Bioinspired Learning for Microarray Gene Selection and Cancer Classification

Sultan H. Aljahdali
Computer Sciences Department, Al-Taif University
Al-Taif -SAUDI ARABIA
aljahdali@tu.edu.sa

Mohammed E. El-Telbany
Computers Engineering Department, Al-Taif University
Al-Taif -SAUDI ARABIA
telbany@eri.sci.eg

Abstract- One major application of microarray technology lies in cancer classification. Thus far, a significant amount of new discoveries have been made and new bio-markers for various cancers have been detected from microarray data. Bioinspired 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 presenting a preliminary result 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, neural networks, and feature selection.

I. 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 underlying 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 mal-transformation 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 researchers 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, due to the nature of the microarray gene expression data, cancer classification has remained a great challenge to computer scientists. Microarray data is characterized with thousands of genes but with only a small number of samples available for analysis. This makes learning from microarray data an arduous task under the effect of curse of dimensionality. These data will yields many distinct class predictors due to the existing of several subsets 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. Moreover, the size of a dataset of microarray is affect the speed and accuracy of most learning algorithms, 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]. Therefore, the goal of feature selection is to find a subset of genes as small as possible, while simultaneously optimizing classification accuracy. This feature subset selection problem is essentially a multi-criterion optimization problem. The multiple criteria to be optimized include the accuracy of classification, cost and risk associated with classification which in turn depends on the selection of genes used to classify the cancer. As shown in Figure 1, there are three main approaches used to feature selection:

1) Filter methods: they are preprocessing methods. They attempt to assess the merits of features from the data, ignoring the effects of the selected feature subset on the performance of the learning algorithm. Examples are methods that select variables by ranking them through compression techniques (like PCA) or by computing correlation with the output Guyon and Elisseeff, 2003].


2) **Wrapper methods**: these methods assess subsets of variables according to their usefulness to a given predictor. The method conducts a search for a good subset using the learning algorithm itself as part of the evaluation function. The problem boils down to a problem of stochastic state space search by utilizing the learning system as a black box to score subsets of features [Blum and Langley, 1997].

3) **Embedded methods**: they perform variable selection as part of the learning procedure and are usually specific to given learning machines. Examples are classification trees, regularization techniques (e.g. lasso) [Breiman et al., 1984].

**FIGURE I: FEATURE SELECTION TECHNIQUES**

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.

**II. 2. Related Work**

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

**III. 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×n gene expression level ratio, where there are several microarray data sets from published cancer gene expression studies. The cDNA 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.**

<table>
<thead>
<tr>
<th>Name</th>
<th>No. of Samples</th>
</tr>
</thead>
<tbody>
<tr>
<td>Breast</td>
<td>7</td>
</tr>
<tr>
<td>Central Nervous System</td>
<td>6</td>
</tr>
<tr>
<td>Colon</td>
<td>7</td>
</tr>
<tr>
<td>Melanoma</td>
<td>8</td>
</tr>
<tr>
<td>Leukemia</td>
<td>6</td>
</tr>
<tr>
<td>Renal</td>
<td>8</td>
</tr>
<tr>
<td>Non-Small-Cell-Lung-Carcinoma</td>
<td>9</td>
</tr>
<tr>
<td>Ovarian</td>
<td>6</td>
</tr>
<tr>
<td>Reproductive</td>
<td>4</td>
</tr>
</tbody>
</table>

**IV. 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:

\[ f : D \rightarrow R \]  

(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 \( \forall S \geq 0 \). The problem is to find the best solution (i.e., classification) \( S' \) such that:

\[
Q(S') = \max_{S} Q(S)
\]

(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. 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 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.

V. 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>
<thead>
<tr>
<th></th>
<th>GA/DT algorithm</th>
<th>GA/MLHD algorithm</th>
</tr>
</thead>
<tbody>
<tr>
<td>Training Accuracy</td>
<td>100.0%</td>
<td>85.37%</td>
</tr>
<tr>
<td>Test Accuracy</td>
<td>100.0%</td>
<td>95.0%</td>
</tr>
</tbody>
</table>

The GA/DT system is able to get a 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].

![Figure 1: The GA/Decision Tree System Architecture](image1)

![Figure 2: Learning rate of GA/DT System](image2)
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.

### Table 3. The Best Predictor Set of Genes.

<table>
<thead>
<tr>
<th>Gene</th>
<th>1</th>
<th>80</th>
<th>319</th>
<th>323</th>
<th>557</th>
<th>641</th>
<th>663</th>
<th>730</th>
<th>743</th>
<th>777</th>
</tr>
</thead>
</table>

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%.

### VI. 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%.

### VII. References


