An Intelligent Hybrid Feature Selection for Big Data Analytics: Application to Microarray Data Analysis

Thavavel Vaiyapuri
Department of Computer Science
Prince Sattam Bin Abdulaziz University
Saudi Arabia

Nasser Alanazi
Department of Information System
Prince Sattam Bin Abdulaziz University
Saudi Arabia

Abstract—Recent advances in Genomics technologies has enabled cancer research to enter the Big Data era. In Big data analytics, both dimensionality reduction and designing an optimal analysis model is a challenging task. This paper presents an effective two phase feature selection model to effectively address this challenge in multi-class cancer classification problem. In the first phase, the model utilizes ensemble of filters to reduce the curse of dimensionality eliminating the irrelevant genes and maintaining the prognostic genes. In second phase, hybrid particle swarm optimization (HPSO) is employed to further reduce the dimensionality of prognostic genes and optimize the classifier structure synchronously to achieve optimal classification accuracy. To confirm the effectiveness of the proposed model, it is compared with other well-known recent methods on five benchmark microarray datasets using 10-fold cross validation. Experimental results demonstrates its superior performance in selecting compact prognostic gene subset while maintaining the classification accuracy.

Keywords—Big data Analytics, Genomics, Feature Selection, Hybrid approach, HPSO, ELM, Cancer Classification

1. INTRODUCTION

Rapid advances in high-throughput genomic technology have revolutionized the field of biomedicine and has enabled biomedical scientist to enter the era of ‘Big Data’ (large datasets) [1]. Genomic data are heterogeneous and high dimensional with thousands of features (genes) for small size of samples. Further, most of these features are irrelevant and redundant in determining the target class for given input samples. These characteristics of genomic data not only pose great computational challenges but also affects the accuracy of data analysis which is essential for correct diagnosis and prognosis of cancer, the life threatening disease in world [2]. Consequently, many literatures have recommended dimensionality reduction technique to combat this phenomenon [3].

Feature selection, one type of dimension reduction technique, has proven to be effective and efficient in determining the most relevant features without altering the physical meaning of the original features and generalizes the performance according to a given problem. Thus, feature selection is widely employed in genomics to identify small subset of features (genes) with high prognostic value for characterizing the heterogeneity of cancers [4]. This process also called as gene selection in the field of genomics is considered as a necessary preprocessing step to reduce the dimensionality and enhance the efficiency of the subsequent data analysis task which will shed light in understanding the biological process and guide better therapeutic decisions for cancer treatment.

In general, feature selection employs evaluation function to find the optimal feature subset. The evaluation function measures relevance of features to discriminate between all target classes. Depending on whether the evaluation function interacts with the classifier, the feature selection methods are categorized into Filters and Wrappers [5]. Filter methods, searches and selects the features based on their discriminative power without interacting with the classifier. Methods that use statistical scores for measuring the discriminative power have shown to be effective. These methods include Signal-to-Noise Ratio (SNR) [6], t-statistics (TS) [7], and F-test [8]. Although filter methods are simple and efficient in searching through the feature space, they fail to consider the correlation between features. While wrapper method encapsulates a search method and classifier in a single approach. The search method explores the feature space for all possible feature subsets and selects the optimal feature subset that can achieve highest performance with the evaluation of a classifier. As result, wrapper method leads to better classification accuracy than filter methods. But the feature selection task in wrapper method is challenging because of the large search space. Great many studies have appeared in literature to solve this problem with variety of search methods such as greedy search based sequential forward selection (SFS) and sequential backward selection (SBS) [9,10]. Unfortunately, these search methods still suffer from local optima and high computational cost. Over past few years, evolutionary computation (EC) that are well known for their global search ability are employed to better address this problem [11,12].

In this context, particle swarm optimization (PSO) has been used increasingly as an effective technique for data analyses [13]. Compared to genetic algorithm, PSO has no complicated evolutionary operators and fewer parameters need to be adjusted [14]. Accordingly, many wrapper models based on PSO have been proposed for gene selection to determine prognostic genes from microarray data [15-17]. These evolutionary based wrappers methods are capable of selecting a compact optimal subset of predictive genes but suffer high computation cost and greater search complexity due to very
high dimensionality of the dataset [15]. In summary, wrapper methods complement each other, where filter methods search through the feature space efficiently and the former provides good accuracy.

Previous studies have proposed hybrid models combining the advantage of filter and wrapper methods in a single approach [18-20]. These hybrid models employ filter method for initially screening out majority of irrelevant features and reducing the computational complexity of wrapper method. But this affects the performance of hybrid model constraining the wrapper method to reply only on those features selected by a single filter which may screen out some relevant biomarkers even from the chance to be considered in the wrapper evaluation [21, 22]. Based on the assumption, there is no single filter that performs best for any given problem, this study considers the decision of ensemble of filters to address this problem.

The main purpose of this research is to present an intelligent hybrid model to solve the feature selection problem and demonstrate its efficiency. The proposed hybrid model employs ensemble of filters with the aim to ensure that all useful features (biomarkers) are selected in initial filter stage. Then employs PSO in wrapper stage with the aim to improve the accuracy of classifier by synchronously optimizing the classifier parameters and the final gene subset. Furthermore, the impact of proposed model is investigated using six publicly available genomic benchmark datasets to provide useful insight on both benefits and limitations to pave a way to explore further the potential of hybrid approaches for feature selection.

II. PRELIMINARIES

A. Hybrid feature selection framework
The schematic diagram for hybrid gene selection procedure is shown in Fig. 1. First, the filter method selects n independent genes from the available gene set. From the selected genes, wrapper method finds the optimal gene subset using searching algorithm such that the performance of the classifier is improved.

Figure 1 Hybrid Model for Feature Selection

B. Extreme Learning Machine
Extreme learning machine (ELM) is a new learning paradigm for training single hidden layer feedforward neural network [23]. Compared to other conventional neural networks, ELM increases its learning speed by choosing hidden layer parameters randomly rather than iteratively. Later calculates its output weights analytically. ELM has recently attracted tremendous attention from various field such as remote sensing, sensor network, internet, business, biology, etc. for its extreme learning speed with good generalization performance. A brief overview of ELM is given below.

Given a training set with N different samples \((x_j, t_j)\) where \(\{x_j, t_j\} \in R^n, t_j \in R^m\) where \(x_j\) is an \(n \times 1\) input vector and \(t_j\) is an \(m \times 1\) target vector. Then the output of ELM network with M hidden neurons and activation function \(g(x)\) are defined as

\[
\sum_{i=1}^{M} a_i g_i(x_j) = \sum_{i=1}^{M} a_i g_i(w_i, x_j + b_i) = O_j \quad \text{where} \quad j = 1, ..., N
\]

(1)

Where \(w_i = [w_{i1}, w_{i2}, w_{i3}, ..., w_{iM}]^T\) is the weight vector connecting the input nodes and the \(i^{th}\) hidden nodes \(\alpha_i = [\alpha_{i1}, \alpha_{i2}, ..., \alpha_{iM}]^T\) is the weight vector connecting the \(i^{th}\) hidden nodes and the output nodes and \(b_i\) is the bias of the \(i^{th}\) hidden node. Assuming that the ELM activation function approaches all N samples by zero error as

\[
\sum_{j=1}^{N} \| O_j - t_j \| = 0
\]

(2)

Then there exist parameters \((w_i, b_i)\) and \(\alpha_i\) such that

\[
\sum_{i=1}^{M} a_i g_i(w_i, x_j + b_i) = t_j \quad \text{where} \quad j = 1, ..., N
\]

(3)

The above N equations can be written compactly as \(H\alpha = T\). Where the hidden layer output matrix \(H\) of the neural network is given as

\[
H = \begin{bmatrix} g(w_1 \cdot x_1 + b_1) & ... & g(w_M \cdot x_1 + b_M) \\ g(w_1 \cdot x_2 + b_1) & ... & g(w_M \cdot x_2 + b_M) \\ \vdots & \ddots & \vdots \\ g(w_1 \cdot x_N + b_1) & ... & g(w_M \cdot x_N + b_M) \end{bmatrix}_{N \times M}
\]

(4)

In ELM, the parameters of hidden-layer nodes, i.e., \(w_i\) and \(b_i\), can be chosen randomly without knowing the training datasets. The output weight \(\alpha\) is then calculated with matrix computation formula \(\alpha = H^T\) where \(H^T\) is the Moore–Penrose inverse of \(H\). The ELM algorithm is given below.

Algorithm 1: ELM

Input: it is a training dataset \((x_j, t_j)\), an activation function \(g(x)\), and the number of hidden nodes \(M\).

Output: it is the weights of hidden layer output layer \(\alpha_i = [\alpha_{i1}, \alpha_{i2}, ..., \alpha_{iM}]^T\).

Step 1: Randomly generate weights and bias \((w_i, b_i)\) \(i = 1, 2, ..., M\)

Step 2: Calculate the hidden layer output matrix \(H\)

Step 3: Calculate the hidden weights using \(\alpha = H^T\)
In contrast to artificial neural networks, ELM has small number of tuning parameters. But random generation of these parameters affects the performances of ELM. Evolutionary research techniques are employed in recent years to address this random selection effect and obtain higher classification accuracy.

C. Particle Swarm Optimization

PSO is an evolutionary optimization technique based on swarm intelligence developed by Kennedy and Eberhart in 1995 [24]. Unlike other evolutionary algorithms, PSO neither has complicated evolutionary operator nor has many parameter for tuning and solves the optimization problem exploiting the feeding characteristics of animals, birds and fish. In recent years, it has proven to solve diverse optimization problems with quick convergence rate and has attracted much attention of many researchers worldwide [13].

PSO algorithm starts with a population (called swarm) of random solutions. This initial population evolves iteratively to find optimal solution for the problem to be optimized. Each individual (called particle) in the population corresponds to a fitness value determined by the function to be optimized and has a velocity, which enables them to move through the search space. Thus, each particle represented by its position \( x_i(t) \) and velocity \( v_i(t) \). During the optimization, particles keep updating its position themselves using its previous position and its current velocity. The current velocity is determined using two cognitive aspects, individual learning (pBest) and learning from a social group (gBest) to move towards the global optimal solution of the problem. These principles are formulated as

\[
x_i(t+1) = x_i(t) + v_i(t+1)
\]

\[
v_i(t+1) = \omega v_i(t) + c_1 r_1 (p_{best} - x_i(t)) + c_2 r_2 (g_{best} - x_i(t))
\]

Algorithm 2 : PSO

1. Initialize the population \( S \) with solution space;
2. Repeat:
   a. Evaluate fitness of each particle position \( (p) \) using objective function
   b. If fitness\( (p) \) > pbest then pbest = fitness\( (p) \).
   c. Set gbest = best of pbest
   d. Update velocity and position of each particle
3. Until convergence is achieved or termination criteria are satisfied
4. Final gbest is optimal solution

III. PROPOSED INTELLIGENT HYBRID MODEL

This article presents an intelligent hybrid model shown in Fig-2 to identify small subset of features which can determine the target class for the given input samples with higher accuracy and lesser computation complexity. The hybrid model works in two stages as described below,

A. Feature Preselection

Generally, the original training dataset are high dimensional with not only informative features but also with redundant and irrelevant features. The key proposal in this stage is to use ensemble of filters instead of one filter to remove unimportant features and increase the stability of the selected features. The reason for this proposal are twofold. First, it enables to overcome the challenge of choosing of a single correct filter for given application based on dataset characteristics. Second, it enables to introduce diversity in feature selection and avoid the possibility of leaving out any important features to be considered in subsequent stages.

This stage starts by applying several filter methods based on different metrics to have diverse set of features. Each filter evaluates each feature (gene) in the training dataset independently and selects predefined number of features based on their discriminative capability without compromising the time and memory requirement of the learning model. Finally the feature subset selected by each filter are combined using union operation.

B. Feature Reselection

In this stage, the wrapper method attempts to improve the classification accuracy taking into consideration two factors, (i) Reselect only compact subset of features but maintains highest classification accuracy and (ii) Optimize the input weights vector and bias vector of ELM structure. An hybrid binary-real PSO algorithm is adopted to solve the above two factors with following optimization components,

Particle Encoding: Each particle in the swarm is encoded with Q discrete values representing the feature selection state and P real values representing the ELM parameters and Q as shown below,

\[
\hat{x} = \left[ w_1, w_2, ..., w_Q, \left[ b_1, b_2, ..., b_Q \right], \left[ F_1, F_2, ..., F_P \right] \right]
\]
Here, the dimension of \( P = (N_i \times N_h) + N_b \), with \( N_i \) and \( N_h \) denoting the number of input neurons and hidden neurons, and \( Q = nf \) with \( nf \) denoting the number of features in XOR part. Since the dimension of input weight vector \((N_i \times N_h)\) and bias vector \((N_b)\) depends on number of hidden neurons \( (N_h)\), this stage as first step determines \( N_h \) using grid search method \((\text{GS})\) in similar fashion as in [25]. Accordingly, the weight vector and bias vector are initialized for optimization process.

**Fitness function:** During optimization process, the fitness of each particle for next generation is computed as in [14] as follows

\[
\text{fitness} (i)=10^4 \times (1 - \text{accuracy} (i)) + k \times \text{GenesNumber}(i) 
\]

Here accuracy(i) is the five-fold cross validation classification accuracy on training data obtained by a ELM structure defined by real part of \( i \)th particle with candidate genes denoted by the binary part of \( i \)th particle along with the candidate gene represented by intersection part. The Genes-Number \((i)\) is the number of selected genes denoted by the \( i \)th particle plus the number of features in intersection part. The parameter \( k \) is a weighted coefficients satisfying the condition \( k > 0 \), and tries to keep a balance between the classification accuracy and the number of the selected genes. As well it gives larger fitness values to feature subset with smaller number of features If the two feature subsets have a same accuracy.

Using the above optimization components and following the steps in algorithm 2, the best and gbest is obtained. Both binary and real part share the same pbest and gbest. Also the velocity update for both binary and real PSO is same as defined in equation (1) but the position update of binary PSO is different from real PSO since it takes only binary values. Therein the position of binary PSO are updated via sigmoid limiting function as applied in [11]–[13]. Other optimization parameters used in this work are given in Table-1. Ultimately, the outputs of this stage are the optimal feature subset and an optimal ELM model. Now, the cancer classification with higher accuracy can be obtained using the indexes of the optimal features from testing datasets as new inputs of the trained ELM model.

### IV. EXPERIMENTAL SETUP AND RESULT DISCUSSION

In this section, we describe experimental design used to evaluate the performance of the proposed method. First subsection deliberates the datasets and parameter settings used in the experiments. This is followed by results and discussion subsection.

#### A. Datasets and Parameter Settings

Two types of biomedical datasets were used in the present study to investigate how the feature selection methods respond to different data structures. The effectiveness and efficiency of the proposed intelligent hybrid feature selection method was evaluated over six benchmark cancer datasets of microarray gene expression data. All of them are publicly available and the main characteristics of these datasets are summarized in Table I. The datasets were normalized with zero mean and unit variance to prevent the gene with higher expression levels from dominating the feature selection or classification processes.

<table>
<thead>
<tr>
<th>DATASETS</th>
<th>NUMBER OF SAMPLES</th>
<th>NUMBER OF GENES</th>
<th>NUMBER OF CLASSES</th>
</tr>
</thead>
<tbody>
<tr>
<td>Colon [26]</td>
<td>2000</td>
<td>62</td>
<td>2</td>
</tr>
<tr>
<td>Breast [27]</td>
<td>24482</td>
<td>19</td>
<td>2</td>
</tr>
<tr>
<td>Prostate [28]</td>
<td>12600</td>
<td>34</td>
<td>2</td>
</tr>
<tr>
<td>Lymphoma [29]</td>
<td>4026</td>
<td>62</td>
<td>3</td>
</tr>
<tr>
<td>SRBCT [30]</td>
<td>63</td>
<td>2308</td>
<td>4</td>
</tr>
</tbody>
</table>

As first step of experimental study, the parameter tuning for the proposed algorithm was conducted to improve its performance concerning the utilization of HPSO. In this process, the population size of HPSO was set to 100 with the maximum number of iteration number of 50. Both the acceleration parameters \( c_1 \) and \( c_2 \) were selected as 1.49 for both real and binary part of HPSO. The parameters that were unique for real and binary part of HPSO are shown in Table II. These parameters were chosen by trial and error to provide results with high degree of confidence and inline with many related work which utilizes HPSO [31,32].

<table>
<thead>
<tr>
<th>PARAMETERS</th>
<th>VALUES</th>
</tr>
</thead>
<tbody>
<tr>
<td>Swarm Size</td>
<td>100</td>
</tr>
<tr>
<td>Number of iterations</td>
<td>50</td>
</tr>
<tr>
<td>( C_1 ) and ( C_2 )</td>
<td>1.49</td>
</tr>
<tr>
<td>REAL PSO</td>
<td>6</td>
</tr>
<tr>
<td>BINARY PSO</td>
<td>1</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Vmax</th>
<th>0.1 x (R_{pmax},R_{pmin})</th>
</tr>
</thead>
<tbody>
<tr>
<td>Weight Inertia</td>
<td>0.9 to 0.3</td>
</tr>
</tbody>
</table>

#### B. Ensemble Design for Feature Preselection

This section of our work for the first time in the literature aims to design an ensemble for feature preselection incorporating diversity of filters to exclude as many irrelevant features as possible. But at the same time not to eliminate any informative features (biomarker) at the pre-selection stage. In achieving this, the present work considers three filters from the broad suite of filters available in literature. These filters are chosen based on different scoring criteria and their wide application in many fields. The theoretical background of the three chosen filters are given below.

1. Fisher Ratio (FR): is a supervised method that selects significant genes based on their class discrimination power [33]. The fisher ratio for each gene is the ratio of between-class scattering to within-class scattering of the respected feature as defined in the following equation.

\[
FR(g_i) = \sum_{j=1}^{c} S_j (\mu_{ij} - \mu_i)^2 / \sum_{j=1}^{c} S_j \sigma_{ij}^2 
\]

Where \( \mu_i \) is the mean vector of the class \( i \) and \( \mu_{ij} \) is the mean vector of the class \( j \).
Where $\mu_i$ is the mean of gene $i$, $S_j$ is the number of samples in class $j$, $\mu_{i,j}$ and $\sigma_{i,j}$ is mean and variance of gene $g_i$ in class $j$. The gene with higher fisher score are selected for feature reselection stage. Motivated by its generalization performance and its robustness to noise and data scarcity [33], FR was considered in this work.

2. Significant Analysis of Microarray (SAM): is a statistical technique that selects genes based on its change in gene expression relative to the standard deviation of repeated measurements for that gene [34]. SAM computes a statistic $d_i$ for each gene $i$, measuring the strength of the relationship between gene expression and the target class as defined below,

$$SAM(i) = \frac{\mu_{i1} - \mu_{i2}}{\sigma_i}$$

$$\sigma_i = \sqrt{\frac{1}{n_{i1}} + \frac{1}{n_{i2}}} \frac{\sigma_{i1}^2 + \sigma_{i2}^2}{n_{i1} + n_{i2} - 2}$$

Where $\mu_{i1}$ and $\mu_{i2}$ are the means of gene expression for gene $i$ in the two target groups. From the above equation, the gene with higher SAM score are regarded as potentially significant and are selected for feature reselection stage. The robust performance of SAM for normality assumption and for controlling the false discovery rate [35] motivated to consider SAM in this work.

3. Information Gain (IG): selects the most informative gene based on its highest correlation with the target class [36]. IG value of a gene $X$ indicates the reduction of uncertainty and is computed based on the entropy of the dataset as follows

$$IG(X,Y) = H(X) - H(X|Y)$$

$$H(X) = \sum P(x) \log(P(x))$$

$$H(X|Y) = \sum P(y) \sum P(x|y) \log(P(x|y))$$

Where $H(Y)$ and $H(X|Y)$ denotes the entropy of dataset $Y$ and conditional entropy on the known gene $X$. Thus the gene with higher IG value is deemed to contribute more class discriminative information. Its performance in many real world application in quickly eliminating large number of irrelevant genes encouraged to consider IG in this work.

**C. Results and Discussion**

Performance study of the Proposed Method: This section demonstrates the efficiency of the proposed hybrid framework over the above described 5 microarray datasets. At first, the designed ensemble of filters discussed in section II is applied to select 100 most significant genes and reduce the search space complexity of the proposed hybrid framework. Then, HPSO with above defined parameter is employed to synchronously select compact gene subset from these 100 genes and optimize the ELM classifier to achieve the highest classification accuracy. Finally, optimized ELM classifier is utilized for cancer detection. Table III shows the performance of the proposed hybrid framework in terms of the number of genes selected and accuracy of the optimized ELM classifier. To demonstrate the stability of the proposed hybrid framework, stability index computed using the equation given below and is listed in the last column of Table III. All these results were obtained by conducting 10-fold cross validation to avoid the over-fitting and to confirm the results is more reliable. From these results, it can be noted that the proposed hybrid model shows highest classification accuracy on most of the evaluation datasets with most compact subset of informative genes. The stability index values ensures that proposed hybrid model is more stable in providing confidence on the compact gene subset returned.

<table>
<thead>
<tr>
<th>DATASETS</th>
<th>NUMBER OF SELECTED GENES</th>
<th>ACCURACY %</th>
<th>STABILITY INDEX</th>
</tr>
</thead>
<tbody>
<tr>
<td>Colon</td>
<td>5</td>
<td>100</td>
<td>0.88</td>
</tr>
<tr>
<td>Breast Cancer</td>
<td>6</td>
<td>87.5</td>
<td>0.81</td>
</tr>
<tr>
<td>Prostate</td>
<td>5</td>
<td>100</td>
<td>0.87</td>
</tr>
<tr>
<td>Lymphoma</td>
<td>2</td>
<td>100</td>
<td>0.91</td>
</tr>
<tr>
<td>SRBCT</td>
<td>6</td>
<td>100</td>
<td>0.83</td>
</tr>
</tbody>
</table>

Comparative study with other feature selection Methods in Literature: This section describes the performance of proposed algorithm for five microarray datasets in comparison with most recent five hybrid feature selection methods reported in literature. Table IV summarizes the performance of the proposed algorithm along with the best results reported in the respective literatures for the five microarray datasets being used for comparison. From this Table, it is quite evident that proposed algorithm performs better than the literature methods in selecting less number of genes with comparably good CV classification accuracy. For instance, on multiclass datasets, SRBCT and Lymphoma, the proposed algorithm significantly outperforms identifying the lowest reported number of genes required to achieve 100% accuracy. The remarkable results of the proposed method concerns the breast cancer results where the proposed algorithm satisfying the two objectives (less number of gene and good classification accuracy) identifies the lowest reported number of genes but only with 87% accuracy. Overall, the comparison opines that the proposed approach enables to identify lowest number of genes to achieve comparable accuracy for all the microarray datasets.
In this paper, an effective two phase hybrid feature selection model for multi-class cancer classification is presented. The model uses ensemble of filters in first phase to identify all relevant genes. Then HPSO is employed in wrapper phase to select optimal gene subset. The main novelty of this model is two-fold: First, its capability to select all relevant genes without eliminating any biomarkers for consideration in the second phase. Second, its ability to synchronously select a compact gene subset and optimize ELM structure to achieve optimal classification accuracy. The 10-fold cross validation was used to evaluate the effectiveness of the proposed model and results were compared with five recent recommended hybrid feature selection approaches in literature. Experimental results confirms the superior performance and stability of the proposed hybrid model over other existing methods in terms of the accurate cancer classification and the number of genes selected. Thus the proposed model could be utilized as an efficient tool to reduce gene dimensionality in microarray data analysis. In future, this research will be extended to investigate the applicability of proposed model with other high dimensional datasets and improve the efficiency in terms of computational complexity.

### REFERENCES


