Available online www.jocpr.com

Journal of Chemical and Pharmaceutical Research, 2015, 7(2):163-168

Research Article

ISSN : 0975-7384
CODEN(USA) : JCPRC5

An ELM-Wrapped GA based multiobjective feature selection for identifying cancer-microRNA markers

Keerthana S., Hemalatha R. J., Hari Krishnan G. and Umashankar G.

Faculty of Bio and Chemical Engineering, Department of Biomedical Engineering, Sathyabama University, Chennai, Tamil Nadu, India

---

ABSTRACT

MicroRNAs (miRNAs) take part a significant role in cancer development and also act as a vital feature in several other diseases. Previously a standard classifier method like SVM classifier exploited for selecting promising miRNAs encompass differential expression in benign and malignant tissue samples. Consequently, the non-dominated sets of capable miRNAs are combined into a single most promising miRNA subset. On the other hand, the most important drawback of such learning techniques is slow learning time. With the aim of overcoming these problems of conventional learning techniques, in this work Extreme Learning Machine classifier is formulated for deciding promising MiRNAs because it only needs modification of one parameter. The performance has been demonstrated on four real-life miRNA expression datasets for ELM and the identified miRNA markers are reported. The experimental results demonstrate that the proposed ELM method outperforms the standard methods.

Key words: MicroRNA marker, Genetic Algorithm, Extreme Learning Machine

---

INTRODUCTION

MicroRNAs (miRNAs) are a new class of small non-coding regulatory RNAs that are engaged in process of regulating gene expression at the posttranscriptional level. These tiny (18-24 nucleotides in length) RNA molecules standardize several biological processes [1] [2].miRNAs, normally transcribed by RNA polymerase II, are primarily made as huge RNA precursors, called pri-miRNAs [5]. Transcription of miRNA genes are regulated by means of the modulation of numerous transcription factors as that of protein-coding genes [6]. miRNAs and their targets seem to generate a complex regulatory network.. There are roughly one third of all human protein-coding genes that are controlled by miRNA in accordance with the computational predictions [7]. Several investigations revealed that miRNA expression appeared to be deregulated in cancer versus normal tissue [8] [9].

Since those initial studies, examples of miRNA deregulation have been revealed in chronic lymphocytic leukemia [10], B-cell lymphoma [11] [12] and breast cancer [13] [14]. The area of analyzing miRNA microarrays has obtained much attention in recent times. One of the major complications in the analysis of miRNA functions was the nonexistence of techniques for quantitative expression profiling. It is feasible to employ the existing marker selection techniques utilized in gene expression studies for miRNA expression data also. On the other hand, miRNA expression datasets have certain features which might require to be taken into account at the time of applying such techniques for miRNA microarray datasets. In most of the cases, the expression profiles of miRNAs obtained from microarray experiments are tissue-specific in nature.

In this paper, a multi objective Genetic Algorithm-based feature selection approach is implemented that encodes a possible feature subset in its chromosomes. Non-dominated Sorting Genetic Algorithm-II (NSGA-II) [15] [16], a well-known multiobjective GA has been exploited as the underlying optimization tool. The fitness of the chromosomes has been assessed by means of Extreme Learning Machine (ELM) classifier, and three objective

functions have been concurrently optimized. The objective functions taken here are the number of features, specificity, and sensitivity. The initial objective is diminished and the other two objectives are maximized. Consequently, the most promising miRNAs accountable for making a distinction among the normal and malignant classes are obtained by a third measure (called -measure). The performance of this method has been assessed on openly available miRNA expression datasets of six dissimilar tissue samples viz., breast, colon, kidney, lung, prostate and uterus. The experimental results establish the effectiveness of the proposed approach. Initially, the experiments have been performed for discovering miRNA markers that make a distinction among the normal and malignant samples globally for all categories of tissue samples. Then, the biological significance tests have been carried out for the selected markers. The result shows that the proposed ELM based classifier is more effective in terms of its accuracy, Sensitivity, Specificity, AUC and F-measure than the existing SVM algorithm.

The rest of the article is organized as follows: The next section provides related work of cancer-miRNA. In Section III, the proposed method is described in detail with the data description and preprocessing. Section IV reports the experimental results. Finally, Section V concludes the article.

EXPERIMENTAL SECTION

Datasets and Preprocessing

A publicly available miRNA expression dataset is acquired from the following website: http://www.broad.mit.edu/cancer/pub/miGCM. The entire dataset includes 251 mammalian miRNAs from several cancer categories. From this website, six datasets were extracted, includes the samples from breast, colon, kidney, lung, prostate, and uterus. Each dataset is described by the entire 251 miRNAs. Every sample vector of the datasets is standardized to have mean 0 and variance 1. The final dataset includes two classes, one indicating the entire normal samples (32 samples) and another indicating the entire tumor samples (57 samples). For preliminary filtering of miRNAs, signal-to-noise ratio (SNR) is employed. Initially for each miRNA, the SNR is calculated. SNR is defined as,

$$SNR = \frac{\mu_1 - \mu_2}{\sigma_1 + \sigma_2}$$

Where μ_i and σ_i, $l \in \{1,2\}$, represent the mean and standard deviation of class l for the equivalent miRNA. It is to be observed that larger absolute value of SNR for a miRNA points out that the miRNA's expression level is elevated in one class and small in another. Therefore this bias is extremely helpful in distinguishing the miRNAs that are expressed differently in the two classes of samples. Subsequent to computing the SNR value of each miRNA with regard to normal and cancerous classes, the miRNAs are organized in descending order of the absolute SNR values.

Then selected from the top miRNAs having absolute SNR values higher than or equal to the mean of all absolute SNR scores. This provides 100 miRNAs which are employed further. The dataset is then arbitrarily segmented into training and test sets with nearly equal distribution. On the other hand, at the time of segmenting into training and test sets, it is ensured that both training and test set include at least one sample from normal and cancerous samples of each of the tissue categories. After making sure this, 40 training samples and 49 test samples were obtained. The feature selection approaches are executed only on the training set. Several real-life optimization complications are multiobjective in nature. Unlike single objective optimization, numerous objectives are concurrently optimized in multiobjective optimization (MOO). Among the available MOO methods, the Genetic Algorithm (GA) dependent approaches like Non-dominated Sorting GA-II (NSGA-II), Strength Pareto Evolutionary Algorithm (SPEA) and SPEA2, Pareto Archived Evolutionary Strategy (PAES) are extremely popular. NSGA-II is an enhancement over its earlier version of NSGA based on computation time. In [17], it has been exposed that NSGA-II executes better compared to several other MOO approaches. Therefore the multiobjective feature selection method is taken here which employs NSGA-II as a principal multiobjective framework. At last, the test set is categorized by the trained ELM on the chosen miRNAs and the classifier performance is reported. It is to be pointed out that the test set is entirely disjoint with the training set.

An ELM-Wrapped GA based Multiobjective Feature Selection

The proposed approach includes two stages. Initially, a multiobjective feature selection approach wrapped with ELM classifier is utilized. Then, the chosen miRNAs in different solutions of the non-dominated set are employed to acquire a single set of most promising miRNAs that make a distinction among the two classes of tissue samples.

ELM Learning Algorithm

Extreme Learning Machine (ELM) is a Single-hidden Layer Feed Forward Neural Network (SLFNN) which arbitrarily picks input weights and hidden neuron biases without any training. The outputs weights are methodically decided using the norm least-square solution and Moore-Penrose inverse of a general linear system, consequently letting a considerable training time reduction. The activation function such as sine, Gaussian, sigmoidal etc., can be selected for hidden neuron layer and linear activation functions for the output neurons. The SLFNN assessed here employs additive neuron design in place of kernel based, thus random parameter selection. Investigations done by Huang et al [18] confirmed that single layer feed forward neural network with arbitrarily assigned input weights and hidden layer biases and with almost any nonzero activation function can commonly approximate any continuous functions on any input data sets. Huang et al [19] put forwarded an alternate approach to train a SHLFN by finding a least square solution β′ of the linear system. The unique minimum norm least square (LS) solution is modelled as

$$\hat{\beta} = H^{\dagger}T$$

where $H^{\dagger}$ represents the MP generalized inverse of matrix H. As investigated by Huang, ELM using such MP inverse technique tends to acquire excellent generalization performance with significantly increased learning speed. The summarization of the ELM algorithm can be as follows:

Provided a training set $N = \{(x_i, t_i) | x_i \in R_n, t_i \in R_m, i = 1, \ldots, N\}$, kernel function $f(x)$, and hidden neuron $\tilde{N}$.

Step 1: Choose appropriate activation function and number of hidden neurons $\tilde{N}$ for the particular problem.

Step 2: Allocate arbitrary input weight w_i and bias b_i $i = 1,2,..,H$

Step 3: Compute the output matrix H at the hidden Layer $H = f \cdot (w \oplus x + b)$.

Step 4: Compute the output weight $\hat{\beta} = H^{\dagger}T$.

Feature Selection Using NSGA-II

A NSGA-II based feature selection approach has been formulated that is wrapped with ELM classifier. In this approach, every chromosome in the population is a binary string having two elements. The first element is of length equal to the number of miRNAs (r) in the dataset. For a chromosome, bit "1" points out that the equivalent miRNA is selected, and bit "0" points out that the equivalent miRNA is not selected. The second element of the chromosome is of length k and it encodes the value of ELM regularization parameter C in binary. The decimal value encoded in bits is mapped in the range [0,100] to acquire the parameter C. Three objective functions are optimized concurrently. For the purpose of computing the objective values for a chromosome, initially the subset of miRNAs that are encoded in the chromosome are obtained. Consider this set is indicated as S. It contains those miRNAs for which the bit position of the chromosome has value "1." The samples in the training set are categorized on the subspace S by means of leave-one-out cross validation by ELM with the intention of finding out the objective function values equivalent to the chromosome. Based on the output of the cross validation, the number of true positives (tp), false positives (fp), true negatives (tn) and false negatives (tn) are figured out. The first objective function is the sensitivity which is given as:

$$f_1 = Sensitivity = \frac{tp}{tp + fn}$$

The second objective f_2 is the specificity which is calculated using the following formula:

$$f_2 = Specificity = \frac{tn}{tn + fp}$$

The third objective is the amount of selected features which is found using:

$$f_3 = |S|$$

This scheme makes an attempt to find the smallest set of miRNAs that accurately classify the benign and malignant tissue samples. In the final generation, a set of non-dominated solutions each encoding a promising features (miRNA) subset is obtained. For the purpose of selection, crowded binary tournament selection method is employed. Subsequent to selection, uniform crossover has been implemented on the chromosomes and later bit-flip mutation is implemented to produce the next generation. Elitism has been integrated to track the fine chromosomes found so far. Elitism is carried out by integrating parent and child population and transferring the non-dominated solutions from the integrated population to the next generation. The process of fitness computation, selection, crossover, and mutation is done for a particular number of generations and the final generation generates a set of non-dominated solutions.

For the purpose of selecting the most promising feature subset that give better values for both specificity and sensitivity, the solution that offers the most excellent F-measure value and precision is selected. Higher value point outs better balance among sensitivity and specificity and consequently point outs better classification. The feature subset encoded in the solution providing the best F-measure is taken as the final set of miRNA markers. Subsequent to selecting the miRNA markers from the training set, then the test samples are classified in accordance with the selected miRNA markers.

RESULTS AND DISCUSSION

In this section, initially the performance of the proposed approach to find out the miRNA markers is analysed. Subsequently, the biological importance of the discovered miRNA markers is analysed. The proposed NSGA-II-based feature selection technique and classification is executed with parameters shown in Table.1.

Table.1.Parameters for proposed work

Parameter	size	Origin	Offset
Training	283 *251	Micro RNA	0
Validation			
Optimization	Genetic with ng=10 np=140		
Selection	151		+- 20
Test	100		+100, 200

The proposed multiobjective feature selection approach is implemented on the pre-processed training dataset for multiple times and for each run, the output set of features is gathered. The classifier performance is measured using the following parameters like accuracy, specificity, sensitivity, F- measure. The wrapper ELM-NSGA-II algorithm creates better accuracy rate, produces high sensitivity, less specificity and high F- measure.When the number of features increases the accuracy of the result is increases

S. No	Parameters	Values
1	Accuracy	84
2	Sensitivity	93
3	Specificity	83
4	Area under curve	0.98
5	F- Measure	91.2

Table 2: Classifier Performance value

The comparative graph of all the parameters is given below.

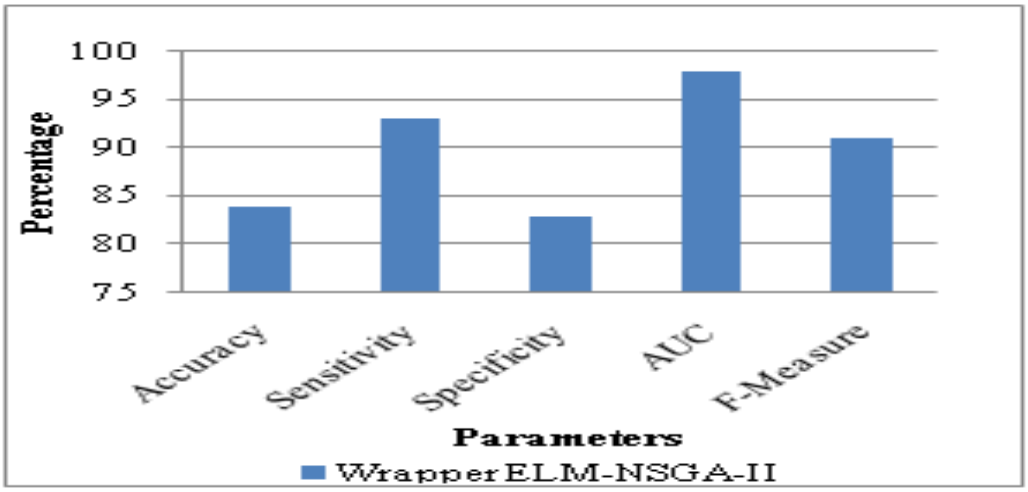

Figure 1(a)

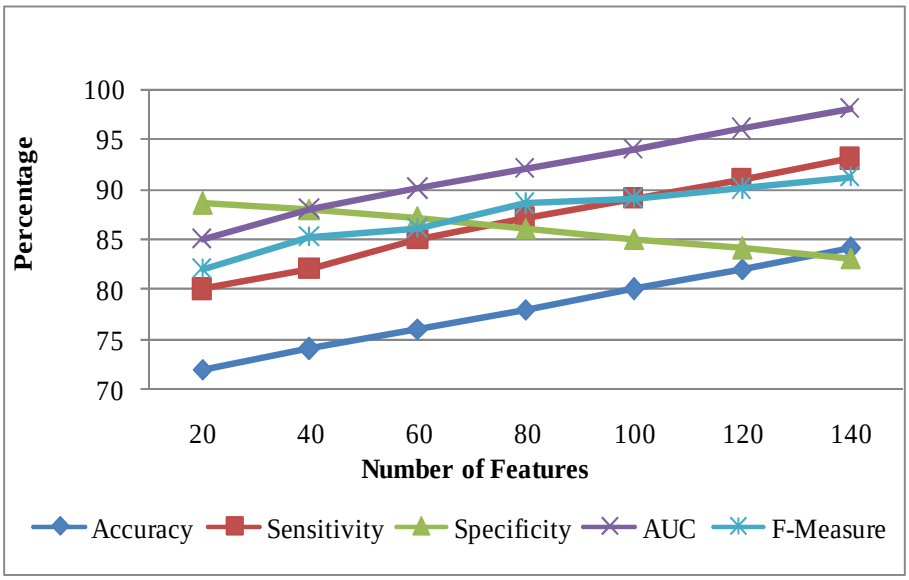

Figure 1(b)

Figure 1(a) and Figure 1(b) gives the graphical representation of classifier performance. The wrapper ELM-NSGA-II algorithm gives better result than the existing methods.

CONCLUSION

In this work, a multiobjective GA-based feature selection approach wrapped with ELM classifier has been formulated for identification of miRNA markers from miRNA expression datasets. This approach optimizes different performance criteria at the same time and evolves the required subset of features (miRNAs). ELM offers an improved computational framework for the classification. Results on real-life miRNA expression datasets of several tissue categories, viz., Breast, Colon, Kidney, Lung, Prostate, and Uterus, have been demonstrated. Furthermore, several recognized miRNA markers are also found to have association with different categories of cancer as per current literatures. Based on the experimental results, the proposed approach is high effective than the existing approaches in terms of computational complexity. This approach takes less computational time of 10 minutes to train and test the dataset than the SVM. The time required for the process is 0.9 seconds, In future, performance of different well-known classifiers, other than ELM, is to be investigated with the use of Swarm intelligence technique.

Acknowledgment
Authors are grateful to the management of Sathyabama University, Faculty of Bio and Chemical Engineering, Department of Biomedical Engineering, Chennai, India for providing required facilities to complete the research successfully

REFERENCES

[1] D.P.Bartel; *Cell*; **2009**, 136(2), 215–33.
[2] W. Filipowicz; S.N.Bhattacharyya; N. Sonenberg; *Nat Rev Genet*, **2008**, 9(2), 102–14.
[3] A.Esquela-Kerscher ; F.J Slack , *Nat Rev Cancer*,**2006**,6(4),259–69.
[4] C. Llave et al., *Science*, **2002**, 297(5589), 2053–6.
[5] X.C Ding, J. Weiler, H .Grosshans, *Trends Biotechnology*, **2009**, 27(1), 27–36.
[6] K.A. O'Donnell et al , *Nature*,**2005**,435(7043),839–43
[7] B.P Lewis, C.B Burge, D.P Bartel, *Cell*, **2005**,120(1),15–20
[8] G.A Calin, C.M. Croce; *Nat Rev Cancer*, **2006**,6(11),857–66.
[9] J. Lu et al, *Nature*, **2005**,435(7043),834–8.
[10] G.A. Calin et al, *Proc Natl Acad Sci U S A*, **2002**, 99(24), 15524–9.
[11] L He. et al; Nature. **2005**,435(7043),828–33.
[12] A. Ota et al, *Cancer Res*, **2004**, 64(9),3087–95.
[13] M.V. Iorio et al, *Cancer Res*, **2005**,65(16),7065–70.
[14] L.F. Sempere et al, *Genome Biol*,**2004**,5(3),R13.
[15] K. Deb; A. Pratap; S. Agrawal; T. Meyarivan, *IEEE Trans. Evol. Comput*,**2002**, 6, 182–197.
[16] S. Bandyopadhyay; R. Mitra; U. Maulik; M. Q. Zhang, *BMC Silence*,**2010**,6.
[17] K. De; S. Agrawal; A. Pratap; T. Meyarivan, *Proceedings of the Parallel Problem Solving from Nature VI Conference, Paris, France*, **2000**, 849–858.
[18] G.B Huang; Q.Y Zhu; C.K Siew, Proc. *Int. Conf. Neural networks, Budapest, Hungary*, **2004**,985-990.
[19] G.B Huang; Q.Y Zhu; C.K Siew, *Neurocomputing*. **2006** ,70(1), 489-501.