{3}}{n z_2 - z_3^2} \quad (5)$$

where, $z_1 = \sum_{i=1}^n q_i$, $z_2 = \sum_{i=1}^n q_i^2$, $z_3 = \sum_{i=1}^n x_i$, $z_4 = \sum_{i=1}^n q_i x_i$.

From this the input vector of the first network can be defined as:

$$input_v = (\hat{q}_i^{1v}, \hat{q}_i^{2v}, \ldots, \hat{q}_i^{nv}) = \frac{q_i^1 - u_{r_v}}{\lambda_{r_v}}, \frac{q_i^2 - u_{r_v}}{\lambda_{r_v}}, \ldots, \frac{q_i^n - u_{r_v}}{\lambda_{r_v}} \quad (6)$$

Most input values can be close to the historical data using this method. The forecasting error increases dramatically due to the big difference between training data and input data. In order to get more accurate results for time series $q_i^1, q_i^2, \ldots, q_i^n, k$ sets of inputs are used and the output vector are $output_v = b_v, v = 1, 2, \ldots, k$. The mechanism for admixture of outputs is presented as follows.

The majority of input values can be close to the historical data only by using this process. Due to this process forecasting error values are dramatically increasing due to the variation among training data and input data. In order to get more accurate results for prediction result $q_i^1, q_i^2, \ldots, q_i^n, k$ sets of inputs are used and the output vector are $output_v = b_v, v = 1, 2, \ldots, k$. The mechanism for a mixture of outputs is represented as follows:

$$Result = \frac{1}{k-1} \sum_{v=1}^k \left(\lambda_{r_v} b_v + u_{r_v}\right) \left(1 - \frac{d_v}{U}\right) \quad (7)$$

$$U = \sum_{v=1}^k d_v \quad (8)$$

where, d_v is the distance among Q_i's v^{th} nearest pattern and Q_i. From Eq. (7), the forecasting result is calculated from b_v $v = 1,2, \ldots, k$ with different weighing coefficients, the better coefficient is specified for closer input data of Q_i. Based on the methodology proposed above, the forecasting scheme can be formulated as shown in Fig. 2.

In the cancer prediction/forecasting schema, focus on one-step ahead pint forecasting Let $y_1, y_2, y_3, \ldots, y_t$ be a time series for classification. At time t for $t \geq 1$, the next value $y_t + 1$ will be predicted based on the observed training results y_t, $y_{t-1}, y_{t-2}, \ldots, y_1$. For ELM model, its result is generally different from time to time because the input weights and hidden biases are randomly selected. It is well recognized that the mean value the forecasting/predication is more reliable. So, a regression based integration schema is proposed in this study to obtain higher prediction accurateness. The final predicated classification result data is only the mean of S predicted classification time series:

$$\overline{y}_{t+1} = \frac{1}{s} \sum_{i=1}^s \tilde{y}_{t+1}^s \quad (9)$$

And the final predicted classification data is only the mean of the *s* classified data series:

$$\overline{y}_{t+1} = \frac{1}{s} \sum_{i=1}^s \tilde{y}_{t+1}^s$$

In order to compare the $AELM$ method with other methods on all the classified data series, adequate error measure method must be selected. The Mean Squared Error (NMSE) is used as the error criterion, which is the ratio of the mean squared error to the variance of the time series. It defined, for a time series y_i, by.

In order to compare the $AELM$ method with other methods on all the prediction accuracy time series measure the error values in the classification. The Normalized Mean Squared Error (NMSE) is used as the error criterion, which is the ratio of the mean squared error to the variance of the time series. It defined, for a time series y_i, by:

$$NMSE = \frac{\sum_{i=1}^M (y_i - \tilde{y}_i)^2}{\sum_{i=1}^M (y_i - \hat{y}_i)^2} = \frac{\sum_{i=1}^M (y_i - \tilde{y}_i)^2}{M\sigma^2} \quad (10)$$

$$\hat{y}_i = \frac{1}{M} \sum_{i=1}^M y_i \quad (11)$$

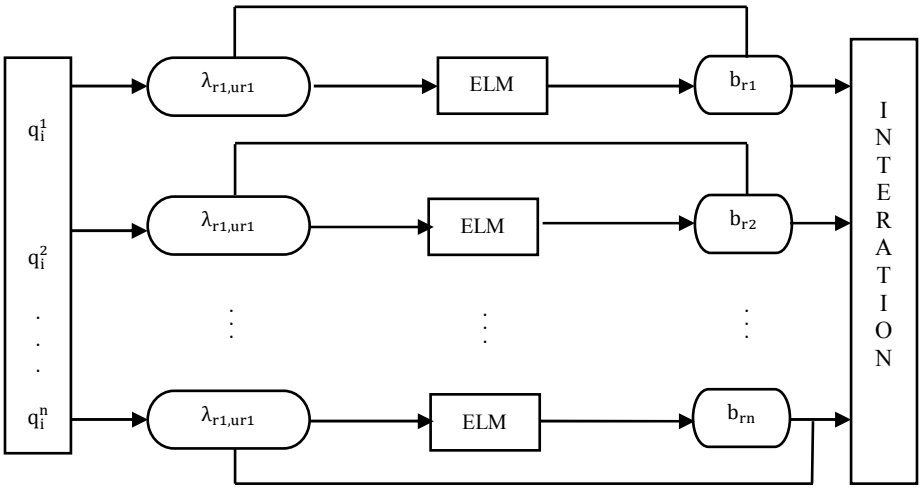


Fig. 2: The forecasting scheme for adaptive extreme learning machine for classification

where, y_i is the source point, $\hat{y}_i$ is the predicted point, M is the number of predicted points and σ^2 is the mean value and variance values are estimated from the input data. A value of NMSE $= 1$ means simply predicting the average. Another important measure is the Mean Absolute Percentage Error (MAPE) to measure the statistical performance of classification, it is represented as:

$$MAPE = \frac{1}{M}\sum_{i=1}^{M}\left|\frac{y_i-\hat{y}_i}{y_i}\right| * 100\% \quad (12)$$

ABC algorithm: Karaboga and Basturk (2007) proposed an Artificial Bee Colony (ABC) algorithm. It is one of the optimization algorithms to find optimal solution or solve optimization problem. The algorithm works based on the honey bee foraging behavior. The general Procedure of ABC algorithm as follows.
Pseudo-code of the ABC algorithm:

Load training samples
Generate the initial population $z_i, i = 1, \ldots . . sN$
Evaluate the fitness (f_i) of the population
Set cycle to 1
Repeat
FOR each employed bee {
    Produce new solution v_i by using (15)
    Calculate the value of f_i
    Apply greedy selection process}
Calculate the probability value p_i for the solution (z_i) by (14)
FOR each Onlooker bee {
    Produce new solution v_i by using (15)
    Calculate the value of f_i
    Apply greedy selection process}
If there is an abandoned solution for scout
Then replace it with a new solution which will be randomly produced by (16)
Memorize the best solution so far
Cycle = cycle + 1
Until cycle = MCN

In general ABC algorithm consists three major bees are present employed bees, onlookers and scouts. A bee is waiting to take decision for choosing the food source is named as onlooker bee and the bee which is previously visited in the food source is named as employee bee. A bee which is used to find new food source in a random way it is called as scout bee. The position of the food source solves the optimization problem with finite solution and the nectar amount of the food source measured according to the fitness value is defined as:

$$fit_i = \frac{1}{1+f_i} \quad (13)$$

The cost function of f_i is calculated according to study (Kavipriya and Gomathy, 2013).

An artificial onlooker bee chooses a food source based on the possibility value associated with that food source, p_i computed by the following Eq. (14):

$$p_i = \frac{fit_i}{\sum_{n=1}^{SN} fit_n} \quad (14)$$

where,
SN = The total number of food sources which is equal to the total number of employed bees in the food source
fit_i = The fitness of the solution given in Eq. (13) which is inversely proportional to the f_i

In order to produce a new candidate food location from the old solution in the memory, the ABC uses the following Eq. (15):

$$v_{ij} = z_{ij} + \emptyset_{ij}(z_{ij} - z_{kj}) \quad (15)$$

where, $k \in \{1, 2, \ldots . SN\}$ and $j \in \{1, 2, \ldots . D\}$ are randomly selected indexes. Even though k is identified randomly, it has to be dissimilar from i. ϕ_{ij} denotes a random number between interval (-1, 1). It controls the invention of neighbor food sources about z_{ij} and represents the assessment of two food positions visually by a bee. As can be seen from (15), as the difference between the parameters of the z_{ij} and z_{kj} decreases, the perturbation on the location z_{ij} gets decreased, too. Therefore, as the search reaches the optimal result in the search space, the step length is minimized.

The food source of the nectar is empty then the bee is replaced with a new food source location found by the scouts. In ABC, this is simulated by generating a location at random and replacing it with the discarded one. In ABC, if a position/location may not be further improved via a fixed number of cycles, then that food source is judged as discarded. The value of fixed number of cycles is an important organizes parameter for ABC algorithm. It is to be implicit that the discarded/abounded source is z_i and $j \in \{1, 2, \ldots, D\}$, then the scout finds a new food source to be replaced with z_i. This operation can be defined as in (16):

$$z_i^j = z_{min}^j + rand(0,1)(z_{max}^j - z_{min}^j) \quad (16)$$

Following each one candidate source position v_{ij} is generated and then predictable by the artificial bee, its performance is measured with that of its old source position. If a new food source position is equal or improved than old source, then it is replaced with the old basis in the memory. If not, the old one is kept as same in memory. Alternatively, a greedy selection method is employed as the selection procedure among the old and the candidate individual. Finally the global optimal result is obtained.

ABC based AELM classifier: Thus, the greatest particle and the food source position of the equations are attained from the fitness value, the first term denotes the current velocity, the second term denotes the local search and the third term is the global search.

The fitness value of the particles is the evaluation of efficiency of the AELM classifier, whose $a_{li} = H$, $b_{li} = V$and RMSE = b is initialized using the particle:

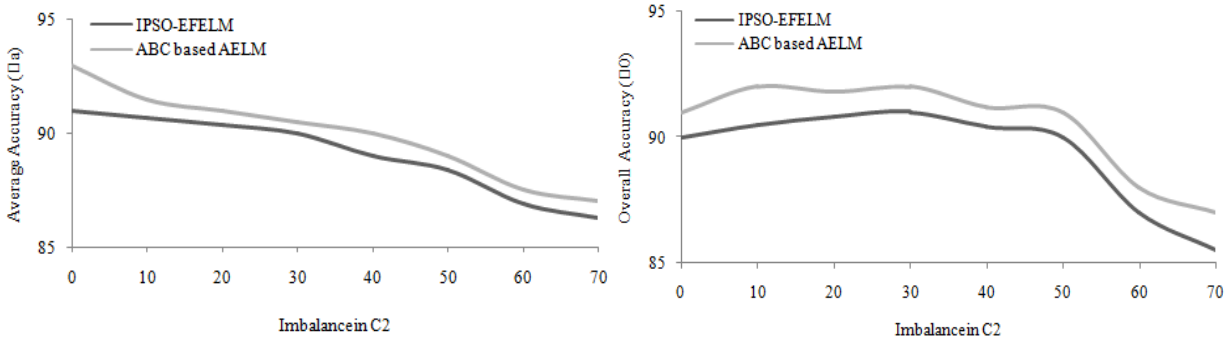
$$F = \eta$$

The ABC searches for the best H, V and b values that systematically computed weight in the ELM classifier which results in improved generalization performance. The cross validation performance of best H, V and b is η^+. The main factor in ABC based AELM classifier is to establish the amount of imbalanced data set that the classifier can handle without losing performance significantly (Guo *et al*., 2011).

Analysis on imbalance data: The sample imbalance handling capability of ABC based AELM classifier is based on the technique in (Suresh *et al*., 2008). The number of samples in one of the class was reduced and performance of the classifier was examined for different imbalance criteria. A similar examination was conducted for the proposed ABC based AELM classifier and the average (η_a), overall (η_0) and individual (η_2) classification efficiencies obtained are shown in Fig. 3.

It is observed that the average and overall classification efficiency of ABC based AELM classifier is almost constant up to 50% sample imbalance in class 2 data. By proper selection of the input weights and bias value, a better classification of performance can be attained. If careful examination is not taken then the classification performance of AELM classifier falls considerably with sample imbalance.

Integer-coded genetic algorithm: Genetic algorithms are widely used to solve composite optimization problems, in which the numeral of parameters and constraint are huge and systematic solutions are not easy to achieve (Michalewicz, 1994). In recent years, a numeral of techniques has been proposed for integrating genetic algorithms and neural networks. Genetic Algorithms are found to be effective in gene selection and classification. The study of selection function and genetic operators of GA are described in (Michalewicz, 1994).



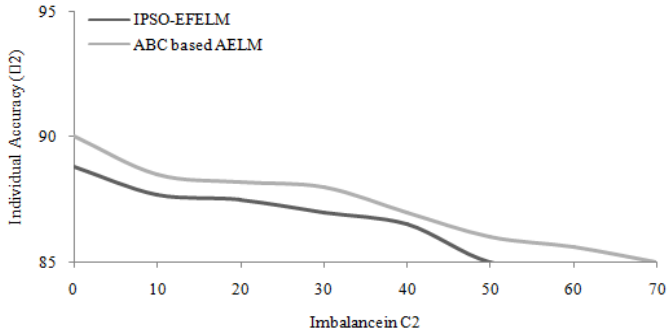


Fig. 3: Effects of the imbalances in data are depicted here, where the performance of the AELM classifier was analyzed for different imbalance conditions

Descriptions of string representation and Fitness are given below.

String representation: In this study, ICGA is used for selection of N best independent features from the given dataset. The characteristic string, which denotes N independent features, is given as:

$$S = [F_1, F_i, F_j, L, F_N]$$

Where the selected features belong to the set S and they are independent.

Fitness: The main aim of feature selection is to determine the features that demonstrates the input output characteristics of the data. The results of the ABC based AELM fivefold cross-validation test are used as fitness criteria, i.e., for the selected features, ABC will identify the best hidden neurons, input weights and biases values and return the validation efficiency obtained by the AELM algorithm along with the best AELM parameters. The features returning the best validation efficiency eventually are chosen as representative of the full data set:

$$F_i = \eta^+$$

The best solution is obtained after a known number of generations are used to expand a classifier (AELM) using the complete training set. This classifier is then used to classify the testing samples.

EXPERIMENTAL RESULTS

In this section, the performance of the proposed approach is compared with other methods based on Global Circulation Models (GCM) data set, in two steps. Initially, with the GCM data set the classification results were compared with other classifiers and then the results for gene selection are compared with other existing results for gene selection. The samples in each class are small with high sample imbalance in GCM data set, that is, large number of classes with high dimensionality requires attention for selection of samples to training and testing. In these experiments, the original data set is dividing into training and testing data.

Global cancer map data: The GCM data is the collection of six different medical institutions around 14 different types of malevolent tumors. It consists of 190 primary complete tumor samples and 8 samples are not used here called metastasis. Each sample contains the virtual expression of 16,063 genes (take for granted a one-to-one mapping from gene to probe set ID). From 190 samples, 144 samples are utilized for gene selection and classifier growth and the left behind 46 samples are used for assessment of the generalization performance. The amount of training samples per class varies from 8 to 24 which are sparse and imbalanced. Based on these notes, the GCM data set is sparse in environment with a high sample imbalance and a high-dimensional feature space for huge number of genes. The main objective is to select sets of genes from the 16,063-dimensional space and identify the smallest number of genes desired to concurrently categorize every tumor types with greater accuracy.

In turn to calculate the classifier performance for sparse and imbalance data set, the results obtained by the proposed ABC based AELM classifier for a given number of genes is compared them with the existing classifiers. Here, 98 genes as selected in Ramaswamy *et al*. (2001) as the source for the classifier performance comparison. The ABC based AELM classifier is ruined to recognize the paramount number of hidden neurons, input weights and bias by means of 144 training data. With the use of best AELM parameters, an AELM classifier is developed by means of the complete training data and the resultant classifier is tested on the remaining 46 samples. In this study the experiment were conducted for a variety of random combinations samples of 144 training and 46 testing set and the results are account in Table 1.

Table 1: Comparative analysis on classification methods for GCM data set using 98 genes selected as explained in Ramaswamy *et al*. (2001)

		Training		Testing	
Various methods	ns	Mean	S.D.	Mean	S.D.
SVM (Suresh *et al*., 2008)	106	96.50	1.85	73.78	5.10
ELM (Zhang *et al*., 2007)	50	92.30	2.25	79.43	6.23
PSO_ELM (Saraswathi *et al*., 2011)	36	94.91	1.42	85.13	4.88
IPSO_E-FELM	30	93.14	1.23	88.45	3.94
ABC based AELM	26	92.85	1.10	89.74	3.24

S.D.: Standard deviation

Table 2: Performance of proposed classifier for the best set of features selected by ABC based AELM with ICGA gene selection approach

	Training efficiency (%)			Testing efficiency		
Genes	Avg.	Max.	S.D.	Avg.	Max.	S.D.
14	94	95	2	74	82	6
28	94	95	2	72	86	6
42	92	96	1	75	98	4
56	92	96	1	88	97	3
70	95	98	2	90	97	4
84	95	98	2	93	97	4
98	94	98	2	94	99	4

Avg.: Average; S.D.: Standard deviation; Max.: Maximum

Table 3: Genes selected from GCM data set that were used for classification by ABC based AELM with ICGA

GCM 42 genes							
Gene	Accession ID	Gene #	Accession ID	Gene #	Accession ID	Gene #	Accession ID
572	D79987_at	1882	M27891_at	7870	AA232836_at	13781	RC_AA403162_at
5836	HG3342-HT3519_s_at	6868	M68519_rnal_at	8034	AA278243_at	13964	RC_AA416963_at
917	HG3432-HT3618 _at	6765	M96132_at	8107	AA287840_at	14565	RC_AA446943_at
5882	HG417-HT417_s_at	3467	U59752_at	8231	AA320369_s_at	14793	RC_AA453437_at
1119	J04611_at	3804	U80017_rna2	8975	AB002337_at	11421	X05978_at
1137	J05068_at	6154	V00565_s_at	9546	H44262_at	476	D50678_at
9731	L13738_at-2	11443	X52056_at-2	9833	M21121_s_at		
1383	L20320_at	4629	X79510_at	10322	R74226_at		
9781	L40904_at	4781	X90872_at	12020	RC_AA053660_at		
5319	L46353_at	4944	Y00815_at	12182	RC_AA100719_s_at		
1655	L77563_at	11606	Z30425_at-2	12717	RC_AA233126_at		
1791	M20530_at	7284	AA036900_at	13541	RC_AA347973_at		

From the Table 1, examine that the ABC based AELM classifier gives better performance than the existing IPSO_E-FELM classifier for 98 genes selected in Ramaswamy *et al*. (2001).

ABC based AELM with ICGA based gene selection and classification results: The proposed approach is called to select 14, 28, 42, 56, 70, 84 and 98 genes, respectively from the original 16,063 genes using a 10-fold cross-validation method on the 144 training samples. The unexploited testing set (46 samples) is worn to assess the generalization performance. ABC based AELM with ICGA is identified best genes for each set. In this experiments, create that the best genes are chosen throughout different runs do not share any common genes. The overlap between the best genes sets (14-98) chosen by proposed approach is insignificant, but their ability to differentiate the cancer classes is more or less similar. These results show that there be real subsets of genes that can discriminate or differentiate the cancer classes efficiently.

The performance of the proposed classifier by creating 100 random trials on the training and testing data sets is done by the optimal gene sets are selected as above. It helps us to predict the classifier sensitivity to data difference. The average, maximum and standard deviations of training and testing performances are given in Table 2.

Performance comparison of proposed ABC based AELM with ICGA classifier with existing methods: The proposed approach for the GCM data set results is compared with other existing methods. Table 3 shows the minimum number of genes needed by each method to attain utmost generalization performance. From the Table 4, the proposed ABC based AELM with ICGA selects a minimum 42 genes with a high average testing accuracy. GA/SVM, selects a minimum of 26 genes which gives results close to ABC based AELM with ICGA performance. It was seen that genes chosen in a variety of runs for any given subset do not have major overlaps also there is no any overlap of genes between any two subsets. Until now, the classifiers improved by means of these sets of selected genes make similar classification performance and were experiential to have the same discriminatory power to classify various

Table 4: Minimum number of genes required by various methods to achieve maximum generalization performance

Data set	Gene selection method	Genes	Avg. testing accuracy (%)
GCM	Proposed ABC based	42	92
	AELM with ICGA	98	95
	ICGA_IPSO_E-FELM	42	90
		98	94
	ICGA_PSO_ELM	42	88
		98	91
	GA/SVM (Eisen *et al*., 1998)	26	85

Table 5: Results for gene selection and classification by ABC based AELM with ICGA for different data sets

			Testing accuracy (%)	
Data set	Classes #	Genes #	Avg.	Best
Lymphoma	6	12	99	100
CNS	2	12	100	100
Breast cancer-B	4	12	93	100

cancer classes. The ABC based AELM with ICGA gene selection and classifier was used to select the minimum number of genes necessary for accurate classification is shown in the Table 4. The average classification accuracies are given in Table 5.

CONCLUSION

In this study, an accurate gene selection and sparse data classification for microarray data is done by using ABC based AELM with ICGA gene selection for multiclass cancer classification is proposed. ICGA selected genes included with optimal input weights and bias values selected by ABC and used by the AELM classifier, to deal with higher sample imbalance and sparse data conditions resourcefully. Hence, ICGA gene selection approach is integrated with the ABC based AELM classifier to identify a dense set of genes that can discriminate cancer types efficiently resulting in enhanced classification results.

REFERENCES

Alba, E., J. Garcia-Nieto, L. Jourdan and E. Talbi, 2007. Gene selection in cancer classification using PSO/SVM and GA/SVM hybrid algorithms. Proceeding of IEEE Congress on Evolutionary Computation (CEC, 2007), pp: 284-290.

Ein-Dor, L., O. Zuk and E. Domany, 2006. Thousands of samples are needed to generate a robust gene list for predicting outcome in cancer. P. Natl. Acad. Sci. USA, 103(15): 5923-5928.

Eisen, M. and P. Brown, 1999. DNA arrays for analysis of gene expression. Method. Enzymol., 303: 179-205.

Eisen, M.B., P.T. Spellman, P.O. Brown and D. Bostein, 1998. Cluster analysis and display of genome-wide expression patterns. P. Natl. Acad. Sci. USA, 95: 14863-14868.

Golub, T.R., D.K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J.P. Mesirov, H. Coller, M.L. Loh, J.R Downing, M.A. Caligiuri, C.D. Bloomfield and E.S. Lander, 1999. Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. Science, 286(5439): 531-537.

Guo, Z., J. Wu, H. Lu and J. Wang, 2011. A case study on a hybrid wind speed forecasting method using BP neural network. Knowl-Based Syst., 24(7): 1048-1056.

Jia, J. and S. Hao, 2013. Water demand forecasting based on adaptive extreme learning machine. Proceeding of 2013 International Conference on Artificial Intelligence and Software Engineering, ISBN: 978-90-78677-71-0.

Karaboga, D. and B. Basturk, 2007. Artificial Bee Colony (ABC) Optimization Algorithm for Solving Constrained Optimization Problems. LNCS: Advances in Soft Computing: Foundations of Fuzzy Logic and Soft Computing, Springer-Verlag, 4529: 789-798.

Kavipriya, P. and C. Gomathy, 2013. An efficient dynamic orthogonal variable spreading factor code allocation approach in WCDMA through swarm intelligence technique. Int. J. Eng. Technol. (IJET), 5(5): 3828-3838.

Koller, D. and M. Sahami, 1996. Toward optimal feature selection. Proceeding of 13th International Conference on Machine Learning, pp: 284-292.

Liao, C., S. Li and Z. Luo, 2006. Gene selection for cancer classification using wilcoxon rank sum test and support vector machine. Proceeding of International Conference on Computational Intelligence and Security, pp: 368-373.

Lipshutz, R., S. Fodor, T. Gingeras and D. Lockhart, 1999. High density synthetic oligonucleotide arrays. Nat. Genet., 21: 20-24.

Michalewicz, Z., 1994. Genetic Algorithm + Data Structures = Evolution Programs. 3rd Edn., Springer-Verlag, New York, pp: 18-22.

Peng, S., Q. Xu, X.B. Ling, X. Peng, W. Dua and L. Chen, 2003. Molecular classification of cancer types from microarray data using the combination of genetic algorithms and support vector machine. FEBS Lett., 555(2): 358-362.

Piatetsky-Shapiro, G. and P. Tamayo, 2003. Microarray data mining: Facing the challenges. SIGKDD Explorations, 5(2): 1-5.

Ramaswamy, S., P. Tamayo, R. Rifkin, S. Mukherjee, C.H. Yeang, M. Angelo, C. Ladd, M. Reich, E. Latulippe, J.P. Mesirov, T. Poggio, W. Gerald, M. Loda, E.S. Lander and T.R. Golub, 2001. Multiclass cancer diagnosis using tumor gene expression signatures. P. Natl. Acad. Sci. USA, 98(26): 15149-15154.

Saeys, Y., I. Inza and P. Larran, 2007. A review of feature selection techniques in bioinformatics. Bioinformatics, 23(19): 2507-2517.

Saraswathi, S., S. Sundaram, N. Sundararajan, M. Zimmermann and M. Nilsen-Hamilton, 2011. ICGA-PSO-ELM approach for accurate multiclass cancer classification resulting in reduced gene sets in which genes encoding secreted proteins are highly represented. IEEE-ACM T. Comput. Bi., 8(2): 452-463.

Shanmugavadivu, T. and T. Ravichandran, 2013. Gene selection for cancer classification using microarrays. Int. J. Comput. Appl. Technol. Res., 2(5): 609-613.

Stolovitzky, G., 2003. Gene selection in microarray data: The elephant, the blind men and our algorithms. Curr. Opin. Struc. Biol., 13(3): 370-376.

Suresh, S., N. Sundarajan and P. Saratchandran, 2008. A sequential multi-category classifier using radial basis function networks. Neuro Comput., 71(7-9): 1345-1358.

Suresh, S., S. Saraswathi and N. Sundararajan, 2010. Performance enhancement of extreme learning machine for multi-category sparse cancer classification. Eng. Appl. Artif. Intel., 23: 1149-1157.

Zhang, R., G.B. Huang, N. Sundararajan and P. Saratchandran, 2007. Multicategory classification using an extreme learning machine for microarray gene expression cancer diagnosis. IEEE-ACM T. Comput. Bi., 4(3): 485-495.