An Enhanced Gene Selection Methodology for Effective Microarray Cancer Data Classification

Dilwar Hussain Mazumder

Department of Computer Science and Engineering
National Institute of Technology Nagaland
Dimapur - 797103, Nagaland, India
Email: dilwar@nitnagaland.ac.in

Ramachandran Veilumuthu

Department of Information Science and Technology
Anna University
Chennai - 600025, Tamil Nadu, India
Email: rama5864@gmail.com

***Abstract* — In cancer classification, only the genes which are highly contributing to the classification process are to be selected due to the problem of 'curse of dimensionality' associated with microarray based gene expression data. Biogeography-Based Optimization (BBO) is a population based evolutionary computation technique successfully applied to many application domains and proved to deliver optimal solutions. The main aim of this paper is to propose Binary Biogeography-Based Optimization Feature Selection methodology for optimal selection of genes. The Artificial Neural Network classifier is applied for cancer classification using the selected genes. The proposed method is validated through experiments on standard gene expression dataset benchmarks. The proposed method yields better results when compared to Improved Binary PSO for feature selection and its modified versions from literature in terms of classification accuracy and optimal gene selection count.**

Keywords - gene selection; cancer classification; biogeography-based optimization; artificial neural network

I. Introduction

Machine learning is one of the prominent areas of research in computer science that specifically deal with whether a program can learn with experience. A learning system is expected to automatically improve its performance as it gains more experience on a specific task. Machine learning algorithms can be classified mainly into three categories: supervised learning, unsupervised learning and reinforcement learning. One of the main important tasks in supervised machine learning is classification, which involves learning a model to correctly predict the class membership of each instance when a set of instances are represented by features or attributes and corresponding class labels[1]. A learnt classifier (model) is needed for classification. The classifier is learnt by a supervised learning (classification) algorithm which uses a set of examples to learn a classifier that is expected to correctly predict the class label of new (unseen) examples. An example of such a supervised learning system is a cancer classification system where the classifier is learnt from patient records of known types of cancer which in turn, used to predict the type of cancer a new patient is suffering from.

Cancer is one of the deadliest diseases that take many lives every year globally. In recent years, due to the advent of microarray technology, classification and diagnosis of cancer got simplified substantially. One of the prominent activities in scientific research using microarray cancer datasets is class prediction. Class prediction focuses on mapping the gene expression profiles to specific classes. The ultimate goal of this work is to use machine learning techniques to perform as accurate cancer prediction as possible. A microarray dataset generally contains few samples of patient records, where each of the samples is represented by expression levels of thousands of genes. However, the small sample size compared to huge gene count of microarray data leads to a complex problem during classification, popularly known as 'curse of dimensionality' which obstruct to build better predictive models and to achieve higher prediction performance. Hence, selection of a small but informative gene subset is imperative in microarray based cancer classification.

Gene selection (also called feature selection) aims to overcome the problem of 'curse of dimensionality' by reducing the irrelevant and redundant genes (features) [2]. Most of the feature selection methods fall under one of the two major categories viz., filter approach and wrapper approach. Filter methods rank each features individually applying a ranking criterion while wrapper methods rely on a learning algorithm to evaluate goodness of feature subsets. Wrapper methods deliver more accuracy while filters are faster. There are hybrid methods which exploit the goodness of both filter as well as wrapper.

In recent years, many population-based Evolutionary Computation (EC) techniques such as Particle Swarm Optimization (PSO) [3], Genetic Algorithms (GA) [4] etc. are applied in the domain of gene selection. One such population based EC technique for global optimization is BBO, proposed by Dan Simon [5]. BBO have been successfully applied to many application domains such power systems, scheduling problems etc. and proved to produce better solutions. While BBO has enough potential to converge towards optimal solution quickly and being newer to GA, PSO etc., BBO is rarely applied to the domain of gene selection problems. In this work, BBO is applied to gene selection for cancer classification. This paper is an extended version of our paper presented at UKSim2018 [6] with new results obtained on more datasets after fine tuning of the parameter settings of BBBOFS.

TABLE I. NOTATIONS AND ACRONYMS

Symbol	Meaning
BBO	Biogeography-Based Optimization
BBBO	Binary Biogeography-Based Optimization
ANN	Artificial Neural Network
EC	Evolutionary Computation
PSO	Particle Swarm Optimization
GA	Genetic Algorithms
HSI	Habitat Suitability Index
SIV	Suitability Index Variables
λ	Immigration Rate
μ	emigration rate
ES	Ecosystem Size
GL	Generation Limit
PD	Problem Dimension
EL	Elites to keep
MP	Mutation Probability
p	Normalization factor
DLBCL	Diffuse Large B-Cell Lymphoma
SRBCT	Small Round Blue Cell Tumor
IBPSO [3]	Improved (modified) binary PSO
IBPSO [10]	Improved binary PSO
IG-GA [4]	Information Gain-Genetic Algorithm
BBBOFS	Binary BBO based Feature Selection (Proposed Method)

The various notations and acronyms that appear throughout this paper are listed along with their meaning in TABLE I. The rest of this paper is organized as follows: in Section II the methodology is presented. Section III presents datasets used and experimental results with discussions. Section IV summarizes this paper by providing its main conclusions.

II. METHODOLOGY

A. Microarry Cancer Datasets

DNA microarray technology is based on the process of hybridization [7]. A photographic film that is sensitive to radiation can be used to visualize the hybridization. The amount of mRNA determines the amount of radiation captured on the photographic film. A microarray is a chip of solid surface called gene chip. Strands of polynucleotide called probes are attached on that surface in specific positions. A proper hybridization is achieved when each probe binds to a quantity of labeled target that is proportional to the level of expression of the gene represented by that probe. This quantity is then read by using a detector (usually a fluorescent microscopy scanner). The detector illuminates the solid surface with laser light to read the quantity by measuring the intensity of the fluorescence over each probe on the array and saves the output in the form an image. A numerical reading of the expression level is then obtained by analyzing the image using image processing algorithms. Such numerical readings of the expression levels of genes are saved in the form of commonly used file formats such .txt, .arff, .mat etc. are known as microarray gene expression datasets. Microarray cancer datasets of various cancer types are available to the research community through a number of public microarray data repositories such as Gene Expression Omnibus, Stanford Microarray Database etc. Six public dataset benchmarks of microarray gene expression data are used in this work to validate the proposed method are briefly summarized in Section III.

B. Artificial Neural Network Classifier

Artificial Neural Networks (ANN) are very efficient and popular classifiers inspired by biological nervous system. ANNs are of various types like feedforward, feedback, radial basis functions etc. based on their architectural arrangement of the neurons. Feedforward networks are simplest among these, consists of an input, a hidden and an output layer of neurons. A feedforward network can map the inputs to the ouputs for any problem if the input is finite. One of the Matlab® implementations of feedforward networks is *feedforwardnet*. Another specialized implementation for pattern recognition problems is *patternnet*. In this paper *patternnet* is used which implicitly trains the generic *feedforwardnet* for classifying the inputs to the target classes. A *patternnet* function is of the form: *patternnet(hiddenSizes,trainFcn,performFcn)* where *hiddenSizes* represent hidden layer sizes (default = 10), *trainFcn* represent the training function (default = '*trainscg*') and *performFcn* represent the performance function (default = '*crossentropy*') and returns a pattern recognition neural network (as shown in Fig. 1) [8].

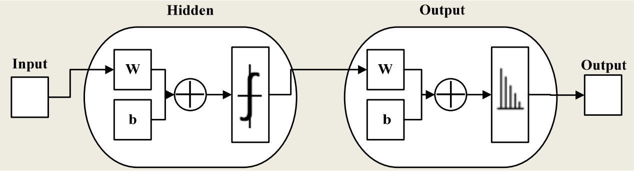

Figure 1. Architecture of a pattern reconition neural network.

C. Binary Biogeography-Based Optimization

Inspired by the biological immigration and emigration of species, Dan Simon [5] proposed a new population based evolutionary computation technique known as Biogeography Based Optimization (BBO). The population is termed as "ecosystem" formed by a number of individual species known as "habitats". Strength of a habitat is measured using habitat suitability index (HSI) which is contributed by a number of favorable conditions known as suitability index variables (SIVs). Stronger habitats have higher HSI and vice versa. Hence, stronger habitats have a tendency to transmit their characteristics which the weaker habitats (with lower HSI) eventually receive as new attributes.

The evolution process in BBO advances through application of migration and mutation on existing habitats to generate newer habitats which eventually leads to the optimal solution after a specified number of generations is reached. Migration operator regulates a habitat probabilistically proportional to the habitat's immigration rate λ and the emigration rate μ computed as follows:

$$\mu_i = \frac{Ei}{n} \tag{1}$$

$$\lambda_i = I\left(1-\frac{i}{n}\right) \tag{2}$$

Where E and I represent the highest possible value of emigration rate and immigration rate respectively, i represent the specie count of the i^{th} individual and n represent the highest species count. The procedure for computation of migration is given below.

```
Procedure1 habitat migration
for i=1 to ES do                    // ES: Ecosystem Size
        Select $H_i$ with probability $\lambda_i$
        if $H_i$ is selected then
          for $j$ = 1 to $SF$ do    // SF: Selected Features
            Select $H_j$ with probability $\mu_i$
            if $H_j$ is selected then
              Randomly select an $SIV_k$ from $H_j$
              Replace a random $SIV$ in $H_i$ with $SIV_k$
            end if
        end for
      end if
end for
```

Mutation operator uses habitat's prior probability of survival to arbitrarily change habitat SIVs with the mutation rate of m calculated as follows:

$$m(S) = m_{\max}\left(\frac{1-P_S}{P_{\max}}\right) \tag{3}$$

Where $m_{\max}$ is a bound specified by the user. Mutation is computed using the following procedure.

```
Procedure2 mutation
for $i$ = 1 to $ES$ do
        for $j$ = 1 to $SF$ do
          Use $\lambda_i$ and $\mu_i$ to compute probability $p_i$
          Select $SIV\ H_i(j)$ with probability $p_i$
          if $H_i(j)$ is selected then
            Replace $SIV\ H_i(j)$ with randomly generated $SIV$
          end if
        end for
end for
```

Mutation promotes newness and variety in ecosystem but involves the risks of ruining better solutions. Hence, in every generation of BBO, best solutions are kept saved as elites to ensure recovery of any possible wreckage caused to their HSI by mutation.

To extend the use of BBO to feature selection problems, Li and Yin [9] proposed a binary coding scheme known as binary BBO (BBBO) where they proposed to use a new binary mutation operation while using the same migration operator of standard BBO. The procedure for binary mutation is outlined below.

```
Procedure3 binaryMutation
    for $i$ = 1 to $ES$ do
      for $j$ = 1 to $SF$ do
        Use $\lambda_i$ and $\mu_i$ to compute probability $p_i$
        Select $SIV\ H_i(j)$ with probability $p_i$
        if $H_i(j)$ is selected then
          Replace $SIV\ H_i(j)$ with $SIV\ 1 - H_i(j)$
        end if
      end for
    end for
```

In binary mutation, an SIV is replaced with 1 – SIV rather than replacing with a randomly generated one as in case of standard BBO. This BBBO is applied in this work with a new objective function to propose a gene selection method for classification of cancer, as described in the next subsection.

D. Proposed Method: Binary BBO based Feature Selection (BBBOFS) for Cancer Classification

In this paper, BBBO is applied to propose a gene selection method, named as BBBOFS, for classification of cancer using ANN classifier. Initially mutual information is used to select sixty numbers of genes. Then, the selected genes are passed through a proposed wrapper built using binary BBO as the search strategy with the new objective function and ANN classifier as the evaluator. Here the objective is to reduce the Mean Squared Error (MSE) of ANN and at the same time minimize the selected gene count.

Mean Squared Error (MSE) of ANN for an individual H_i is calculated as follows:

$$MSE(H_i, \mathrm{O}_i, \mathrm{ANN}) = \frac{1}{N}\sum_{i=1}^{N}(\mathrm{T}_i - \mathrm{O}_i)^2 \tag{4}$$

Where T_i is the target output and O_i is the obtained output for individual H_i by the ANN with N output neurons.

Accordingly the proposed objective function used to calculate the fitness of an individual H_i can expressed as follows:

$$Fitness(H_i) = MSE(H_i, \mathrm{O}_i, \mathrm{ANN}) + pC(H_i) \tag{5}$$

Where $MSE(H_i, \mathrm{O}_i, \mathrm{ANN}) \in [0,1]$ is the Mean Squared Error evaluated by ANN using the genes in the gene subsets of individual H_i. $C(H_i)$ is the number of selected genes in habitat H_i. 'p' is an user given constant used to normalize the value of $C(H_i)$ also to the range [0,1].

Each individual is represented using a binary (0/1) bit string of features; 1 indicates that a gene (feature) at that particular position is included and 0 indicates that gene is not included in the individual. Fitness of that individual is then evaluated applying the proposed objective function (5). The first part of the objective function i.e., the classification error rate of ANN evaluator is computed with 70:30 training:test split ratio of the dataset. That is 70 percent of the randomly selected samples are used to train the neural network, remaining 30 percent of the samples are equally divided to validate and test the trained network. Finally, the best subset of selected genes returned by the wrapper is used for classification by ANN classifier with the same train-test percentage split ratio.

The process of the proposed system is illustrated in flowchart as shown in Fig. 2 and the major steps are detailed as below:

1: Data loading: The gene expression data are loaded.

2: Mutual information selector: The sixty top genes with the highest scores are selected as the gene subset.

3: BBBO and ANN wrapper:

(i) Subset search: In this step, the algorithm searches for better solutions by the migration and mutation model.

(ii) Subset evaluation: The objective function in (5) is evaluated using ANN classifier.

5: Stopping condition: The final feature subsets are selected, and then output the final best feature subset.

6: Classification: The best subset of selected genes returned by the wrapper is used for classification by ANN classifier.

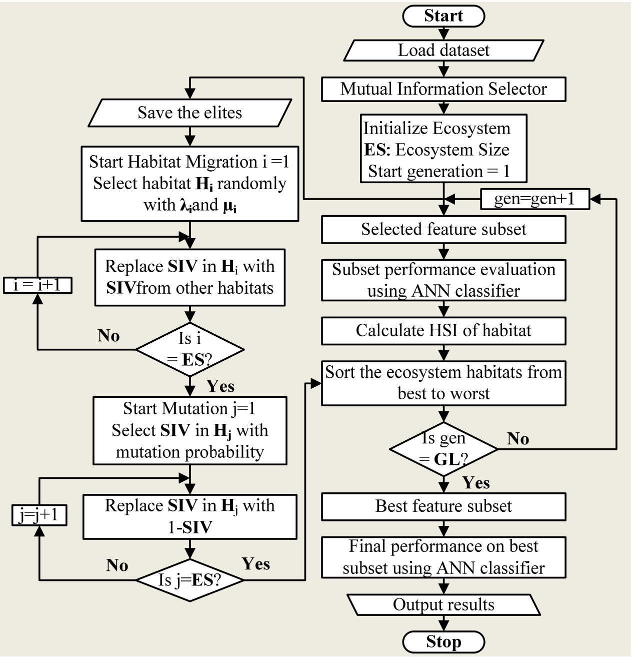
Start
Load dataset
Mutual Information Selector
Initialize Ecosystem
ES: Ecosystem Size
Start generation = 1
gen=gen+1
Selected feature subset
Subset performance evaluation using ANN classifier
Calculate HSI of habitat
Sort the ecosystem habitats from best to worst
Is gen = GL?
No
Yes
Best feature subset
Final performance on best subset using ANN classifier
Output results
Stop
Save the elites
Start Habitat Migration i =1
Select habitat H_i randomly with $\lambda_i$and μ_i
Replace SIV in H_i with SIVfrom other habitats
i = i+1
No
Is i = ES?
Yes
Start Mutation j=1
Select SIV in H_j with mutation probability
Replace SIV in H_j with 1-SIV
j=j+1
No
Is j=ES?
Yes

Figure 2. Architecture of the cancer classification system using the proposed gene selection method.

III. Results and Discussions

A. Datasets and Experimental Setup

In the present study, four high dimensional microarray gene expression cancer dataset benchmarks (downloaded from gems-systems.org [10]) are used for the experiments which include both binary and multiclass gene expression data of tumor samples, brain tumor, lung cancer and prostate tumor samples. These datasets are summarized in TABLE II. In the experiments the datasets are randomly portioned into 70 percent training samples, 15 percent validation and 15 percent test samples. Experiments are conducted on a standalone PC with Intel i3 CPU and 3GB of RAM with the implementations been carried out in Matlab® 2014a programming environment. The best gene subset returned by the proposed BBBOFS method is used to find out the classification accuracy using the ANN classifier with 70:15:15 training versus validation and test ratio of samples. Classification accuracy and the selected gene count are the two criteria considered for the evaluation of the performance of BBBOFS. The various parameter values of BBBOFS are listed in TABLE III.

TABLE II. DATASETS SUMMERY

Dataset Name	Description		
	Samples Count	*Gene Count*	*Class Count*
9_Tumors	60	5,726	9
Brain_Tumor1	90	5,920	5
DLBCL	77	5,469	2
Lung_Cancer	203	12,600	5
Prostate_Tumor	102	10,509	2
SRBCT	83	2,308	4

TABLE III. PARAMETER SETTING FOR BBBOFS

Parameter Name	Values
Ecosystem Size	100
Generation Limit	300
Problem Dimension	50
Elites to keep	3
Mutation Probability	0.05
Normalization factor	0.02

TABLE IV. EXPERIMENTAL RESULTS

Dataset Name	Results		
	Accuracy % (#Acc)	*Selected Gene Count (#SGC)*	*Percentage of Selected Genes*
9_Tumors	86.21	24	0.42
Brain_Tumor1	95.56	21	0.35
DLBCL	100	8	0.15
Lung_Cancer	99.51	15	0.12
Prostate_Tumor	99.02	7	0.07
SRBCT	100	7	0.30
Average	***96.72***	***13.67***	***0.24***

B. Experimental Results and Discussions

The results of experiments of BBBOFS on the six datasets are listed in TABLE IV. These are results obtained in a single run with ANN classifier with train:test percentage split method. *#Acc* denotes the classification accuracy in percentage returned by ANN classifier with the selected gene subset and *#SGC* denotes the Selected Gene Count i.e., the total number of genes present in the selected gene subset. The strength of the proposed method may be highlighted under the following four headings in the light of the obtained results of experimentations.

(i) Efficiency of Selection:

It can be observed from TABLE IV that the classification accuracies of more than 95% is achieved for all the datasets except for 9_Tumors (86.21%). It can also be observed that the average percentage of genes selected by BBBOFS is 0.24% which is sufficient to achieve more than 96% average accuracy over all the datasets. This establishes the fact that only a small fraction of genes is necessary for correct prediction of cancer, while most of the genes are irrelevant to the prediction process. Rather presence of these irrelevant genes reduces the prediction performance. Further the increase in classification accuracy indicates that for high dimensional datasets (gene expression cancer data) BBBOFS is appropriate for selecting a small number (subset) of genes.

(ii) Diversity of Search:

A higher classification accuracy means a lower classification error rate. This means BBBOFS can reduce the MSE to the desired level. The fitness value of BBBOFS lowers substantially after few generations on all the datasets which proves that BBBOFS does sufficient exploration to find a better solution (as shown in Fig. 3).

(iii) Uniformity of Prediction:

The ROC curves for the datasets are shown in Fig. 4. It can be observed from Fig. 4 that the results of BBBOFS are well balanced and uniformly distributed among the class labels in all the datasets.

(iii) Stability of Model:

The final ANN classifier model built using the best gene subset returned by BBBOFS is stable enough which is established by the cross-entropy plots across the epochs of training, validation and test (as shown in Fig. 5). In all the datasets the stable pattern is observed for training, validation as well as for test.

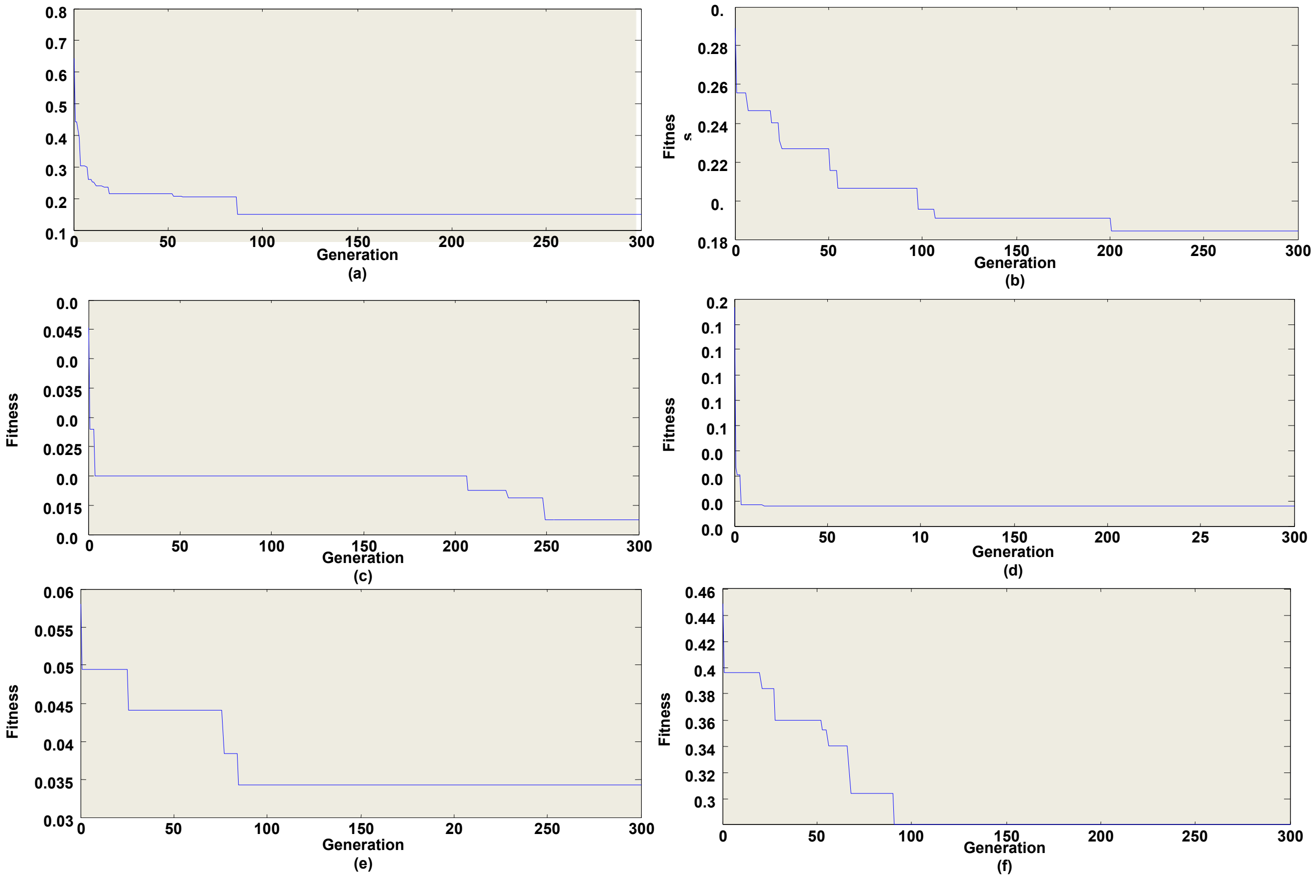

Figure 3. Relation between fitness values and number of generations for BBBOFS: (a) 9_Tumors (b) BrainTumor_1 (c) DLBCL (d) Lung_Cancer (e) Prostate_Tumor and (f) SRBCT.

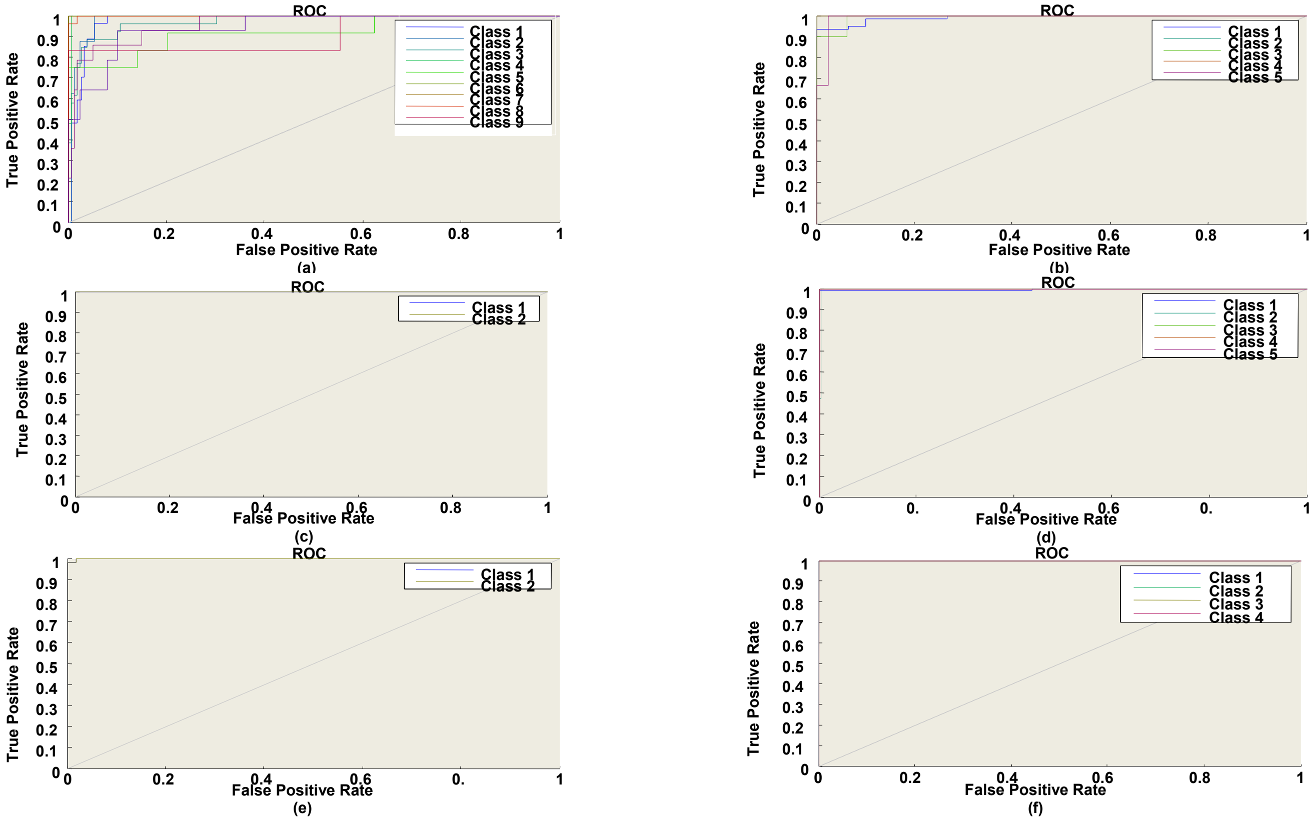

Figure 4. ROC Curves: (a) 9_Tumors (b) BrainTumor_1 (c) DLBCL (d) Lung_Cancer (e) Prostate_Tumor and (f) SRBCT.

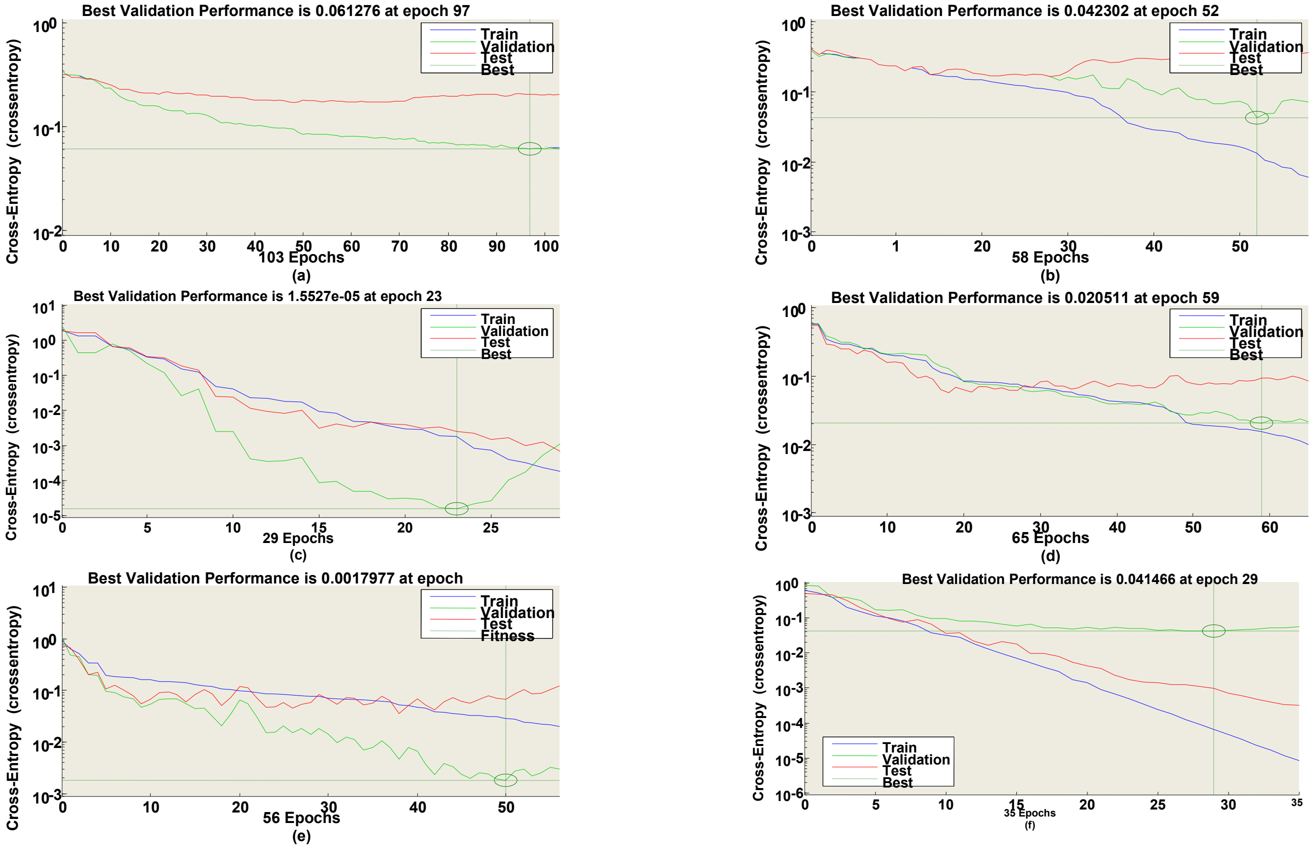

Figure 5. Validation Performance Curves of ANN: (a) 9_Tumors (b) BrainTumor_1 (c) DLBCL (d) Lung_Cancer (e) Prostate_Tumor and (f) SRBCT.

A comparison of BBBOFS with other state of the art evolutionary techniques of feature selection viz., IBPSO [3], IBPSO [11], BPSO [3] and IG-GA [4] is presented in TABLE V. Best results are shown bold faced. Results of average of a number of runs are shown in italic. BBBOFS achieves higher classification accuracies than all other previous methods under comparison in all the six datasets and the number of selected genes is also smaller except for the results reported in the work [3]. The number of selected genes of work [3] is slightly smaller for Brain_Tumor1 dataset. By and large, BBBOFS has produced better results than the previous related works in terms of classification accuracy and number of selected genes.

TABLE V. COMPARISON OF BBBOFS WITH OTHER METHODS

Dataset Name	***Measure***	**Methods**				
		BBBOFS (present work)	***IBPSO [3]***	***IBPSO [10]***	***BPSO[3]***	***IG-GA [4]***
9 - Tumors	*#Acc*	**86.21**	*75.50*	78.33	*77.33*	85
	#SGC	**24**	*240.6*	1280	*236*	52
Brain_Tumor1	*#Acc*	**95.56**	*92.56*	94.44	*92*	93.33
	#SGC	21	***11.2***	754	*236*	244
DLBCL	*#Acc*	**100**	*100*	100	*100*	100
	#SGC	8	***6***	1042	*230.1*	107
Lung_Cancer	*#Acc*	**99.51**	*95.86*	96.55	*96.6*	95.57
	#SGC	**15**	*22.3*	1897	*228.7*	2101
Prostate_Tumor	*#Acc*	**99.02**	*97.94*	92.16	*98.04*	96.08
	#SGC	**7**	*13.6*	1294	*231.5*	343
SRBCT	*#Acc*	**100**	*100*	100	*100*	100
	#SGC	**7**	*17.50*	431	*221.3*	56

#Acc and #SGC denote the classification accuracy and the selected gene count respectively.

IV. CONCLUSION

This paper proposed Binary Biogeography-Based Optimization Feature Selection methodology for optimal selection of genes. The parameters used in the feature selection methodology are fine tuned so as to extract smaller set of genes which are needed with higher classification accuracies compared with other previous methods reported. This is established by the experimental results which demonstrate that he proposed algorithm can obtain the higher accuracy in all the six datasets and smaller number of genes selected in four of the six microarray datasets. Overall, the present work has outperformed the previous related works in terms of classification accuracy and number of selected genes.

REFERENCES

[1] L. Breiman, J. Friedman, R. Olshen, and C. Stone, Classification and regression trees. Wadsworth International Group, 1984.

[2] M. Dash and H. Liu, "Feature selection for classification," Intelligent Data Analysis, vol. 1, no. 4, pp. 131–156, 1997.

[3] M. S. Mohamad, S. Omatu, S. Deris, and M. Yoshioka, "A modified binary particle swarm optimization for selecting the small subset of in-formative genes from gene expression data," IEEE Trans. Inf. Technol. Biomed., vol. 15, no. 6, pp. 813–822, 2011.

[4] Yang, C.H., Chuang, L.Y., Yang, C.H.: IG-GA: A Hybrid Filter/Wrapper Method for Feature Selection of Microarray Data. Journal of Medical and Biological Engineering, vol. 30, pp. 23-28 (2010).

[5] D. Simon, "Biogeography—Based optimization," IEEE Trans. Evol. Comput., vol. 12, no. 6, pp. 702–713, 2008.

[6] D. H. Mazumder and V. Ramachandran, "Binary Biogeography-Based Optimization Applied to Gene Selection for Cancer Classification Using Artificial Neural Network", Proceedings of the twentieth IEEE UKSim-AMSS International Conference on Modelling & Simulation, Vol. 20, pp. 43–48, Cambridge University, UK, 2018. DOI 10.1109/UKSim.2018.00020.

[7] Knudsen, S. (2006). Cancer Diagnosis with DNA Microarrays. Wiley.

[8] https://in.mathworks.com/help/nnet/ref/patternnet.html.

[9] X. Li and M. Yin, "Multi-objective binary biogeography based optimization for feature selection using gene expression data," IEEE Trans. Nano Biosci., vol. 12, no. 4, pp. 343–353, Apr. 2013.

[10] http://www.gems.systems.org.

[11] L. Y. Chuang, H. W. Chang, C. J. Tu, and C. H. Yang, "Improved binaryPSO for feature selection using gene expression data," Comput. Bio.Chem., vol. 32, pp. 29–38, Feb. 2008.