CLASSIFICATION OF GENE EXPRESSION DATA WITH THE AID OF OPTIMIZED FEATURE SELECTION

Pankaj Bhambri¹, Manpreet Singh², Amit Jain³, Inderjit Singh Dhanoa⁴, Vijay Kumar Sinha⁵, Shyam Lal⁶

¹Assistant Professor, Department of Information Technology,
Guru Nanak Dev Engineering College, Ludhiana. pbhambri@gmail.com
²Assistant Professor, Department of Information Technology,
³Assistant Professor, Department of Computer Science & Engineering,
⁴Assistant Professor, Department of Computer Science & Engineering, Guru Nanak Dev Engineering College, Ludhiana.
⁵Associate Professor, Department of Computer Science & Engineering, Chandigarh University, Chandigarh.
⁶Assistant Professor, Department of Mathematics,
Akal Degree College, Sangrur.

ABSTRACT

A Gene Expression (GE) analysis has received a lot of interest from researchers in the bio-informatics field. The GE degree allows for the likelihood of diagnosing multiple illnesses, such as cancer. However, the presence of a large number of genes and very few available patient samples hinders the classification or Machine Learning techniques from producing accurate classification results. Most of these are immaterial and redundant, which may depreciate the performance of the classification. This work proposes an efficient technique for classifying GE Data (GED) with the assist of Hybrid PSO-GA optimized feature selection. Firstly, the data-set is pre-processed by performing normalization of the dataset. Then Feature Selection (FS) is performed with the aid of Hybrid PSO-GA Optimization, in which Particle Swarm Optimization along with Genetic Algorithm are hybridized. The fact is that not every characteristic is necessary for classification, and certain obsolete and immaterial attributes that also act as outliers. To dispose the outliers, Feature Reduction (FR) is done with the aid of Euclidean Distance (ED). And finally, classification is executed utilizing the Artificial Neural Network (ANN). The experimental outcomes illustrate that the proposed work effectively classifies whether the sample is normal or cancer affected and can obtain better classification results on considering the conventional techniques.

Keywords: Gene Expression Analysis, Bioinformatics, Euclidean Distance

I. INTRODUCTION

Bioinformatics (BI) is an interdisciplinary research field that involves new algorithms and applications for the representation of biological data. Some of the key problems in this field involve data collection / generating, sorting, preservation and retrieval, as well as the security of biological data. It extends to the pharmaceutical, academia, biotechnology, pharmaceutical, agrochemical and food industries and numerous other related sectors. BI is a fast-growing research field because of the DNA microarrays technology’s arrival. In the area of BI, Genomics deals with the entire set of genetic material inside an organism. GE micro-array is a higher-throughput genomic technology in research and clinical management that allows recording and monitoring the thousands of genes' expression levels concurrently within a few disparate samples. The advent of DNA micro-array technology has made it possible for biologists to investigate the behavior of gene chilaides (features) in one experiment. These data provide useful knowledge, including the GE stage, which may well be used as a guide for cancer diagnosis and classification.
The application of micro-array technology has emerged as the most groundbreaking breakthroughs in molecular biology, especially in the identification of cancer cells. Nevertheless, utilizing the data contained in the microarray genome, the diagnosis or identification of cancer is a big problem for computer scientists. This is since, in fact, the amount of genes involved in the microarray data is in the range of thousands. But the amount of samples of microarrays is on the order of thousands, as the procedure is highly expensive and time-consuming. The research gap in GED classification is to solve the higher dimensionality of the sample in limited sample sizes and the experimental variations as per the different GE levels. For microarray gene profiles, mRNA tests are used to calculate the degree of expression of the proteins, which can be in the hundreds of thousands. It, in effect, renders identification and analysis challenging due to the higher data dimensionality.

Among the molecular cancer info, GE is the most widely used cancer detection tool because of the fact that tumor tissues often have a particular pattern distinct from standard GE tissues. The extent of expression of a gene is similar to the total copies of the RNA formed in a cell; in addition, it is related to the amount of equivalent protein created. It indicates that the exact pattern of GE occurs in different biological states, such as cell development, and during normal physiological reactions. Scrutiny of the GE profile can provide insight into the classification defined on the GE indications of the cells under consideration.

The issue of efficient retrieval of GE patterns in larger datasets has gained extensive attention from the data mining society because of its application to BI and medical fields. Modern cancer detection is based on clinical and morphological evidence, although these approaches have been documented to have restricted diagnostic capabilities. Cancer diagnosis based on genomic data was used to resolve certain established drawbacks. The microarray can gauge the expressions of several chilidiads of genes concurrently. This ability makes the microarray a more influential tool than other customary methods for diagnosis, accurate class prediction and identification of subclasses. While many ML methods have been applied, the classification of micro-array data is still a very difficult job as the data include noise and missing genes. Therefore, the question of correct labeling of data is of concern.

Section 2 describes the surveys of the relevant works on the planned research. A detailed debate on the suggested approach is provided in Section 3; Section 4 analyzes the findings of the inquiry and Section 5 presents the conclusion of this article.

II. CLASSIFICATION OF GENE EXPRESSION DATA USING OPTIMIZED FEATURE SELECTION

GE micro-array is a higher-throughput genomic technology in research and clinical management that allows recording and monitoring the expression levels of chilidiads of genes concurrently within a few different samples. In general, GE micro-arrays comprise a vast number of genes and extremely few samples that represent a grave challenge for disease prediction and also the diagnosis. Because cancer and normal tissue have specific GE, GED may well be used as an efficient cancer detection tool. Nevertheless, the correct diagnosis of cancer using the initial GE remains difficult due to the higher-dimensional inherent function and also the smaller scale of the test samples. This work proposes an effective technique for classifying GED with the assist of Hybrid PSO-GA optimized FS. Firstly, the data-set is pre-processed by performing normalization of the dataset. Then FS is done with the aid of Hybrid PSO-GA Optimization, in which the PSO Algorithm and Genetic Optimization Algorithm are hybridized. The fact is that not all characteristics are important for classification; in addition, certain obsolete and immaterial features can also act as outliers. To get clear of the outliers, FR is done with the aid of ED. And finally, classification is done using ANN, which classifies the sample as a normal gene or cancer affected gene. The proposed methodology’s framework is revealed in Fig 1.
Let \( V = [v_1, v_2, \ldots, v_n] \in \mathbb{R}^{d \times n} \) is the initial data collection in which each column is the d-dimensional vector of the sample and the total number of samples. The suggested approach first embraces GE Dataset as input patterns and proceeds to take more steps.

### 2.1 Pre-Processing

The pre-processing stage is combined to render the GED appropriate for FS and classification and to let the algorithm run faster. Pre-processing helps to suppress the genes in tissue samples that are found in ambiguities. In pre-processing, normalization of the dataset is done. The normalization is applied on every facet of the provided input data matrix.

#### 2.1.1 Normalization

The critical stage of pre-processing data is to convert all symbolic features to numerical values that are supposed to be normalized. Data normalization is a method of multiplying the importance of each element into a well-proportioned gamut, such that the bias in favor of attributes of greater values is removed from the dataset.

Normalizing the data means that every attribute is assigned an equal weight. Each attribute in each record shall be normalized by means of the corresponding maximum value and shall fall within the same range as \([0, 1]\). Normalization therefore reduces the training error, thereby proving the accuracy of the classification issue. At this point, the expression rates for each gene are normalized to \([0, 1]\) using the standard protocol shown in equation 1.

\[
v = \text{lower} + \left[ \text{upper} - \text{lower} \times \frac{\text{val} - \text{val}_{\min}}{\text{val}_{\max} - \text{val}_{\min}} \right]
\]

Here, among all the gene’s expression levels in consideration, \(\text{val}_{\max}\) is the maximum original value, \(\text{val}_{\min}\) is the minimum original value, \(\text{upper} \ (\text{lower})\) is 1 (0) and \(g\) is the normalized expression level. So for all genes subsequent to normalization, \(\text{val}_{\max}\) will be ‘one’ and \(\text{val}_{\min}\) will be ‘zero’.

### 2.2 Feature Selection

This is defined as a method that selects a limited subset of features from the initial category of features, such that the space function decreases the maximum emphasis on such selection criteria. The FS is used with an array of valuable knowledge from broad databases, and meta-heuristic algorithms will look for an optimal answer by scanning the correct areas of the storage room. The key goal of optimizing the classification accuracy rate with the smaller size of the apps is known to be the optimization issue. It is a mechanism that allows the most descriptive genes to be selected and can precisely differentiate classes of cancer forms. The suggested methodology uses a mixture of PSO and Genetic Optimization Algorithm for the collection of functions.

#### 2.2.1 Hybrid PSO-GA Optimization

It is the mixture of the PSO and the GA. The steps below are given:

www.turkjphysiotherrehabil.org
Step 1: Assign the input population / particle parameters to random locations and also the velocity in the d dimensions of the issue area. Every parameter is treated as a particle.

Step 2: Evaluate the desired fitness optimization feature on d variables for each particle.

Step 3: Compare particles’ fitness assessment with particles $p^{\text{best}}$. If its value is enhanced from $p^{\text{best}}$, then set $p^{\text{best}}$ value equivalent to the current value and set the $p^{\text{best}}$ location with the current position on d-dimension space.

Step 4: Build the latest best fitness value (BF) with the total inputs considered. If this current BF value is enhanced on considering the global best $g^{\text{best}}$ then allot the current BF value to $g^{\text{best}}$ and the current coordinates to $g^{\text{best}}$ coordinates.

Step 5: Pick a particle with a BF meaning, re-initialize its location. In addition, evaluate the particle with the worst fitness value and decide if its new place obtained is sufficient. If it is in an acceptable range then update this position or else a new position is allotted to the particle randomly in its neighbor and then renew the position $p$, together with a velocity $x$, of other particles utilizing the expression proffered below,

$$x_i(t_{n + 1}) = x_i(t_n) + d_1e_1(g_i(t_n) - h_i(t_n)) + d_2e_2(g_i(t_n) - h_i(t_n))$$  

$$p_i(t_{n + 1}) = p_i(t_n) + x_i(t_{n + 1})$$  

In the equ (2), $d_1$ and $d_2$ signifies the acceleration constants that are requisite for integrating every particle with the $p^{\text{best}}$ along with $g^{\text{best}}$. Once the velocity together with the position is evaluated, crossover & mutations are done to make the optimization more effective.

Step 6: The two-point cross-over is chosen from the different forms of crossovers. In the two-point cross-over, 2 points are chosen for the parent chromosomes. The genes at the two points are interchanged with the parent chromosomes. Therefore, the children are produced by chromosomes. The cross-over points shall be calculated as follows:

$$x_1 = \frac{|v^{(i)}_p|}{3}$$  

$$x_2 = x_1 + \frac{|v^{(i)}_p|}{2}$$  

These children's chromosomes are processed separately and their associated indexes are preserved.

Step 7: The mutation is then done by replacing a variety of genes with new genes for each chromosome. Substituted genes are spontaneously generated genes with no recurrence in the genome. After that, chromosomes that are selected for cross-over operation, along with chromosomes that are reached as mutations, are joined together and, as a result, the population pool is packed with chromosomes.

Step 8: The process is replicated before a solution with a greater health benefit is produced. The customized apps are therefore acquired.
2.3 Feature Reduction

FR deals with retaining the relevant information while reducing the measure of information necessary to represent the variation. In FR, the optimized features are reduced utilizing ED.

2.3.1 Euclidean distance

FR is centered on the calculation of ED between the optimized features. ED is a standard metric for geometrical problems. This selects the necessary features centered on the distance measure for the optimal classification. The ED between the features is stated as,

\[ D_k = \sqrt{\sum_{k=1}^{n} (u_k - v_k)^2} \quad (6) \]

Where, \( n \) implies the number of dimensions (attributes), \( D \) implies the ED, and \( (u_k - v_k) \) is the difference between two features.

The values of \( D_k \) show the average distance between the two optimized features.

2.4 Classification using ANN

ANN is a training algorithm that can be analyzed to address complicated problems such as training results, and requires a collection of input pairs along with desired outputs (targets). This consists of a number of neurons (signified by functions) connected to others arranged on separate layers where each layer consists of neurons.
problem to be addressed is the input patterns that are sent through the layers. The data is represented using equal synaptic weights.

For ANN, the weights are successively adjusted on the basis of the input category and the corresponding set of output objectives needed. Synaptic weight adaptation consists of growing its importance until the desired behavior is reached. Improved repetition consists of 2 bars: forward activation to generate an answer, and backward replication of the measured error to adjust weights. The forward and backward sweeps are conducted continuously until the ANN solution responds to the appropriate value inside the pre-stated tolerance.

Back propagation algorithm is used to train the neural network, which is defined in the following sections.

**Step 1:** Generate arbitrary weights within the range [0, 1]. Assign it to the secret & outgoing neuron row. Maintain a unit weight value for the whole neurons of the input sheet.

**Step 2:** Pass the training data to the classifier and evaluate the BP error as follows

\[ BP_{err} = Z_{tar} - Z_{out} \]  

(7)

In Eq. (1), \( Z_{tar} \) implies the target output and \( Z_{out} \) is the network output, which can be determined as

\[ Z_{out} = [X_{2}^{(1)} \ X_{2}^{(2)} \ldots \ X_{2}^{(N)}], \]

\( X_{2}^{(i)}, \ X_{2}^{(2)}, \ldots, X_{2}^{(N)} \) are the network outputs. The network outputs can be determined as

\[ X_{2}^{(i)} = \sum_{r=1}^{N_{i}} w_{2r1} X_{1}(r) \]  

(8)

where,

\[ X_{1}(r) = \frac{1}{1 + \exp(-w_{11r} Z_{in})} \]  

(9)

Eqs. (2) along with (3) implies the activation function performed on the output layer and hidden layer correspondingly.

**Step 3:** Regulate the weights of the entire neurons as

\[ w = w + \Delta w \]

where, \( \Delta w \) is the change in weight which can well be determined as

\[ \Delta w = \gamma \cdot X_{2} \cdot BP_{err} \]  

(10)

In Eq. (4), \( \gamma \) implies the learning rate, usually, it ranges from 0.2 to 0.5.

**Step 4:** Repeat the process from step 2, until BP error gets minimized to the least value. Practically, the criterion to be met is \( BP_{err} < 0.1 \).

3. RESULT AND DISCUSSION

This segment describes the efficiency evaluation of the PSOGA ANN GE dataset classification on the basis of automated FS. The new system utilizes the repositories of the National Center for Biotechnology Information (NCBI) to provide links to medicinal and genomic knowledge. GE Omni-bus (GEO) is a service provided by the GSPI. This is an multinational public library that collects and often liberally distributes higher-throughput GE and other usable genomics datasets. The sample amount, the gene amount and the designated parts for the data on leukemia and colon tumors are displayed in Table 1.

**Table 1:** Gene Expression Profile

<table>
<thead>
<tr>
<th>Data Set</th>
<th>Sample Number</th>
<th>Gene Number</th>
<th>Sample</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>Positive</td>
</tr>
</tbody>
</table>

www.turkphysiotherrehabil.org
As seen in Table 1, the GE profile of each study contains two labelled groups, respectively positive and negative for the Leukemia and Colon tumor dataset. Colon Tumor data collection consists of 62 samples and the Leukemia data set includes 38 samples.

### 3.1 Performance Metrics

The tests are calculated in comparison to predictive success metrics such as Positive instances and even Negative instances, including True Positive, True Negative, False Positive and even False Negative. True Positive (TP) is the highest positive condition accurately treated. Real Negative (TN) is a true adverse incident that has been correctly treated. False Positive (FP) is assumed to be the cumulative bad instances defined as positive. False Negative (FN) stands for the total positive instances identified as negative. These metrical outputs are mainly computed and then used to quantify Accuracy, Specificity, Precision, Recall and F-Measurement of the algorithm.

\[
\text{Accuracy} = \frac{(TP + TN)}{(TP + FP + TN + FN)} \tag{11}
\]
\[
\text{Specificity} = \frac{TN}{FP + TN} \tag{12}
\]
\[
\text{Sensitivity} = \frac{TP}{TP + FN} \tag{13}
\]
\[
\text{GMean} = \sqrt{\text{Specificity} \times \text{Sensitivity}} \tag{14}
\]

### 3.2 Performance Evaluation

The performance of the Hybrid PSO GA based ANN (PSOGA ANN) classification is weighted down by classifiers such as Support Vectors Machines (SVM), Naïve Bayes (NB), K Nearest Neighbors (KNN) for Accuracy, Sensitivity, Specificity, and Geometric Mean (GMean).

#### 3.2.1 Leukemia dataset

The performance metrical like Sensitivity, Accuracy, Specificity, along with GMean is calculated for the Leukemia data-set. The proposed PSOGA_ANN technique is contrasted with the existent classifiers say SVM, NB, and KNN.
Figure 3: Comparison of current and planned classifiers in terms of Accuracy for Leukemia Dataset

Figure 3 analyzes the existing classifiers’ performance like NB, SVM & KNN and the proposed PSOGA_ANN classifier in respects of accuracy. It can be inferred that the existing NB encompass the lowest accuracy and the proposed PSOGA_ANN has the highest accuracy. Compared with NB, the SVM & KNN accuracy is high, but the proposed technique shows the highest accuracy than the other existing classifiers.

Figure 4: Comparison of current and planned classifiers in terms of Specificity and Sensitivity for Leukemia Dataset

Figure 4 analyzes the existing classifiers' performance such as NB, SVM, and KNN and the proposed PSOGA_ANN classifier in respects of specificity along with sensitivity. On considering Specificity, the existing SVM shows very poor performance, whose specificity value is 0.21. Also, the existing NB and the proposed PSOGA_ANN shows approximately similar specificity values. Yet, the proposed technique shows the highest specificity. Now, considering Sensitivity, the existent NB performance is very low and the sensitivity of the existent SVM and the proposed PSOGA_ANN are approximately similar. Yet, the proposed technique shows the highest sensitivity.

Figure 5: Comparison of current and planned classifiers in terms of GMean for Leukemia Dataset

Figure 5 examines the existent classifiers’ performance like NB, SVM along with KNN and the proposed PSOGA_ANN classifier in respects of GMean. GMean implies the balance betwixt classification performances in the minority and majority class. This metric ponder both the sensitivity, which is the accuracy in the positive cases and also the specificity, which is the accuracy on the negative examples. This can be concluded from the graph that the current NB has the lowest GMean and the new PSOGA ANN has the maximum GMean.

3.2.2 Colon cancer dataset

www.turkjphysiotherrehabil.org
Quality measures such as Sensitivity, Accuracy, Specificity, and GMean are computed for the Colon Cancer data collection. The suggested PSOGA ANN methodology is compared with current classifiers such as NB, SVM, and KNN.

Figure 6: Comparison of current and planned classifiers in terms of Accuracy for Colon Cancer Dataset

Figure 6 analyzes the existing classifiers’ performance like NB, SVM, KNN and the proposed PSOGA_ANN classifier in respects of accuracy evaluated using colon cancer dataset. It can well be reasoned that the existing NB shows lower performance on considering the remaining classifiers. Also, the existing KNN and also SVM encompass the accuracy values of 0.58 and 0.61 respectively. But, the proposed POSGA_ANN shows the uppermost performance when compared with another existing classifiers.

Figure 7: Comparison of current and proposed classifiers with respect to specificity and Sensitivity for Colon Cancer Dataset

Figure 7 analyzes the present classifiers' performance such as NB, SVM, and KNN and the proposed PSOGA_ANN classifier in respects of specificity and sensitivity evaluated using colon cancer data-set. For Specificity, the classifiers have almost acceptable specificity values. But among all those contrasted classifiers, the proposed technique has the greatest specificity values. For Sensitivity, the existing NB has very poor sensitivity and the sensitivity of SVM and KNN are almost similar. Yet, the proposed PSOGA_ANN has the greatest sensitivity values.
Figure 8: Comparison of current and potential classifiers with regard to GMean for Colon Cancer Dataset

Figure 8 compares the existing classifiers' premiere and the potential PSOGA_ANN classifier in respects of GMean evaluated using colon cancer data-set. It can well be perceived from the figure that the existing NB has the lowest GMean value and so it shows poor performance. Contrary to NB, the proposed PSOGA_ANN shows the highest GMean values and thus it reveals superior performance.

III. CONCLUSION

The explosion of DNA microarray data-set in the scientific repository is encouraging inter-disciplinary research on computer science, ecology, and BI. GE micro-arrays contain a vast number of genes and very limited tests, which is a crucial barrier to disease detection and diagnosis. This work proposes an efficient technique for classifying GED with the aid of Hybrid PSO-GA optimized FS. The proposed work is proven to be an effective way for classifying the GED as a normal one or cancerous one. The proposed approach’s performance is estimated centered on the metrics say, sensitivity, accuracy, specificity together with GMean. The evaluated finding indicates that the new approach is best adapted to all the other current strategies.

REFERENCES