INTERNATIONAL RESEARCH JOURNAL OF PHARMACY
www.irjponline.com
ISSN 2230 – 8407

Research Article

# ESBO: CANCER CLASSIFICATION IN MICROARRAY DATA USING SBO ALGORITHM WITH SELF-ADAPTIVE EMISSION RATE

R. Balamurugan *, N. Narayanan Prasanth, WI. Suresh Kumar
Associate Professor, School of Computer Science and Engineering (SCOPE), Vellore Institute of Technology, Vellore, Tamil Nadu, India
*Corresponding Author Email: balacse05@gmail.com

Article Received on: 14/08/19 Approved for publication: 12/09/19

**DOI: 10.7897/2230-8407.1010303**

**ABSTRACT**

The complete diagnosis of sub type of cancers underpins the care of individual cancer patients. In the perspective of cancer classification, gene expression data analysis has been used to more precisely classify tumors. However, selection of non-redundant but relevant genes from microarray gene expression data is computationally tough task. It has become increasingly clear that the traditional approach to cancer classification is insufficient. The major motto of this article is to originate a deterministic approach to pick the highly relevant genes from the microarray data for cancer diagnosis. This article presents a modified nature-inspired algorithm namely Stellar-Mass Black hole with Self-Adaptive Emission Rate (ESBO) to choose genes from microarray data that are able to classify various cancer sub-types with high accuracy. The experiment done and results are analyzed with five reputed Micro array benchmark datasets. The results depict that ESBO is outshines than SBO and other deterministic methods. The experiment also proves that adopting dynamic adjustment of emission rate with SBO is more effective than applying them individually.

**Keywords:** Cancer, Classification, Stellar Mass Black hole, Gene Expression data, Emission Rate, Optimization.

## INTRODUCTION

The microarray is a recently developed lab-oriented biochip technology in pharmacological treatment of diseases such as oral premalignant lesions and in cancer research. Technologists can investigate and compute the expression levels for many number of genes in a single experiment with the assist of microarray technology in a competentway.[1] Usually, Microarray technique consists of many-stages: First, the RNA is take out from the sample using a column or solvent like phenol-chloroform; Next, reverse transcription of the mRNA is done to create cDNA (complementary DNA strand); Then, in order to each cDNA gets hybridized to its complementary strand, the labeled cDNAs from both the samples are placed in the DNA microarray.[2] However, Due to curse of dimensionality gene expression dataset which makes the classification task for a given sample become more complex.[3]

Without prior knowledge, it's really challenging task to detect which genes are more useful. Supervised learning is a part of approach in machine learning.[4] One of the well-known supervised learning techniques is called Classification. It is the process of predicting the class of given data points. Usually, High dimensional gene expression data consists of a few amounts of irrelevant genes which are not useful for classification. Therefore, an effective approach of finding genes for cancer diagnosis is analytically necessary. The main objective of the gene selection process is to choose the minimum number of important genes that are more predictive in classification process. The feature selection process uses a different deterministic approach to select a subset of the relevant features, which has the most useful information, from the large gene microarray data.[5]

The widely accepted gene selection methods by researchers can be divided in to three groups as follows: filter, wrapper and embedded one. Zhang *et al*. proposed support vector machine based on recursive feature elimination and parameter optimization (SVM-RFE-PO).[6] This approach computes the feature ranking score by remove a signature attribute with the smallest ranking coefficient in each iteration, and finally obtains the decreasing order of all signature attributes. Pan *et al.* presented an ensemble learning algorithm. It uses the variable importance of random forest algorithm to sort the features and then uses the sequence backward search method.[7] Adding importance score of each feature along with mutual information for classification purpose, qualitative mutual information (QMI) method is proposed by Arpita Nagpal *et al*.[8] Most gene selection methods identify only single-class specific signature genes and cannot identify multiple-class specific signature genes easily. Next wrapper method for gene selection from microarray data is invented by Alanni *et al*. using an innovative Gene Selection Programming (GSP) method.[9]

The next Based on Gene Expression Programming (GEP) algorithm, Azzawi *et al*. proposed an innovative Structural Binary Classification (SBC) strategy for classifying lung cancer subtypes using microarray data.[10] Recently a hybrid algorithm based on interaction information for microarray-based cancer classification.[11] This method employs interaction information to rank candidate genes to add into a gene subset. Due to the high dimensionality of microarray data, and traditional gene selection algorithms are filter-based. Usually, the optimal gene selection problem is considered as Non-deterministic polynomial-time hard (NP-hard) problem.[12] The problem is even more complex when considering the large number of microarray data sets available whose properties, in terms of both the number of

features and the number of patients, can vary significantly. Therefore, taking into account the complexity of gene expression data, we directed to apply nature-mimic algorithms in order to solve this problem. Hence, we derived a new modified approach, which combine a method called Stellar Mass Black hole (SBO) optimization with self-adaptive emission rate, to choose more relevant gene subset from cancer microarray gene expression data.

## MATERIALS AND METHODS

### Problem statement

Recently, microarray technology is recognized as a high scalable technology in industry. It has been used in clinical, research and development for deep understanding of molecular mechanisms and effective treatment of complex diseases. Generally, gene expression data can be viewed as a row-column matrix. Each row directs genes and column represents sample (condition). The number of columns is very lesser than that of rows in gene expression datasets, so the results of the current algorithms are not accurate enough. There is a necessity for selection methods to pick significant genes from expression data. It takes in to find a subset of the relevant genes. Therefore, instead of classifier built from the entire set of features, a classifier built with this subset of genes would perform better. So, in order to maximize the accuracy of the classifier, the proposed methods should remove irrelevant features.

However, various heuristic methods have been proposed to extract informative and relevant cancer genes from microarray gene expression data and meanwhile reduce the number of noise and irrelevant genes as well as to achieve accurate classification. Recently to handle the high dimensions data, many of the classification methods are not scalable. It may fail to extract the useful information from the raw gene expression microarray data. Generally; overall correctness of the classifier is known as classification accuracy. Ratio between total number of classifications and sum of correct cancer classifications is called as classification accuracy. Its mathematical representation is shown below:

$$Classificationaccuracy = \frac{T}{C} \times 100 \quad (1)$$

Where; C - True positive + False positive + False negative + True negative Overall instances in the initial microarray dataset.
T = True positive + True negative- correctly classified instances.

Classifier performance is evaluated by its one of the measures is called classification accuracy. Here, the main objective of this article is to find the significant genes, which improve the classification accuracy. By maximizing the accuracy values in each iteration, Black hole with Adaptive emission rate the global optimum value could be found to be the optimum result.

### SBO Algorithm with Self Adaptive Emission rate

Due to their capability in searching for the optimal or near-optimal solutions on large and complex spaces of possible solutions, Bio-inspired evolutionary algorithms are more applicable and accurate than other deterministic gene selection methods. In recent years, most of the engineering optimization problems have been resolved by SBO method. However, for a complex optimization problem, classical SBO algorithm could not be nimble to cause convergence behavior. The proposed algorithm recommends remarkable explore and exploit search capacities in the way of named absorption and emission operators. Regardless of its unique characteristics, SBO has some limitations from an evolutionary computing point of view. Obviously, abortion and emission rate of SBO state that, to refine their current populations, it implements a series of random walks around some of the best solutions; though this procedure looks to be suitable for the need of local search, where as excessive search of solutions in search space are known to unhelpfully escalation work load, a occurrence that is distasteful to almost any optimization technique. Nevertheless, this way of approach could be useful in some contexts, but it produces an imbalance between diversification and intensification which could significantly degrade its performance.

As we know very well, an absorption and emission rate is the two parameters play vital role in SBO, to fix the mass of the block hole. The emission rate is one of the operators which have a substantial impact on the correctness of solutions. A large likelihood of emission value good turn to get new block hole, thus increases the exploration of the search space. Whereas a small value of emission rate leads the solution towards to obtain local optimum value. However, the next parameter absorption value is very useful in the form of creating diversity of solution in search space. The constant emission value in traditional SBO algorithm not enhances the convergence rate. Thus, in order to overcome the above drawback and improve the optimization efficiency of SBO, adaptive emission rate is incorporated. Hence, emission rate is dynamically adjusted using the following Equation 1.

$$\beta_i = \left\{ \beta_{max} - (\beta_{max} - \beta_{min}) \times \left( \left(\frac{2t}{NI}\right) - \left(\frac{t}{NI}\right)^2 \right) \right. \quad (2)$$

Where $\beta_{max}$ and $\beta_{min}$ are the maximum and minimum of emission rate respectively. *NI* is the total number of iterations and *t* is the current iteration.

Furthermore, simultaneous dynamic absorption and emission value can cause a twist of global and local search. It can be concluded that an opposite candidate solution has a better chance to be closer to the global optimum solution than a random candidate solution.

### Pseudo code for Stellar-Mass Black Hole Optimization

Initialize a population of N candidate solutions $B_k$
While not (termination criterion)
For each $B_k$ find fitness of $B_k$ , growth rate $\rho_k$ and radiation rate $e_k$
For each $B_k$, update the mass
Rank the solutions
Replace fraction of worst black holes with new black holes
Collide the nearest black holes and generate black holes for left outs in the population
Keep the best solution
Next generation

The population size N of candidate solutions is a tuning parameter. The optimization performance of SBO will suffer when N is too small or too large. In typical implementations of SBO, the value of N is somewhere between 20 and 200. Usually, the zeroth iteration black hole solutions are randomly generated. The termination criterion is problem dependent, as in any other nature inspired algorithm. Black holes with lesser mass emit more energy, finally, at end of their lives. So the black holes with lowest fitness values will be replaced by highest fitness value. It creates the diversity in solution. If the two black holes are close, then the merging of the black holes forms a new black hole. It avoids premature convergence. The new black holes are generated for those left out because the population size is fixed.

## Experimental design and result analysis

### Datasets

Based on these considerations, the aim of the present work is to provide a broad experimental comparison on the feature-selection and classification techniques applied to different DNA microarray datasets. Kent ridge biomedical data repository is used for study of experimental.[13] Table 1 depict the details of the datasets: The column of table means that name of the datasets, samples, genes and classes respectively. Classes denote explicitly the outcome is binary. In order to examine our proposed algorithm, we used accuracy measure. Also, every result obtained over 50 epochs.

**Table 1: Summary for the test databases**

| Datasets | Samples | Genes | Classes |
|---|---|---|---|
| Lung | 181 | 12533 | 2 |
| Colon | 62 | 2000 | 2 |
| Prostate | 102 | 12600 | 2 |
| Leukemia | 72 | 7129 | 2 |
| Lymphoma | 77 | 7129 | 2 |

Initially, well-known statistic techniques such as T-Statistics, SNR and F-Statistics are used to rank the genes in the microarray gene expression data. Due to curse of dimensionality, three level of top ranked (top-50, top-100 and top-200) genes are selected from the gene expression data by applying the statistical measures. As for the classifier considered in this study, ESBO is used to select informative genes from the top-m ranked genes. The Genetic Algorithm (GA), Cuckoo Search Algorithm (CSA), Ant Colony Optimization (ACO) and Particle Swarm Optimization (PSO) are used as the classifiers to measure the performance of ESBO.[14-17] The selected genes are applied to CSA, PSO, GA, and ACO to evaluate the performance. Figure 1 depicts ESBO accuracy value over average of 50 time independent runs on the datasets described and only it shows best of ESBO convergence on prostate dataset with top 50, 100 and 200 genes.

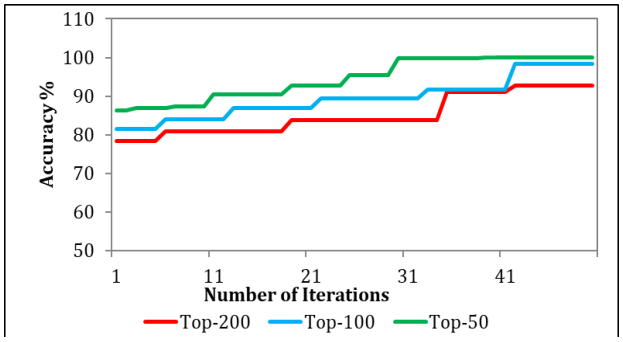


**Figure 1: optimum convergence of ESBO algorithm for prostate dataset**

Figures 2-4 show the accuracy achieved for selected top 200, 100 and 50 samples from CSA, ACO, GA, PSO, SBO and ESBO for lung, colon, leukemia, lymphoma and prostate datasets. The results represent that the proposed ESBO algorithm outperforms existing statistical methods and ESBO in all five cancer gene expression data sets.

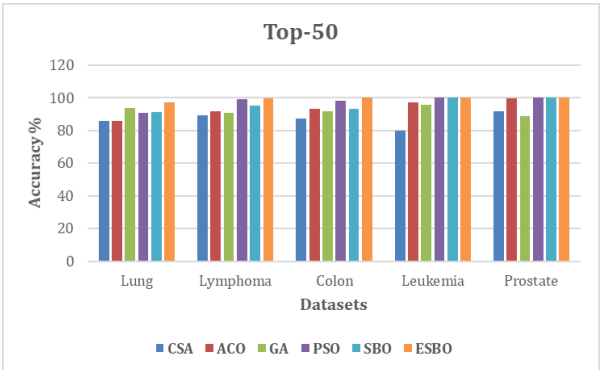


**Figure 2: Datasets versus accuracy on top 50 genes**

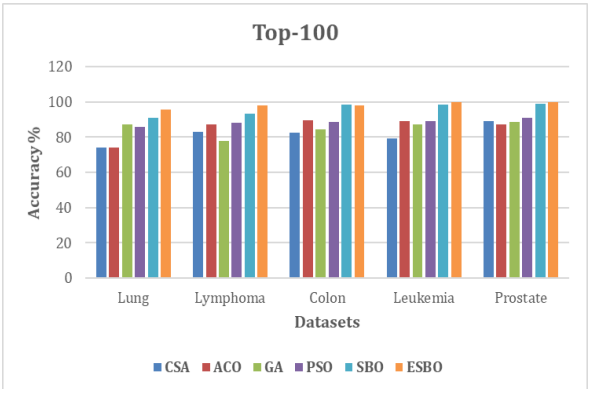


**Figure 3: Datasets versus accuracy on top 100 genes**

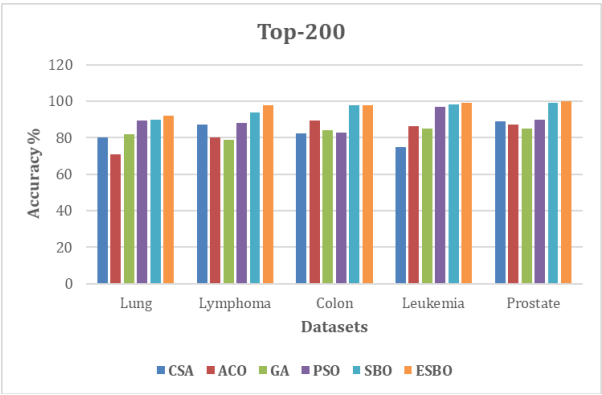


**Figure 4: Datasets versus accuracy on top 200 genes**

### Comparative Analysis

Table 2 depicts Comparison highest accuracy in 50 independent executions of proposed method with four gene selection algorithms on five selected benchmark datasets. Underline indicates the best results. We compared the efficiency of the proposed algorithm with ACO, CSA, PSO and GA. Since these four Bio-inspired algorithms are implemented in the same environment, data sets and parameters. So, results are unquestionably comparable. The performance comparison table shows that, compared to ACO, CSA, PSO and GA proposed method has an obvious advantage. The proposed feature selection method gives 100% classification accuracy for Prostate dataset. For Colon and Leukemia dataset the classification accuracy obtained by the proposed classifier is 98% and 99% respectively. The performance results of ESBO and SBO method (see Table 2) on Colon dataset was as same. This is probably due to the fact that this model focuses on filter the number of irrelevant genes,

but already number of genes is minimum as compared with other datasets. In terms of the correct rate, the search space of solution capability of ESBO is better than the other four approaches.

**Biological significance of Selected Genes for Lung Cancer Data**

To identify functional association of genes with an example of positively selected genes, we used the classification of biological processes based on the p-values from the likelihood ratio test.[18] For Lung Cancer Data, the top 5 genes with the highest selection frequency of each microarray data are presented in Table 3. For example, the roles or activity of the selected gene STX10 is involving in vesicular transport from the late endosomes to the trans-Golgi network. It clarifies that they have a strongest evidence for positive selection and acting as biomarkers for the disease. Hence, these features are evidenced to be the potential reason for Lung Cancer.

**Table 2: Comparison of proposed algorithm with other state of art algorithms**

| Methods | Lung | Lymphoma | Colon | Leukemia | Prostate |
|---|---|---|---|---|---|
| **CSA** | 71 | 80 | 89.34 | 86.23 | 87.32 |
| **ACO** | 80 | 87.32 | 82.54 | 75 | 89 |
| **GA** | 82 | 79 | 84.33 | 85 | 85 |
| **PSO** | 89.54 | 88.27 | 83 | 97 | 90 |
| **SBO** | 90 | 94 | 98 | 98.23 | 99.01 |
| **ESBO** | 92 | 98 | 98 | 99 | 100 |

**Table 3: Top 5 genes with the highest selection frequency of Lung data set**

| Gene Name | Gene id | Functional Associations |
|---|---|---|
| PCDH10 | 57575 | inhibiting cancer cell motility |
| LINC00226 | 338004 | long intergenic non-protein coding RNA |
| PRR16 | 51334 | Increases mitochondrial mass and respiration |
| STX10 | 8677 | involved in vesicular transport from the late endosomes to the trans-Golgi network |
| MRI1 | 84245 | catalytic activity, promotes cell invasion in response to constitutive RhoA activation |

## CONCLUSION

This work proposed a modified nature inspired method named Stellar-Mass Black hole with Self Adaptive Emission Rate or ESBO. It identifies genes which are mostly relevant. With the use of our newly proposed searching algorithm, more numbers of genes were obtained in the high dimensional gene expression data. Hence, Self-adaptive emission rate helps in obtaining discriminative value for each feature. The results of ESBO performance evaluations show that ESBO can make an informative gene for cancer classification on each dataset. Applying the Biological Analysis to the dataset provided a reliable result for external validation. The comparative analysis from the classification performance and the biological activity showed that the proposed searching strategies are quite effective in selecting of significant genes as biomarker for lung cancer. More effort needs to be made on microarray data processing before applying the ESBO model to achieve better results.

## REFERENCES


1. Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton, GG. The Sequence of the Human Genome. Science 2001; 291: 1304-1351.
2. Shalon D, Smith SJ, Brown, PO. A DNA microarray system for analyzing complex DNA samples using two-color fluorescent probe hybridization. Genome Research 1966; 6: 639-645.
3. Pengyi Y, Yoo PD, Juanita F, Zhou BB, Zhang Z, Zomaya, AY. Sample Subset Optimization Techniques for Imbalanced and Ensemble Learning Problems in Bioinformatics Applications. IEEE Transactions on Cybernetics 2014; 44: 445-455.
4. Chiang TS, Watada J, Ibrahim Z, Khalid M. Evolutionary Fuzzy ARTMAP Neural Networks for Classification of Semiconductor Defects. IEEE Transactions on Neural Networks and Learning Systems 2015; 26: 933-950.
5. Kim KJ, Sung-Bae C. Meta-classifiers for high-dimensional small sample classification for gene expression analysis. Pattern Analysis and Applications 2015; 18: 553-570.
6. Zhang Y, Qingchun D, Wenbin L, Xianchun Z. An Efficient Feature Selection Strategy Based on Multiple Support Vector Machine Technology with Gene Expression Data. Bio Med Research International 2018; 2018: 1-10.
7. Pan X, Shen H. Robust Prediction of B-Factor Profile from Sequence Using Two-Stage SVR Based on Random Forest Feature Selection. Protein and Peptide Letters 2009; 16: 1447–1454.
8. Nagpal A, Singh A. A Feature Selection Algorithm Based on Qualitative Mutual Information for Cancer Microarray Data. Procedia Computer Science 2018; 132: 244-252.
9. Alanni R, Hou J, Azzawi H, Xiang Y. A novel gene selection algorithm for cancer classification using microarray datasets. BMC Med Genomics 2019; 12: 1-10.
10. Azzawi H, Hou J, Alanni R, Xiang Y. SBC: A New Strategy for Multiclass Lung Cancer Classification Based on Tumor Structural Information and Microarray Data. In 17th International Conference IEEE/ACIS on Computer and Information Science (ICIS); 2018. p. 68–73.
11. Nakariyakul S. A hybrid gene selection algorithm based on interaction information for microarray-based cancer classification. PLoS ONE 2019; 14: 261-272
12. Liu KH, Muchenxuan T, Shu TX, Vincent TYN. Genetic Programming Based Ensemble System for Microarray Data Classification. Computational and Mathematical Methods in Medicine 2015; 2015: 1-10.
13. datam.i2r.a-star.edu.sg. New York: Kent ridge biomedical data repository, Inc. http://datam.i2r.a-star.edu.sg/datasets/krbd/.
14. Goldberg, DE. Genetic algorithms in search. Optimization and machine learning. 1st Ed. Boston: Addison-Wesley; 1966.
15. Yang, XS, Deb, S. Cuckoo search via lévy flights. Proceedings of World Congress on Nature and Biologically Inspired Computing, India, IEEE Publications, USA; 2010. p. 210-214.
16. Dorigo M, Di Caro G, Gambardella LM. Ant Algorithms for Discrete Optimization. Artificial Life 1999; 5: 137–172.

17. Kennedy J, Eberhart, R. Particle Swarm Optimization. Proceedings of IEEE International Conference on Neural Networks; 1995. p. 1942–1948.
18. Sarah MA, Saleh AI, Labib LM. A new distributed feature selection technique for classifying gene expression data. International Journal of Biomathematics 2019; 12: 368-381.

**Cite this article as:**

R. Balamurugan *et al*. ESBO: Cancer classification in microarray data using SBO algorithm with self-adaptive emission rate. Int. Res. J. Pharm. 2019;10(10):82-86 http://dx.doi.org/10.7897/2230-8407.1010303

Source of support: Nil, Conflict of interest: None Declared