# Bio-inspired Machine Learning in Microarray Gene Selection and Cancer Classification

Sultan H. Aljahdali
Computer Sciences Dept., Taif University
Taif -Saudi Arabia
aljahdali@tu.edu.sa

Mohammed E. El-Telbany
Computers Engineering Dept., Taif University
Taif -Saudi Arabia
telbany@tu.edu.sa

***Abstract -*** Microarray technology today has the ability of having the whole genome spotted on a single chip. It allows the biologist to inspect thousands of gene activities simultaneously. Machine learning approaches are suited and used to discovering the complex relationships between genes under controlled experimental conditions and classify microarray data by identifying a subset of informative genes embedded in a large data set that involves multiple classes and is infected with the high dimensionality noise. In this paper, a hybrid system integrates genetic algorithms and decision tree is proposed for genes expression analysis and prediction to their functionality for cancer classification. The learning capacity of decision trees used in the base learning systems is boosted by feature selection method. Experiments present preliminary results to demonstrate the capability of hybrid system to mine accurate classification rules for classifying prediction in comparable to traditional machine learning algorithms.

***Keywords -*** bioinformatics, classification, genetic algorithms, decision tree, and feature selection.

## 1. INTRODUCTION

The human genome project provides the capstone for efforts in the past century to discover genetic information and a foundation of efforts to understand it using the computational techniques. This integration is known as bioinformatics [19]. The bioinformatics work would have profound long-term consequences for medicine, leading to the explanation of the underling molecular mechanism of diseases and thereby. However, no much is known about the structure, function, expression and regulation of more than 80% of human genes [19]. In order to assign a function to many genes, there is a need for computational method for functional prediction for unknown genes, since the experimentally determining the function of a protein is time-consuming. One method of predicting the gene functionality is the study of its expression pattern. The expression of genes is complex and highly controlled and regulated process. The relatively recently developed high potential DNA microarray technology has been used for gene expression profiling of normal and malignant cells in several tumors including clone [1], leukemia [14], Lung cancer [3], and breast [21]. These studies may provide mechanist insight into maltransformation and help to in identifying biomarkers for cancer classifications. Due to the huge number of genes examined at once and only a fraction may present distinct profiles for different classes of cancers. The machine learning approaches are suited and used to endow the doctors with the capability of discovering the complex relationships between genes under controlled experimental conditions and classify microarray data by identifying a subset of informative genes embedded in a large data set that involves multiple classes and is infected with the high dimensionality noise. A previously presented knowledge-based computational machine learning techniques include the use of *k*-nearest neighbors (KNN) [17], artificial neural networks [15], [11], [5], support vector machines [4], [11], [5], hierarchical clustering [9] *K*-mean clustering [24] and self-organizing map [14]. However, the microarrays data suffer from the *curse of dimensionality*, where the microarrays data consists of a large number of genes and a small number of samples, which unlike the conventional data sets, that consists of a large number of samples and few parameters. These data will yields many distinct class predictors due to the existing of several subset of genes that can distinguish between different classes of samples. These subsets competing near-optimal solutions and in order to find the global optimal solution any algorithm must be able to deal robustly with the dimensionality of this feature space by identifying the subset of genes that can potentially discriminate efficiently between different classes of sample. From above discussion, we can conclude that the size of a dataset of microarray is so large that learning might not work as well before removing these unwanted features. Reducing the number of irrelevant/redundant features drastically reduces the running time of a learning algorithm and yields a more general classifier [2],[10]. In this paper, a hybrid system integrates genetic algorithms [13] and decision tree [22] is proposed for genes expression analysis and prediction to their functionality for cancer classification. Since the knowledge of the characteristic expression patterns of functional classes of genes can be utilized in the annotation of the unknown genes. The decision trees as data structure used for classification function, while genetic algorithms implement the selection process for the informative genes. The learning capacity of decision trees used in the base learning systems is boosted by feature selection method.

## 2. RELATED WORK

Recently, many methods for selecting a subset of informative genes for sample classification have

978-1-4244-5950-6/09/$26.00 ©2009 IEEE

proposed. Golub *et al.*, [14] successfully applied neighborhood analysis to identify a subset of genes that discriminate between acute myeloid leukemia (AML) and acute lymphoblastic leukemia (ALL), using a separation measure similar to *t*-statistic. Li and Yang [18] ranked the genes as had been done in the first analysis [14] and used the top ranked genes. They varied the number they included and found no clear indication of any optimal number and get the conclusion obtained by Golub *et al.*, [14]. Alaiya *et al.*, (2000) and Khan *et al.*, [15] used the principle component analysis (PCA) to identify a subset of genes. Li *et al.*, [16] used the genetic algorithms to choose a relatively few subset of genes for testing based on the valuation function. [6] in this case the KNN. Deutsch [7][8] used a replication algorithm to evolve an ensemble of predictors, to generate a set of optimal predictors as a form of generative procedure [6]. Ooi and Tan [20] used a hybrid technique combining genetic algorithm with maximum likelihood to select the optimal number of genes.

## 3. DNA MICROARRAY DATA SET

The microarray data obtained in parallel gene expression experiments provides the expression levels of *n* genes of interest, which measured under different conditions in *m* experiments. The data points form a $m \times n$ gene expression matrix, where an *n*-element expression vector represents each gene expression level ratio. There are several microarray data sets from published cancer gene expression studies. The *c*DNA microarrays data sets reported by [20] that are selected from the NCI60 dataset of cell lines which corresponding the nine tumor types [23] will be used in this study as shown in Table 1. The data set contains a normalized expression data for 1000 genes, which have the highest standard deviation value. The normalization is occur by subtracting the mean of the Cy5/Cy3 ratio of the control spot and divides the result by the standard deviation of the Cy5/Cy3 ratio of the control spot [20].

**Table 1: The classes in microarray data set.**

| Name | No. of Samples |
|---|---|
| Breast | 7 |
| Central Nervous System | 6 |
| Colon | 7 |
| Melanoma | 8 |
| Leukemia | 6 |
| Renal | 8 |
| Non-Small-Cell-Lung-Carcinoma | 9 |
| Ovarian | 6 |
| Reproductive | 4 |

## 4. HYBRID GA/DECISION TREE SYSTEM

The classification problem of microarrays genes can be formulated an *optimization* and *search* problem where the result is a complicated function $f$ that needs to be optimized. This function can be represented as follows:

$$f : D \rightarrow R \tag{1}$$

Where $D$ is the set of possibilities and best choices are those for the function $f$ is optimal, the result is a complicated objective function $Q(S)$ that needs to be optimized, where $Q(S)$ represents the quality measurement for a solution $S$ given $\forall S \; Q(S) \geq 0$. The problem is to find the best solution (i.e., classification) $S'$ such that:

$$Q(S') = Maxs\, Q(S) \tag{2}$$

In our proposed implementation, the solution $S$ which is the classification function represented by decision tree data structure, while the searching technique for the best solution is implemented using genetic algorithms as shown in Figure 1.

Searching for informative genes set as a preprocessing step prior to the application of learning algorithm is important for many reasons. One reason is, that the prediction accuracy of the decision tree (i.e., C4.5) decreases when irrelevant or radiant features are added. Another problem particularly affecting the computation time is the lacking scalability of the learning decision tree. Hybridization the genetic algorithms with the decision tree algorithm (i.e. C4.5) as feature selection method will boost the learning capacity of the decision tree and increasing their accuracy and scalability.

### *4.1 Genetic Algorithms*

Genetic algorithm is an iterative optimization technique. Instead of working with a single candidate solution in each iteration, genetic algorithm works with a number of candidate solutions (collectively known as a population) in each iteration. In the absence of any knowledge of the problem domain, a genetic algorithm begins its search from a random population of solutions. Before a GA can be run, a suitable coding (or *representation*) for the problem must be devised. We also require a *fitness function*, which assigns a figure of merit to each coding solution. During the run, if the termination condition is not satisfied, parents must be selected for reproduction, and recombined to generate offspring using the reproduction, crossover and mutation operators to update the population of candidate solutions.

Since the genetic algorithm is responsible for genes selection, the solution is represented as a set of 20 genes indices of a subset of genes picked from the truncated 1000 gene dataset. The genes indices can be duplicated to indicate the same gene. The genetic algorithm uses the *steady-state* model which uses overlapping populations with a user-specifiable amount of overlap.

978-1-4244-5950-6/09/$26.00 ©2009 IEEE

The initial population is generated by creating 50 random set of genes. The crossover probability is 0.8, mutation probability is 0.001, and percentage of population replacement is 25%. The experiments are done using GAlib which a library of genetic algorithms in C++ [12] for 51 generations. The fitness function calculated using a more quality metric is classification accuracy of the decision tree using the selective set of genes.

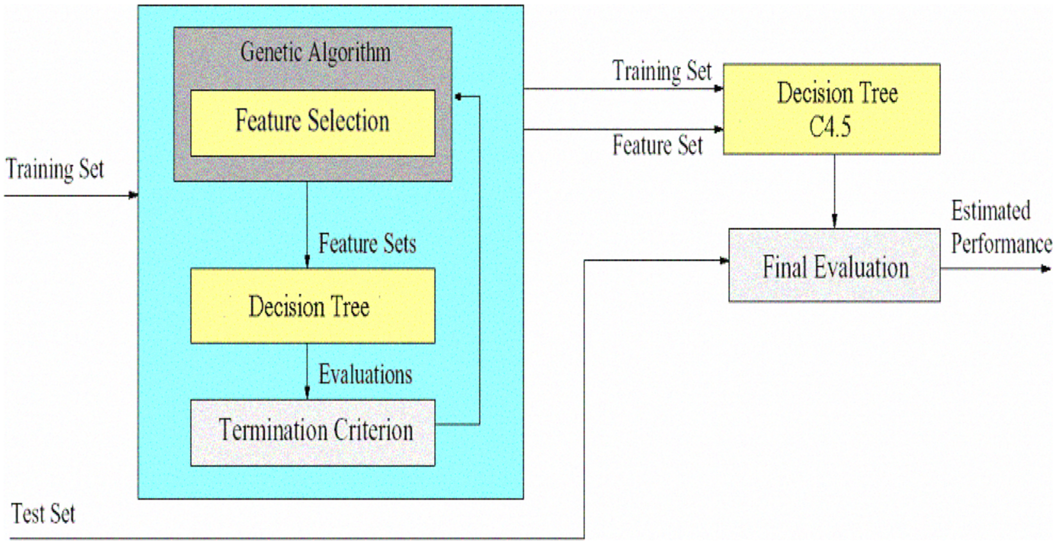

**Fig. 1. The GA/Decision Tree System Architecture**

## 5. EXPERIMENTAL RESULT

The hybrid GA/DT system is tested on the microarray data set and we arbitrarily took the same number of training and test sets reported by Ooi and Tan [20] which consists of 41 training sample and 20 test samples. Multiple runs are conducted with and the results from various runs are presented and compared with the best results reported by Ooi and Tan [20] which represents the best results obtained so far. The accuracy of results of training and test data of the two systems is presented in Table 2.

**Table 2. The Classification Accuracy Comparison.**

| | GA/DT algorithm | | GA/MLHD algorithm | |
|---|---|---|---|---|
| | Training | Test | Training | Test |
| Accuracy | 100.0% | 100.0% | 85.37% | 95.0% |

The GA/DT system is able to get classification accuracy 100.0% and 100.0% on training and test set data. Using GA/DT can effectively create comprehensive tree with greater predictive power with a few learning iterations as shown in Figure 2. The best predictor set of genes obtained using the GA/DT system is listed in Table 3. However, these set is different from the ones reported in by Ooi and Tan [20].

The effect of the crossover and the mutation probabilities on the system performance is studied and the higher values of both probabilities increase the system performance as shown in Fig. 3.

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 80 | 319 | 323 | 557 | 641 | 663 | 730 | 743 | 777 |

## 6. Conclusion and Discussion

This paper explores the synergy of GA and C4.5 learning algorithms in comparison with GA/MLHD classifier [20] for classification the microarray data. The GA/MLHD it best results is able to get a classification accuracy 95.0% on data of test set however, using GA/DT can effectively create comprehensive tree with greater predictive power that reduce the error to 0.0%.

## REFERENCES

[1] Alon U., Barkai N., Notterman A., Gish K., Ybarra S., Mack D., and Levine J., (1999). Broad patterns of gene expressions revealed by cluster analysis of tumor and normal colon tissues probed by oligonucleotide arrays. Proc. Natl. Acad. Sci., **96** (12), pp. 6745-6750.

[2] Bala J., Huang J., Vafaie H., DeJong K., and Wechsler (1995). *Hybrid Learning Using Genetic Algorithms and Decision Trees for Pattern Classification.* IJCAI conference, Montreal.

[3] Bhattacarjee A., Richards A., Staunton J., Li C., Monti S., Vasa P., Ladd C., Beheshti J., Bueno R., Gillette M., Loda M., Weber G., Mark E., Lander E., Wong W., Johonson B.,

Golub T., Sugarbarker D., and Meyerson M., (2001) *Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses*. Proc Natl. Acad. Sci., 98, pp. 13790-13795.

[4] Brown M., Grundy W., Lin D., Cristianini N., Sugent C., Furey T., and Jr A., (2000). *Knowledge-based analysis of microarray gene expression data by using support vector machines*. Proc. Natl. Acad. Sci., **97**(1), pp. 262-267.

[5] Cho S., and Won H., (2003). *Machine Learning in DNA Microarray analysis for cancer Classification.* In First Asia-Pacific Bioinformatics Conference, Australia.

[6] Dash M., Liu H., (1997). *Feature Selection for Classification*. Intelligent Data Analysis, **1**, pp. 131–156.

[7] Deutsch J., (2001). Algorithm for finding an Optimal Gene Set in Microarray prediction. AvXiv: physics/0108011 v1.

[8] Deutsch J., (2003) Evolutionary Algorithms for finding Optimal Gene Set in Microarray prediction. Bioinformatics, 19(1), pp. 45-52.

[9] Eisen B., Spellman T., Brown O., and Botstein D., (1998). *Cluster analysis and display of genome-wide expression patterns*. Proc. Natl Acad. Sci., **95**, pp. 14863-14868.

[10]El-Telbany M., Lichtenegger J., Abdelwhab Ashraf H., and Sheta A. (2001). *Detection of Oil Spill Using a Genetic Recognition System (GRS).* In The International Conference on Industrial Electronics, Technology & Automation, Cairo, Egypt.

[11]Furey T., Cristianini N., Duffy N., Bednarski W., Shummer D., and Haussler D., (2000) *Support vector machine classification and validation of cancer tissue samples using microarray expression data*. Bioinformatics, **16**, pp. 906-914.

[12]GAlib: Matthew's C++ Genetic Algorithms Library., http://lancet.mit.edu/ga/GAlib.html

[13]Goldberg D., (1989). *Genetic Algorithms in Search, optimization, and Machine Learning.* Addison-Wesley.

[14]Golub T., Slonim D. Tamayo P., Gaasenbeek M., Mesirov J., Collor H., Loh M., Downing J., Caligiuri M., Bloomfield C., and Lander E., (1999). *Molecular classification of cancer: class discovery and class prediction by gene expression monitoring*. Science, **286**, pp. 531-537.

[15]Khan J. Wei S., Ringner M. Saal H., Ladanyi M., Westermann F., Berthold F., Schwab M., Antonescu R., Peterson C. and Meltzer S. (2001). *Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks*. Nature Medicine, 7(6) pp. 673-679.

[16]Li L., Pedersen L., Darden T., and Weinberg C., (2001). *Class Prediction and Discovery Based on Gene Expression Data*. Technical Report.

[17]Li L., Weinberg C., Darden T., and Pedersen L., (2001). Gene Selection for Sample classification based on gene expression data: study of the sensitivity to choice of parameters of the GA/KNN method. Bioinformatics **17**(12), pp., 1131-1142.

[18]Li W., and Yang Y. (2001). How many genes are needed for a discriminate microarray analysis?. ArXiv: physics/0104029 v1.

[19]Luscombe N., Greenbaum D., and Gerstein M., (2001). *What is bioinformatics? An Introduction and Overview*. IMIA.

[20]Ooi C., and Tan P., (2003). Genetic algorithms applied to multi-class prediction for the analysis of gene expression data. Bioinformatics, **19**(1), pp. 37-44.

[21]Perou C., Jeffey S., Van de Rijn M., Rees C., Eisen M., Ross D., Pergamenschikov A., Willias C., Zhu S., Lee J., Lashkari D., Shalon D., Brown P., and Botstein D., (1999) *Distinctive gene expression patterns in human mammary epithelial cells and breast cancer*. Proc. Natl. Acad. Sci., **96** (16), pp. 9212-9217.

[22]Quinlan R., (1993). C4.5: Programs for machine learning. Morgan Kaufmann.

[23]Ross T., Scherf U., Eisen B., Perou M., Rees C., Spellman P., Iyer V., Jeffrey S. Van deRijn M., and Waltham M., (2000) *Systematic Variation in Gene expression Patterns in Human Cancer Cell Lines*. Natural Genet., 24, pp. 227-235.

[24]Tavazoie S., Hughes D., Campbell J., Cho J., and Charch M., (1999). *Systemic determination of genetic network architecture*. Nat Genet, **22**(3), pp. 281-285.

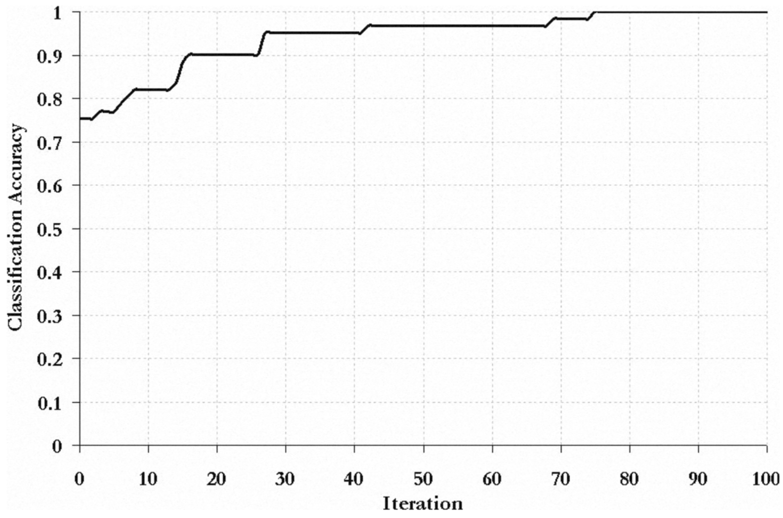

**Fig. 2. Learning rate of GA/DT System**

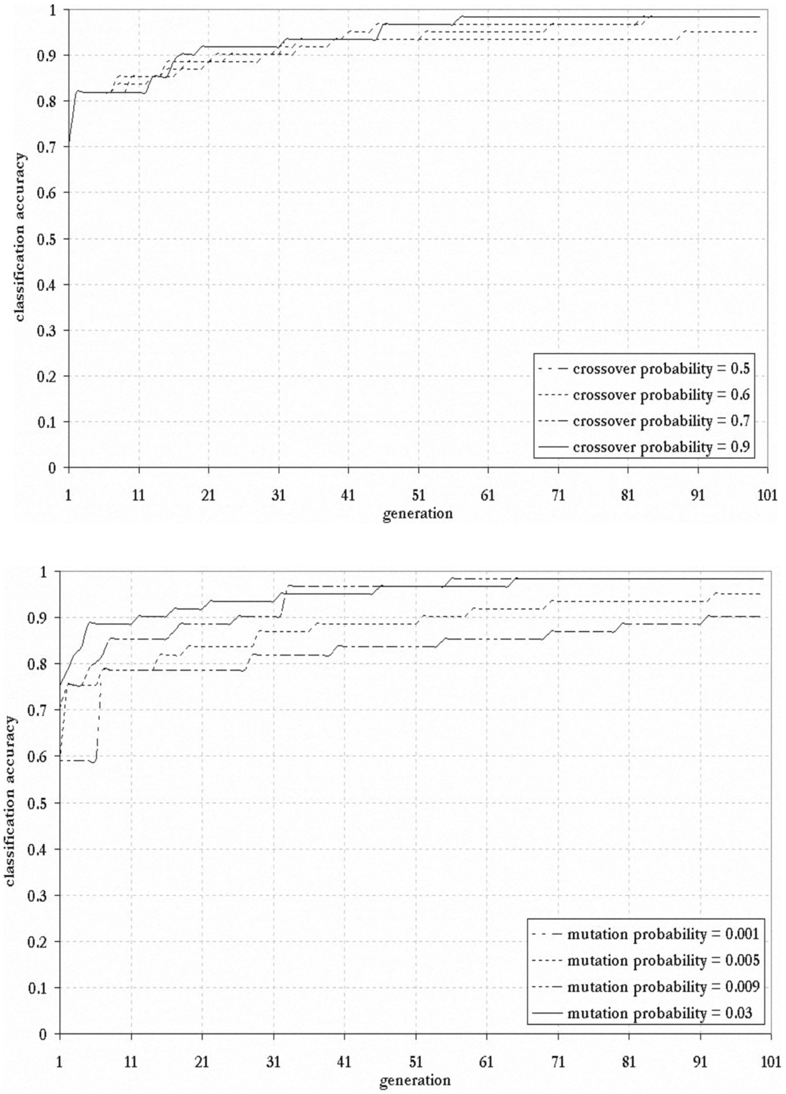

**Fig. 3. The effect of crossover and mutation probabilities on the GA/DT system performance**