The Performance of Bio-Inspired Evolutionary Gene Selection Methods for Cancer Classification Using Microarray Dataset

Hala M. Alshamlan, Ghada H. Badr, and Yousef A. Alohali

***Abstract*—Microarray based gene expression profiling has become an important and promising dataset for cancer classification that are used for diagnosis and prognosis purposes. It is important to determine the informative genes that cause the cancer to improve early cancer diagnosis and to give effective chemotherapy treatment. Furthermore, find accurate gene selection method that reduce the dimensionality and select informative genes is very significant issue in cancer classification area. In literature, there are several gene selection methods for cancer classification using microarray dataset. However, most of them did not concern on identifying minimum number of informative genes with high classification accuracy. Therefore, in our research study we discuss the performance of Bio-Inspired evolutionary gene selection method in cancer classification using microarray dataset. And, we prove that the Bio- Inspired evolutionary gene selection methods have superior classification accuracy with minimum number of selected genes.**

***Index Terms*—Bio-inspired evolutionary methods, cancer classification, microarray, gene selection, gene expression.**

I. Introduction

Gene expression profiling or DNA microarray dataset has enabled the measurement of thousands of genes in a single RNA sample by hybridized to a labeled unknown molecular extracted from a particular tissue of interest. It offers an efficient method of gathering data that can be used to determine the patterns of gene expression of all the genes in an organism in a single experiment [1], [2]. DNA microarrays can be used to determine which genes are being expressed in a given cell type at a particular time and under particular conditions, to compare the gene expression in two different cell types or tissue samples, and then we can determine the more informative genes that are responsible to cause specific disease or cancer [3].

Recently, microarray technologies have opened up many windows of opportunities to investigate cancer diseases using gene expressions. In our research, we focus on microarray data classification and cancer microarray dataset classification in particular. The primary task of microarray data classification is to determine a computational model from a given microarray data that determines the class of unknown samples. Accuracy, quality, and robustness are important elements of microarray classification models. The accuracy of microarray dataset classification depends on both the quality of the provided microarray data and the utilized classification method. However, microarray dataset suffers from the curse of dimensionality, the small number of samples, and the level of irrelevant and noise genes, makes the classification task for a given sample more challenging [4] [5]. Those irrelevant genes not only introduce some unnecessary noise to gene expression data analysis, but also increase the dimensionality of the gene expression matrix, which results in the increase of the computational complexity in various consequent researches such as classification and clustering [6]. As a consequence, it is significant to eliminate those irrelevant genes and identify the informative genes, which is a feature (genes) selection problem crucial in microarray data analysis.

Therefore, the first step of classifying the microarray data is to identify a small subset of genes that are primarily more predictive for the cancer [5]. In literature, there are several gene selection methods for cancer classification using microarray dataset. However, most of them did not concern on identifying minimum number of informative genes with high classification accuracy. In literature, up to our knowledge, there are several gene selection methods that are based on Bio-Inspired evolutionary algorithm such as Genetic Algorithm (GA), Particle Swarm Optimization (PSO), or Ant Colony Optimization (ACO). These gene selection methods are capable of searching for optimal or near-optimal solutions on complex and large spaces of possible solutions. Thus, in this paper we will evaluate the performance of Bio-Inspired evolutionary gene selection methods in cancer classification using microarray gene expression profile. And, we prove that the Bio-Inspired evolutionary gene selection methods have superior classification accuracy with minimum number of selected genes.

The rest of this paper is organized as follow: In Section II, first we define what is the gene selection process in cancer classification using microarray gene expression profile, followed by description of the various kinds of gene selection methods. The performance of Bio-Inspired evolutionary gene selection method is discussed in section III. In Section IV, we present our contribution. We conclude the paper in Section V.

II. Gene Selection Methods in Cancer Classification

Gene selection is the process of selecting the smallest subset of informative genes that are most predictive to its relative class using a classification model. This maximizes

Manuscript received December 25, 2013; revised March 18, 2014.

The authors are with Computer Science Department, King Saud University, Saudi Arabia (tel.: +96612309954; e-mail: halshamlan@ksu.edu.sa, badrghada@hotmail.com, yousef@ksu.edu.sa).

DOI: 10.7763/IJBBB.2014.V4.332

the classifiers ability to classify samples accurately. In this kind of studies, one aim to identify the genes that contribute the most to cancer diagnosis and this would assist in drug discovery and early diagnosis. It is much cheaper to focus on the expression of only a few genes rather than focusing on thousands of genes. In addition, the feature selection reduces the dimensionality problem, and this leads to a reduction in the classification computational cost [7].

The optimal feature selection problem has been shown to be NP-hard [8]. Therefore, it is better to use heuristics approaches in order to solve this problem. However, the selection of significant genes in microarray studies is challenging. This is because of the very large number of gene expression profiles, when compared to a typically small number of experiments. The low numbers of experiments are due to the high cost for each experiment or the lack of samples. The challenge increases in multi-class microarray, because it is not easy to identify few numbers of genes that are causing multi-type of cancers.

Depending on how feature selection techniques are combined with the classification algorithm, they can be classified into three categories: Filter, Wrapper, and Hybrid gene selection methods. Some techniques do not assume any specific distribution model in the gene expression data. They are referred as a model-free gene selection methods or usually called Filter methods. Others are model-based gene selection methods or called Wrapper methods [9]. When both techniques are combined in one approach, they can be called Hybrid methods. They mainly focus on achieving the best possible accuracy performance with a similar time complexity of Filter techniques. In these methods, the selection of optimal subset of features is usually built inside the classifier, and they iteratively use classifier parameters to select subsets of features [10], [11].

III. Bio-Inspired Evolutionary Gene Selection Methods

The Bio-Inspired evolutionary algorithm is a heuristic optimization algorithm using procedure inspired by mechanisms from organic evolution such as mutation, recombination, and natural selection to find an optimal configuration for a specific system within specific constraint. In lecture, there are many gene selection methods that are based on Bio-Inspired evolutionary algorithm such as Genetic Algorithm (GA), Particle Swarm Optimization (PSO), or Ant Colony Optimization (ACO). These gene selection methods are capable of searching for optimal or near-optimal solutions on complex and large spaces of possible solutions. Thus, in this section, we discuss the performance of Bio-Inspired evolutionary gene selection methods as much as we know from the state-of-art for cancer classification using microarray gene expression profile. In Table I, the performance of the Bio-Inspired evolutionary gene selection methods under study are summarized when applied with cancer classification

Lee *et al.* [12] introduced an effective gene selection method named Adaptive Genetic Algorithm (AGA). The authors evaluate the classification accuracy for this proposal method in Colon cancer microarray dataset by using K Nearest Neighbor KNN algorithm. The results of this study indicate that AGA/KN can be provided as an effective tool with excellent performance for dimension reduction and gene selection on microarray dataset. Recently, Lee and Leu [13] applied Genetic algorithm with dynamic parameter setting (GADP) to select more predictive genes. After selecting the set of significant genes, they use an SVM to test the prediction accuracy using the set of the selected genes. When compared with other methods in literature, the proposed GADP method selects fewer genes with higher prediction accuracy for GCM dataset. Additionally, the GADP method is faster than a traditional GA method in execution time. It is worth mentioning that Huang *et al.* [14], and Huang and Chang [15], presented an efficient evolutionary approach for gene selection from microarray data, which is named Intelligent Genetic Algorithm (IGA) that can be combined with the optimal design of various binary and multi class classifiers, such as Maximum Likelihood [14], and Evolutionary support vector machine (ESVM) [15]. Notably, the high performance of IGA mainly arises from an efficient intelligent crossover operation [14].

Abderrahim *et al.* [16] invented a novel method to identify a small subset of informative genes that could be used to classifying the microarray samples with high accuracy. In this study, the authors combined the Genetic Quantum Algorithm (GQA) with the Support Vector Machines (SVM) classifier for gene selection and classification of high dimensional microarray dataset, and they named this algorithm GQASVM. Notably, the Quantum operators that are applied in this novel method strongly depend on the evolutionary algorithm that is used for data encoding. In order to prove the efficiency of proposed algorithm, the authors make a comparison between the proposed approaches and different other methods in the literature, particularly GASVM and PSOSVM in [3]. In their experiments they used six benchmark two-class cancer microarray datasets (Leukemia, Breast, Colon, Ovarian, Prostate, and Lung). The experimental results show that the proposed method acquires a very good performance. Also, results show that the GQASVM approach is able to determine the minimum number of informative genes with a high classification accuracy.

Alba *et al.* [3] proposed two wrapper feature selection method that used meta- heuristics, and they combined them with SVM classification technique: Particle Swarm Optimization (PSO) and another one that is based on the popular GA using a specialized Size-Oriented Common Feature Crossover Operator (SSOCF). Both approaches (PSOSVM and GASVM) were experimentally evaluated on six benchmark microarray cancer datasets discovering new and challenging results, and select specific genes that are considered as significant ones. Comparisons with several state of art methods show competitive results according to high classification accuracy rate (Around 100%) where a few number of genes are generated per subset (3 or 4) in most experiments.

Recently, hybrid of particle swarm optimization technique and support vector machine (PSOSVM) is developed for gene selection and classification. It is worth mentioning that most versions of PSO have operated in continuous and real-number spaces [17]. It is noted that an improved PSOSVM or modified discrete PSO algorithm would be considered a good solution to solve the over-fitting problem [18]. Thus, Shen *et al*. [17] combined the modified discrete PSO and support vector machines (SVM) for cancer classification. In [17], the modified discrete PSO is utilized to select the informative genes, while SVM is applied as classifier or evaluator. In this study, the authors used colon cancer microarray dataset to evaluate the performance of the proposed method, and they are compared with SVM without adopted any gene selection method. The result shows that the efficiency of SVM model was much improved in classification accuracy by the PSOSVM analysis from 83 to 91.7% with small number of selected genes.

TABLE I: THE PERFORMANCE OF BIO-INSPIRED EVOLUTIONARY GENE SELECTION METHODS

Ref	Gene Selection Method	Classification Method	Data Set	Genes	Accuracy
[1]	Genetic Quantum Algorithm (GQA)	SVM	Leukemia	2	100
			Breast Colon Lung Ovarian Prostate	5	98.97
				1	100
				2	100
				2	100
				2	100
[2]	PSO	SVM	Leukemia	3	100
			Breast Colon Lung Ovarian Prostate Leukemia	4	90.72
			Breast Colon Lung Ovarian Prostate	2	100
				4	99.44
				4	100
				4	100
	GA	SVM		4	100
				4	100
				3	100
				4	100
				4	100
				4	100
[7]	Intelligent GA fold	Evolutionary Sup- port Vector Machine (ESVM)	Brain1	6	96.67
			Brain2	4	100
			DLBCL Leukemia Lung Prostate SRBCT	3	100
				3	100
				6	95
				3	100
				5	98.7
[8]	Intelligent GA (IGA)	Maximum Likelihood	NCI	13	91.75
			Brain1	7	99.2
			Brain2	5	99.65
			Leukemia Lung Prostate	4	100
				7	99.45
				4	99.4
[10]	GADP	(OVO) SVM	Colon	8	100
			Breast	5	100
			Leukemia	5	100
[12]	Adaptive Genetic Algorithm (AGA)	K-nearest neighbor method	Colon	9	100
[17]	Modified discrete PSO	SVM	Colon	4	91.7
[20]	Adaptive Ant Colony Optimization (AACO)	SVM	Colon	4	96.77
			Leukemia	3	100

Xiong and Wang [19] introduced new feature selection algorithm that is a wrapper approach that is based on self-adaptive Ant Colony Optimization (ACO) algorithm when combined with Support Vector Machine (SVM) algorithm. Normalization of distance may blur the discrimination of feature importance score. Thus, they did not normalize the distance of features to accelerate the convergence of ACO.

In order to attain better performance, Xiong and Wang [20] proposed two techniques to enhance the adaptive ACO proposed in [19]. In the first one, each ant starts from a different feature as a start node. While in the second enhancement, ants use an independent random generator (seed2) different from another (seed1) used for paths selection to create the random number of nodes. Moreover, in this study [20], the researchers utilized two statistical classification methods: Random Forests and T-test method. They applied Random Forests first to calculate the feature importance score named RF-Imp. Then, they used t-test score based on statistic theory to provide feature important score for two-class problem. The experimental results on Colon tumor and Leukemia datasets show that the proposed algorithm is effective when applied to microarray data and that it obtains higher classification accuracy with a small selected feature subset.

IV. ANALYSIS AND DISCUSSION

In our study, we observed the most researcher in cancer classification used Bio- Inspired evolutionary gene selection methods and they achieve high classification accuracy with minimum number of selected genes. Moreover, Bio-Inspired evolutionary algorithms such as (GA, PSO, and ACO) are more applicable and accurate as wrapper gene selection method, because it is capable of searching for optimal or near-optimal solutions on complex and large spaces of possible solutions. Furthermore, it is allow searching of these spaces of solutions by considering multiple interacting attributes simultaneously, rather than by considering one attribute at a time. For this motivation, bio inspired evolutionary algorithms may represent a helpful and effective tool in the binary and multi class cancer classification based on microarray gene expression data. In order prove that, we evaluate the performance of Bio-Inspired evolutionary gene selection methods presented in state of art for four cancer microarray dataset (Colon, Leukemia, Lung, and Prostate), and our evaluations are illustrated in Table II and Fig. 1.

TABLE II: THE EVALUATION OF BIO-INSPIRED EVOLUTIONARY GENE SELECTION METHODS

Cancer Dataset	No. Of Genes	Classification Accuracy
Colon	4	98.2
Leukemia	4	100
Lung	5	99.7
Prostate	4	99.7

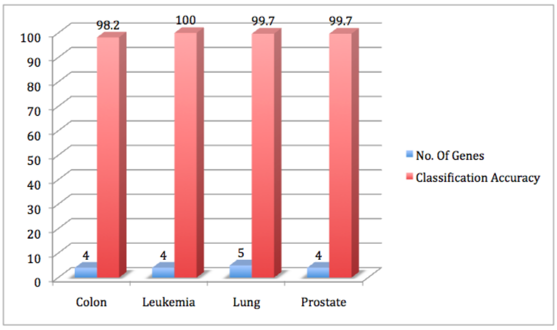

Fig. 1. The Evaluation of Bio-Inspired evolutionary Gene Selection Methods.

Furthermore, Bio-Inspired evolutionary gene selection methods allow searching of the spaces of solutions by considering multiple interacting attributes simultaneously, rather than by considering one attribute at a time. Other advantages of the Bio-Inspired evolutionary approach are that it automatically determines the optimal predictor set size and the delivery of predictive accuracies that are comparable to other methods using classifier gene sets of substantially fewer features than previously required. For these benefits, Bio-Inspired evolutionary gene selection methods may represent a helpful and useful tool in cancer classification based on microarray gene expression data.

Based on our analysis presented in Table II and Fig. 1, we can conclude that Bio-Inspired evolutionary gene selection methods could significantly improve the accuracy of classification algorithms with minimum number of selected genes. In other hand, compared with other approach Bio-Inspired evolutionary approach generally obtains one gene subset with better classification performance but much more computational cost [3]. Bio-Inspired evolutionary gene selection methods can be impractical for some computationally expensive algorithms such as SVM or artificial neural networks [2]. However, the classification accuracy in cancer research in significant, because this reason often uses Bio- Inspired evolutionary gene selection methods in cancer prediction [3].

V. CONCLUSION

Cancer is one of the dreadful diseases, which causes considerable death rate in humans. DNA Microarray based gene expression profiling has been emerged as an efficient technique for cancer classification, as well as for diagnosis, prognosis, and treatment purposes. In recent times, DNA microarray technique has recently gained more attention in both scientific and in industrial fields. It is also important to determine the informative genes that cause the cancer to improve cancer classification and early cancer diagnosis. Classifying cancer microarray gene expression data is challenging task because microarray has a high dimensional-low sample dataset with a lots of noisy or irrelevant genes and missing data.

Bio-Inspired evolutionary algorithms such as (GA, PSO, and ACO) are more applicable and accurate as wrapper gene selection method, because it is capable of searching for optimal or near-optimal solutions on complex and large spaces of possible solutions. In this research study, we prove that Bio-Inspired evolutionary gene selection methods achieve high classification accuracy with minimum number of selected genes for cancer microarray gene expression profile. In future, we will compare the performance of Bio-Inspired evolutionary methods with filter and hybrid gene selection methods in cancer classification based on microarray dataset.

REFERENCES

[1] A. Abderrahim, E. Talbi, and M. Khaled, “Hybridization of genetic and quantum algorithm for gene selection and classification of microarray data,” in *Proc. IEEE International Symposium In Parallel Distributed Processing*, , pp. 1–8, 2009.

[2] E. Alba, J. Garcia-Nieto *et al*., “Gene selection in cancer classification using pso/svm and ga/svm hybrid algorithms,” *Evolutionary Computation*, pp. 284–290, 2007.

[3] C. Alonso, I. Moro-Sancho, A. Simon-Hurtado, and R. Varela-Arrabal, “Microarray gene expression classification with few genes: Criteria to combine attribute selection and classification methods,” *Expert Systems with Applications*, vol. 39, no. 8, pp. 7270 – 7280, 2012.

[4] C. Feng and W. Lipo, “Applications of support vector machines to cancer classification with microarray data,” *International Journal of Neural Systems*, vol. 15, no. 6, pp. 475–484, 2005.

[5] L. M. Fu and C. S. Fu-Liu, “Multi-class cancer subtype classification based on gene expression signatures with reliability analysis,” *FEBS Letters*, vol. 561, no. 13, pp. 186–190, 2004.

[6] S. Ghorai, A. Mukherjee, S. Sengupta, and P. Dutta, “Multicategory cancer clas- sification from gene expression data by multiclass nppc ensemble,” in *Proc. International Conference on Systems in Medicine and Biology*, pp. 4–48, 2010.

[7] H.-L. Huang and F.-L. Chang, “Esvm: Evolutionary support vector machine for automatic feature selection and classification of microarray data,” *Biosystems*, vol. 90, no. 2, pp. 516–528, 2007.

[8] H.-L. Huang, C.-C. Lee, and S.-Y. Ho, “Selecting a minimal number of relevant genes from microarray data to design accurate tissue classifiers,” *Biosystems*, vol. 90, no. 1, pp. 78–86, 2007.

[9] Y. Kun, C. Zhipeng, L. Jianzhong, and L. Guohui, “A stable gene selection in microarray data analysis,” *BMC Bioinformatics*, vol. 7, no. 1, pp. 1–16, 2006.

[10] C.-P. Lee and Y. Leu, “A novel hybrid feature selection method for microarray data analysis,” *Soft Computer Application*, vol. 11, pp. 208–213, Jan. 2011.

[11] C.-P. Lee, W.-S. Lin, Y.-M. Chen, and B.-J. Kuo, “Gene selection and sample classification on microarray data based on adaptive genetic

algorithm/k-nearest neighbor method," *Expert Systems with Applications*, vol. 38, no. 5, pp. 4661 – 4667, 2011.

[12] M. Mohamad, S. Omatu, M. Yoshioka, and S. Deris, "An approach using hybrid methods to select informative genes from microarray data for cancer classification," in *Proc. International Conference in Modeling Simulation*, pp. 603–608, 2008.

[13] N. Patrenahalli, "A branch and bound algorithm for feature subset selection," *IEEE Transactions on Computers*, vol. 26, no. 9, pp. 917–922, 1977.

[14] Q. Shen, W.-M. Shi, W. Kong, and B.-X. Ye, "A combination of modified particle swarm optimization algorithm and support vector machine for gene selection and tumor classification," *Talanta*, vol. 71, no. 4, pp. 1679 – 1683, 2007.

[15] G. Sheng-Bo, L. Michael, and L. Ming, "Gene selection based on mutual information for the classification of multi-class cancer," pp. 454–463, 2006.

[16] R. Simon, "Analysis of DNA microarray expression data," *Best practice and research Clinical haematology*, vol. 22, no. 2, pp. 271–282, 2009.

[17] W. Xiong and C. Wang, "Feature selection: A hybrid approach based on self- adaptive ant colony and support vector machine," in *Proc. International Conference on Computer Science and Software Engineering*, vol. 4, pp. 751–754, 2008.

[18] W. Xiong and C. Wang, "A hybrid improved ant colony optimization and random forests feature selection method for microarray data," in *Proc. International Conference on Networked Computing and Advanced Information Management*, pp. 559–563, 2009.

[19] H. Yu and S. Xu, "Simple rule-based ensemble classifiers for cancer dna microarray data classification," in *Proc. International Conference in Computer Science and Service System (CSSS)*, pp. 2555–2558, 2011.

[20] S. Yvan, I. Aki, and L. Pedro, "A review of feature selection techniques in bioinformatics," *Bioinformatics*, vol. 23, pp. 2507–2517, Sept. 2007.

Hala M. Alshamlan is a lecturer in Information Technology Department, King Saud University Riyadh, Saudi Arabia. Currently, she is a Ph.D. Candidate in Computer Science Department, King Saud University. Her research of Interest in bioinformatics, artificial intelligent, data mining, and machine learning. and, she has many publications in these areas.

Ghada H. Badr completed her Ph.D. in 2006 in computer science at Carleton University, School of Computer Science, Ottawa, Canada. She was the winner of the university Senate Medal for outstanding research achievements. From 2006 to 2007, she worked as a research associate at the National Research Council (NRC) of Canada for the language technology group in Gatineau, Canada, in the field of Machine Translation. In 2007, she won the prestigious NSERC Postdoctoral Fellowship in Canada. Since 2007 till 2011, she worked as a Postdoctoral fellow at the University of Ottawa, Ottawa, Canada, where her research focused in the field of Data mining in bioinformatics. She worked, and still working, on another research project in other bioinformatics aspects, where it involves discovering and localizing interacting RNA secondary structures. At KSU, she was able to establish the Bioinformatics Research group (BioInG), where she is the coordinator for the group since Fall 2012. Through the group she was able to attract a lot of researchers from different departments and to develop a lot of activities and workshops.

Yousef A. Alohali is an associated professor in Computer Science Department at King Saud University. He completed his Ph.D in computer science at Concordia University, Montreal, Canada. Her research of interested included: artificial intelligence: human computer interaction, data mining, office automation, intelligent solutions, and optimization techniques.