Hybrid Dimension Reduction Techniques with Genetic Algorithm and Neural Network for Classifying Leukemia Gene Expression Data

S. Jacophile Susmi*, H. Khanna Nehemiah¹ and A. Kannan²

¹Ramanujan Computing Centre, Anna University, Chennai – 600025, Tamil Nadu, India; jacophile.susmi@gmail.com, nehemiah@annauniv.edu
²Department of Information Science and Technology, Anna University, Chennai – 600025, Tamil Nadu, India; kannan@annauniv.edu

Abstract

Background/Objectives: This paper presents a hybrid framework for classification of leukemia gene expression data. The framework used in this work consists of three subsystems, namely, class based dimension reduction subsystem, feature selection subsystem and classification subsystem. Methods/Statistical Analysis: This work uses class based dimension reduction techniques by employing PCA and Canonical Correlation Analysis (CCA) to the leukemia gene expression dataset. Acute Lymphoblastic Leukemia (ALL) class is subjected to Principal Component Analysis (PCA) and Acute Myeloid Leukemia (AML) class to CCA thus obtaining dimension reduced data. The feature selection subsystem uses Genetic Algorithm (GA) to select an optimal subset of informative genes. The classification subsystem utilizes these informative genes to train the NN and the classifier is obtained. Findings: The performance of the hybrid framework, GA-PCA and CCA, is analyzed and compared with that of single dimension reduction techniques, namely, GA-PCA and GA-CCA. The experimental results show that the proposed framework achieved accuracy of 88.23%. The sensitivity of the system is 85% and specificity of the system is 92.85%. This aids in determining the informative genes that are relevant to leukemia gene expression data. Applications/Improvements: The classification accuracy of GA-PCA and CCA has shown improvement when compared to that of single dimension reduction technique. Hence, combining more than one method yields higher classification accuracy and aids in identification of new classes.

Keywords: Cancer Classification, Canonical Correlation Analysis (CCA), Dimensionality Reduction (DR), Genetic Algorithm (GA), Neural Network (NN), Principal Component Analysis (PCA)

1. Introduction

In this paper, a framework is presented for classification of leukemia microarray gene expression data. The framework utilizes two-dimension reduction techniques, PCA and CCA with GA and neural network. The gene expression dataset considered as input is reduced by its dimension using PCA and CCA. The PCA reduces ALL class and CCA reduces AML class. The reduced feature set from both the techniques is combined to form initial population for genetic algorithm. The dimension reduced feature is then optimized using Genetic algorithm to find the best set of features for classification using feed Forward Back-Propagation Neural Network (FFBNN). The FFBNN performs the classification with optimized set of best feature and gives the classification performance.

*Author for correspondence
Hybrid Dimension Reduction Techniques with Genetic Algorithm and Neural Network for Classifying Leukemia Gene Expression Data

Huerta et al. have proposed a hybrid approach to select genes for cancer classification from microarray data. Their approach combined genetic algorithm with Fisher’s Linear Discriminant Analysis (LDA) for classification. Their algorithm used LDA classifier in its fitness function, as well as LDA’s discriminant coefficients in its crossover and mutation operators. In addition, the fitness function is used to either minimize the number of selected genes or to maximize the prediction accuracy. They tested their algorithm with seven public datasets namely leukemia, colon cancer, lung cancer, prostate cancer, Central Nervous System embryonal tumor (CNS), ovarian cancer and lymphoma (DLBCL) and they have achieved an accuracy of 100%, 91.9%, 99.3%, 96%, 86.6%, 100% and 100% respectively. The proposed algorithm attained high prediction accuracy with 3 genes for leukemia, 4 genes for colon cancer, 10 genes for lung cancer, 53 genes for prostate, 34 genes for CNS, 18 genes for ovarian and 4 genes for DLBCL respectively.

Yao and Tian have proposed a supervised dimension reduction and feature extraction method to extract hyper spectral image for different remote sensing applications. Genetic-algorithm-based selective principal component analysis (GA-SPCA) is used. The genetic algorithm is used to select subset of the original image bands to reduce the data dimension. In the selection phase, the original image band number is first reduced. Then, the features are extracted from the resulting eigen image, thereby reducing the feature space with several principal component bands. Subsequent image processing is performed with improved accuracy. The performance analysis was carried out using three data of image bands. By removing image bands that do not contribute to information extraction for a specific application, the resulting correlation was improved. The GA-SPCA method provided a standard approach for hyper spectral image dimension reduction and feature extraction. It also provided useful information for imaging sensor development. Results demonstrated using GA-SPCA clearly shows that the reduction ratios in the band selection stage were 36.7%, 51.7%, and 71.7% respectively.

Liu proposed a hybrid method for selection of genes based on microarray profiles. This approach uses wavelet feature extraction method for selecting a set of new basis for transforming the data. The purpose of selecting the set of wavelet basis is to detect the features contained in microarray data. Wavelet decomposition is performed on a microarray vector to extract the coefficients, which signifies the changing information and reduce vector dimensionality. To identify the significant genes, wavelet details are reconstructed based on detail coefficients. The wavelet analysis for gene expression data, represents a sum of wavelets at different time shifts and scales using discrete wavelet analysis (DWT). The DWT is capable of extracting the local features. SVM classifiers are designed to classify the wavelet features into different diagnostic classes. Experiments are carried out on four independent datasets namely Leukemia, mixed-lineage leukemia (MLL), Prostate cancer, Subtypes of ALL with six diagnostic classes. The performance was analyzed using twofold cross-validation that achieves 97.06% accuracy for Leukemia dataset.

Ghorai et al. proposed a binary non-parallel plane proximal classifier (NPPC) ensemble for gene microarray expression analysis. A new proximity profile based combination of multiple experts (CME) has been proposed for NPPC ensemble. NPPC ensemble technique shows good discriminating power in gene expression analysis. NPPC ensemble intends to search the discrimination ability of NPPC in different subspaces with the best gene subset in that higher dimension. Genetic algorithm (GA) is used to select the best gene set of the NPPC by maximizing the k-fold cross validation accuracy. The performance of GA-based selection of NPPC method is analyzed with seven datasets namely ALL-AML, colon cancer, lung cancer, breast cancer, lymphoma, liver cancer and prostate cancer. The response of the system has obtained classification accuracy of 94.52%, 82.77%, 96.38%, 81.21%, 86.8%, 96.77% and 90.16% respectively. NPPC ensemble method attained high classification accuracy for test samples in a Computer-Aided Diagnosis (CAD) framework.

Das et al. presented a framework for selection of gene set using two approaches, namely, statistical and information based approaches used for ranking genes. The subset of genes is selected using gene ranking. Initially the dataset is pre-processed using principal component analysis. Statistical approach is based on Euclidian Distance (ED) and Pearson Correlation (PC). Mutual information based approach is based on Information Gain (IG) and Dynamic Relevance (DR). The mutual information approach generates the subset using a threshold. Different methods generate a number of subsets according to the threshold value. The subset generated is then applied on the classifier and the performance is analyzed. The gene selection methods used is experimented on four publicly available data sets such as, breast cancer, leukemia, hepatitis and dermatology. The resultant subset of genes is fed to two classifiers namely Naive-Bayes and Support
Vector Machine (SVM). Results demonstrated show the improvement in the performance of the classifiers.

Compared to the works discussed in the literature, the proposed work is different in the following ways: The framework utilizes a class based dimension reduction subsystem by employing PCA and CCA techniques based on its class variants. The system framework is different from the existing ones in the way of hybridization. Rather than using single dimension reduction technique this system makes use of two-dimension reduction techniques. ALL class is subjected to PCA and AML class is subjected to CCA thus obtaining dimension reduced data. Further, the genetic algorithm is used to select optimal subset of relevant genes for classification. In GA fitness function for the feature selection is computed from the error rate obtained from neural network. The classification subsystem uses the optimal subset of relevant genes as input to train the feed forward neural network and the trained neural network is used for classifying the leukemia gene expression data. The search for optimum subset of relevant genes improves the classification accuracy.

2. System Framework

The proposed hybrid framework for classification of leukemia gene expression data is illustrated in Figure 1. The framework comprises of three subsystems, namely, class based dimension reduction subsystem, feature selection subsystem and classification subsystem.

![Figure 1. Framework showing the hybrid dimension reduction and classification.](image)

In class based dimension reduction subsystem dimension reduction subsystem comprises of both dimension reduction subsystem using PCA and dimension reduction subsystem using CCA. The feature selection subsystem uses GA and classification subsystem uses FFBN. In feature selection subsystem, an optimal subset is selected by evaluating fitness for each chromosome. The subset of features with maximum fitness is selected to train the feed forward back-propagation neural network and classify the selected genes as acute lymphoblastic leukemia and acute myeloid leukemia. The performance of the trained classifier is evaluated using test data.

2.1 Class based Dimension Reduction Subsystem

The class based dimension reduction subsystem reduces the high dimensional gene expression data, which has 7129 features (genes) to a lower dimension data by employing PCA and CCA techniques based on its class variants. It comprises of two subsystems, namely, dimension reduction subsystem using PCA and dimension reduction subsystem using CCA. Dimension reduction subsystem using PCA takes ALL as input and dimension reduction subsystem using CCA takes AML class as input. Let the initial leukemia gene expression data of 7129×72 be \( S \) represented as gene expression data matrix. The leukemia gene expression data comprises of ALL and AML data. Thus the matrix \( S \) comprises of two matrices namely, matrix \( X \) and matrix \( Y \).

\[
S = \{ D_n \}_{n \times b} \\
\text{and} \\
X = \{ x \}_{n \times k} \\
Y = \{ y \}_{n \times k}
\]

Eq (1)

\( X \) holds the acute lymphoblastic leukemia class and \( Y \) holds acute myeloid leukemia class. The procedure of the dimension reduction subsystem using PCA and CCA is detailed below.
2.1.1 Dimension Reduction Subsystem using Principal Component Analysis

PCA is a multivariate statistical technique used for dimensionality reduction\(^5\). In this work, PCA re-builds the dimension by integrating the original gene set of genes into a new set of integrated gene set\(^6\). The PCA dimension reduction subsystem considers all 7129 gene expression levels taken over 47 ALL samples. The procedure of dimension reduction subsystem using PCA is summarized as follows.

Let there be \(n\) samples with \(k\) genes constituting an \(n \times k\) data matrix \(X\). The data matrix \(X\) represents the matrix of ALL data as shown in (2).

\[
X = \begin{bmatrix}
    x_1 & x_2 & \cdots & x_{1k} \\
    x_2 & x_2 & \cdots & x_{2k} \\
    \vdots & \vdots & \ddots & \vdots \\
    x_{n1} & x_{n2} & \cdots & x_{nk}
\end{bmatrix}
\quad Eq (2)
\]

\[
X_j^* = \frac{x_j - \bar{x}_j}{\sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (x_j - \bar{x}_j)^2}}
\quad Eq (3)
\]

where; \(\bar{x}_j = \frac{1}{n} \sum_{i=1}^{n} x_j\)

Equation (3) computes the mean vector for the dimension of original gene set. The correlation matrix \(R = (r_{ij})_{2k\times2k}\) is then calculated for \(n \times k\) gene data matrix. The eigen values \(\lambda\) is calculated using Equation given in (4).

\[
|\lambda E - R| = 0
\quad Eq (4)
\]

In Eq (4) \(E\) represents eigen vectors and \(\lambda\) represents eigen values. The eigen values can be obtained from the corresponding eigen vectors. From the eigen vectors the number of principal components is determined. The contribution ratio is computed is using (5).

\[
\alpha_i = \frac{\lambda_i}{\sum_{i=1}^{k} \lambda_i}
\quad Eq (5)
\]

If the ratio value is high, then the ability to reflect the information present in the original indexes is strong. The Eigen values are arranged in the descending order, and calculate their accumulated contribution ratio is calculated. When the accumulated contribution ratio reaches a certain value, then \(m\) variables can reflect sufficient information of the original \(k\) genes, \(\lambda_i\). The parameter \(\lambda_i\) is taken according to the requirement of the system. The number of principal components is obtained by the inequality relation defined in eqn. (6), where \(\sigma\) is the scoring coefficient.

\[
\sum_{i=1}^{m} \lambda_i \geq \sigma \sum_{i=1}^{k} \lambda_i
\quad Eq (6)
\]

Thus, the new \(m\) variables can take the place of the original \(k\) genes.

\[
Z_1 = a_1 x_1 + a_2 x_2 + \cdots + a_k x_k
\]

\[
Z_2 = a_1 x_1 + a_2 x_2 + \cdots + a_k x_k
\]

\[
\vdots
\]

\[
Z_m = a_{1m} x_1 + a_{2m} x_2 + \cdots + a_{km} x_k
\quad Eq (7)
\]

where, \(1 \leq m \leq k\), \(Z_1\) is the first principal component; \(Z_2\) is second principal component and \(Z_m\) is the \(m^{th}\) principal component. As a result, 7129×47 input genes are dimension reduced into 200×47 output genes comprising of 200 expressions.

2.1.2 Dimension Reduction Subsystem using Canonical Correlation Analysis

This works uses the conventional CCA presented by Sun et al.\(^8,9\) for performing the analysis of gene expression. The CCA dimension reduction subsystem considers 7129 gene expression levels taken over 25 samples.

There are two different representations of the same set of genes used. CCA computes two projection vectors, \(\omega_x \in \mathbb{R}^d\) and \(\omega_y \in \mathbb{R}^k\), such that the correlation coefficient \(\rho\) given by eqn. (8) is maximized.

\[
\rho = \frac{\omega_x^T X^T \omega_y}{\sqrt{(\omega_x^T X^T \omega_x)(\omega_y^T X^T \omega_y)}}
\quad Eq (8)
\]

Since, \(\rho\) is invariant to the scaling of \(\omega_x\) and \(\omega_y\), CCA can be formulated equivalently as

\[
\sum_{i=1}^{m} \lambda_i \geq \sigma \sum_{i=1}^{k} \lambda_i
\]
\[
\max_{\omega_i, \omega_j} \omega_i^T X \omega_j \quad \text{Eq (9)}
\]

where; \(\omega_i^T X \omega_j = 1\) and \(\omega_j^T X \omega_j = 1\).

\(\omega_i\) and \(\omega_j\) are projection vectors.

As a result, 7129×25 input genes are dimension reduced into 200×25 output genes comprising 200 expressions. After obtaining the individual results of the dimension reduction subsystem, union operation is performed from which 20 expression levels were obtained common in both PCA and CCA dimension reduction subsystem. The reason behind the union operation performed on the dimension reduced data results is eliminating redundant genes. Thus, the output of class based dimension reduction subsystem obtains 380 features which serves as an input to feature selection subsystem.

2.2 Feature Selection Subsystem

The feature selection subsystem uses the dimension reduced 380 features from class based dimension reduction subsystem as an input. Genetic algorithm is used for selecting an optimal subset of informative genes from the dimension reduced leukemia gene expression data set. A subset of genes is selected for population initialization, fitness evaluation, crossover and mutation. The obtained feature subsets (chromosomes) are used to train the FFBNN and thereby the fitness of the chromosome is evaluated. The FFBNN is trained with the optimal feature subset. The steps involved in feature selection subsystem are detailed below.

The steps involved in the feature selection subsystem are detailed below:

**Algorithm**

**Input**

380 feature set obtained from class based dimension reduction subsystem

**Process**

**Step 1:** The output of class based dimension reduced data is the initial population to genetic algorithm.

Each chromosome is represented as a string of binary digits. The presence of ‘1’ and ‘0’ indicates the presence and absence of the gene in that corresponding feature subset.

**Step 2:** The fitness is computed for subset of gene data that are present in the population pool using the fitness function defined in Equation (10).

\[
f = \frac{C}{Y} \times 100 \quad \text{Eq (10)}
\]

The fitness \(f\) is calculated using the above mathematical model in Eq (10) where, \(Y\) is the desired output and \(C\) is the actual output. The output of the FFBNN is obtained by computing the output value of the output neuron as defined in Equation (11).

\[
C = \sum_{j=1}^{M} \frac{w_j^O}{1 + \exp(-\sum_{i=1}^{N} F_i w_j^I)} \quad \text{Eq (11)}
\]

In Equation (11), \(F_i\) is the input selected subset of gene value, \(w_j^I\) is the weight assigned between input and hidden layer, \(w_j^O\) is the weight assigned between hidden and output layer and \(M\) is the number of hidden neurons. The hidden layer and output layer uses the non-linear activation function.

**Step 3:** The chromosomes with optimal subset of features are ranked based on their fitness values. The fitness values are computed using the fitness function \(f\) defined in Equation (7.10). The chromosomes with maximum fitness are selected for crossover and mutation.

**Step 4:** Crossover is performed over subset of features in order to produce two new offspring (chromosome). This work uses a single point crossover operation with a crossover rate \((C_c)\) of 0.8.

**Step 5:** The mutation is performed over the child chromosomes by rearranging the feature set which is modified and the new solution is obtained. This work uses bit swap mutation.
Step 6: Repeat the process from step 2 to step 5, until new generation chromosome produces the error e<0.1.

Output
Subset of informative genes
The genetic algorithm initializes the population set with 100 solutions randomly with each solution comprising of 50 features. The fitness function is computed using feed forward back propagation neural network. The fitness for each solution is calculated and ranked from maximum to minimum (descending order) with respect to the fitness value. The first 50 solutions are selected and genetic operation, crossover is performed over the 50 solutions and 100 offsprings are obtained. The fitness of the 100 offsprings obtained by performing crossover is computed and ranked from maximum to minimum with respect to the fitness value. The first fifty offspring’s are selected and mutation is performed on the offsprings. Through repeated experimentation it is inferred that the optimal subset with 50 features is obtained at the 20th iteration. The output of feature selection subsystem obtained is the optimal subset with 50 features which is further used by classification subsystem.

2.3 Classification Subsystem
The classification subsystem uses feed forward back-propagation neural network (FFBNN). FFBNN consists of three layers with input layer, hidden layer and output layer as illustrated in Figure 2. The FFBNN structure uses back-propagation algorithm for training and testing. The network structure comprises of three layers, namely, input layer, hidden layer and output layer. The input layer takes fifty features as input $g_1, g_2, ..., g_{50}$ obtained from feature selection subsystem to train the FFBNN as shown in Figure 2. As a result, the input layer has fifty neurons each neuron corresponding to each feature. The second layer is the hidden layer with hundred neurons. The number of neurons in the hidden layer is computed and approximated using $2n+1$ (Morshed and Kaluarachchi 1998) where n is the number of input features. The third layer is the output layer with one output neuron. The neuron in the output layer corresponds to the class label. The hidden layer and the output layer uses sigmoidal and purelin activation functions respectively. The FFBNN is trained using to train each subset of solutions obtained.

Figure 2. Structure of feed forward back-propagation neural network.

The optimal subset of leukemia gene expression data obtained is used for both training and testing. In training phase, the FFBNN is trained using the optimal feature subset obtained from genetic algorithm. In testing phase, the trained neural network is used to classify the selected genes in test data as either ALL or AML.

3. Experimental Results and Discussion
The proposed hybrid framework system for the dimensionality reduction of gene expression data and classification is implemented using Matlab (version 2013a) and tested with leukemia gene expression data obtained from broad institute cancer program website. The leukemia gene expression dataset taken into consideration contains 7129 genes and 72 samples belonging to two leukemia classes, namely, acute lymphoblastic leukemia and acute myeloid leukemia. The 72 samples comprises of 47 ALL samples and 25 AML samples.

The microarray gene expression data considered as input is first subjected to class based dimension reduction subsystems. In class based dimension reduction subsystem, two-dimension reduction techniques are employed based on the class variants. First, acute lymphoblastic leukemia samples are subjected to dimension reduction techniques with genetic algorithm and neural network for classifying leukemia gene expression data.
subsystem using PCA and acute myeloid leukemia samples are subjected to dimension reduction subsystem using CCA. The output of class based dimension reduction subsystem obtains 380 genes which is the initial population to genetic algorithm. In genetic algorithm, fitness is computed for each optimal subset of feature set. The feature set with maximum fitness is given as input to the neural network. An optimal subset of 50 genes (features) were obtained from genetic algorithm. For training 38 samples were used and for testing 34 samples were used.

The performance of the hybrid technique is analyzed and compared against the performance of dimensionality reduction techniques like GA-PCA and GA-CCA. The performance of the designed hybrid framework is evaluated using the statistical performance measures namely accuracy, sensitivity, specificity, Positive Predictive Value (PPV), Negative predictive value (NPV), False Positive Rate (FPR) and False Discovery Rate (FDR) for the different dimensionality reduction techniques. The mathematical models for the performance measures are defined in Eq (11-17). The performance measure of the hybrid framework with different techniques is presented in Table 1.

![Figure 3. Performance of GA-PCA&CCA in terms of accuracy.](image)

As shown in Figure 3 the comparison graph result h Hybrid Dimension are shown that our proposed hybrid framework that combines GA-PCA and CCA achieves better performance rate when compared to GA-PCA and GA-CCA respectively.

### 4. Conclusion

The proposed system presents a hybrid dimension reduction technique with genetic algorithm for selecting subset of informative genes, thereby reducing the dimension of gene expression data. The selected genes can be used to classify ALL and AML genes from leukemia gene expression data. This work focuses in feature selection subsystem that selects informative genes thereby identifying and eliminating irrelevant feature before classification. It has been observed that this process of dimension reduction in gene expression data has improved the classification accuracy. There are many interesting aspects for future work. Further, this work can be extended to classification on multi-class leukemia problems with different other gene expression datasets. In addition, there are many other optimization algorithms other than genetic algorithm, namely, particle swarm optimization algorithm, ant bee colony optimization algorithm, cat swarm optimization algorithm, firefly algorithm and still more which can be used for selection.

<table>
<thead>
<tr>
<th>Techniques</th>
<th>Accuracy</th>
<th>Sensitivity</th>
<th>Specificity</th>
<th>PPV</th>
<th>NPV</th>
<th>FPR</th>
<th>FDR</th>
</tr>
</thead>
<tbody>
<tr>
<td>GA-PCA</td>
<td>85.29%</td>
<td>80%</td>
<td>92.85%</td>
<td>94.11%</td>
<td>76.47%</td>
<td>7.1%</td>
<td>5.8%</td>
</tr>
<tr>
<td>GA-CCA</td>
<td>85.29%</td>
<td>85%</td>
<td>85.71%</td>
<td>89.47%</td>
<td>80%</td>
<td>14.28%</td>
<td>10.52%</td>
</tr>
<tr>
<td>GA-PCA &amp;CCA</td>
<td>88.23%</td>
<td>85%</td>
<td>92.85%</td>
<td>94.11%</td>
<td>81.25%</td>
<td>7.1%</td>
<td>5.4%</td>
</tr>
</tbody>
</table>
of informative genes. Further, the temporal behaviour of genes can also be used for improving the diagnosis.

5. References