IAES International Journal of Artificial Intelligence (IJ-AI)
Vol. 1, No. 2, June 2012, pp. 91~102
ISSN: 2252-8938

A Projection Algorithm to Detect Cancer Using Microarray

Nazario D. Ramirez-Beltran*, Joan M. Castro, Harry Rodriguez*****
* Department of Industrial Engineering, University of Puerto Rico
** Department of Electrical and Computer Engineering, University of Puerto Rico
***Cordis LLC a Johnson & Johnson Company, San German, Puerto Rico

Article Info

Article history:

Received May 21, 2012
Revised June 15, 2012
Accepted June 20, 2012

Keyword:

microarray,
discriminant analysis,
neural networks, and
logistic regression

ABSTRACT

The projection algorithm to classify tissues with a large number of genes and a small number of microarrays is proposed. The algorithm is based on the angle formed by two vectors in the n-dimensional space, and takes advantages of the geometrical projection principle. The properties of known tissues can be used to train the algorithm and distinguish between the cancer and normal gene expressions. The gene's percentiles from an independent data set can be used to create a third vector, which is projected into the previously trained vectors to classify the third vector in one of the two populations, cancer or normal population. The proposed algorithm was implemented to detect cervical cancer in a microarray data set, which contains 8 normal and 25 cancerous tissues, which were randomly selected one thousandof times using a combinatory strategy. The algorithm was compared with three existing algorithms that have been used to solve the microarray classification problem: Fisher discriminate function, logistic regression, and artificial neural networks. Results show that the proposed algorithm outperformed the selected algorithms.

Corresponding Author:

Nazario D. Ramirez-Beltran,
Departement of Industrial Engineering,
University of Puerto Rico,
P.O. Box 9000, Mayaguez, Puerto Rico, 00681-9000.
Email:nazario.ramirez@upr.edu

1. INTRODUCTION

Cancer is one of the most devastating and impacting illnesses of the human being. Although, large efforts have been conducted throughout several generations the solution has not been accomplished yet. This research effort will contribute in the developing of an algorithm for cancer diagnosis through exploring microarray analysis. Deoxyribonucleic acid (DNA) is a nucleic acid that contains the genetic instructions used in the development and functioning of all known living organisms with the exception of some viruses. A DNA microarray is a genomics tool used to study many genes in an organism, at the same time in only one experiment [1]. With this technique it is possible to obtain the genetic catalog of a cell through the hybridization process, which consists of extracting all messenger ribonucleic acid (mRNA) molecules for every expressed gene in a given cell and build complementary DNA (cDNA) molecules. A DNA microarray is a plate made of thousands of spots, where each spot contains copies of a unique DNA sequence that corresponds to a single known gene.

During a microarray experiment the built cDNA molecules are marked with a color molecule and then deposited in a microarray plate where each cDNA molecule couples with its corresponding DNA sequence in the spots. The plate is scanned with a laser light to identify the genes within the cell in question, and the identified genes are the expressed genes of a given cell. Thus, if the cDNA molecules from two cells are marked with different colors it is possible to identify the difference in a gene expression of the two cells.

This technique can be used to characterize the cells from cancerous and normal tissues and then be able to diagnose cancer in patients.

The microarray technology has provided an opportunity to develop methods to treat cancer and other diseases. The availability of microarray data has motivated to derive biomarkers for disease diagnosis and prognosis, and the identification of efficient treatments. According to Pang [2] there are three statistical problems in cancer genomics research: the identification of subclasses within a particular tumor type; the classification of patients into known class; and the selection of biomarkers, i.e., genes that characterize a particular tumor. In this research we will focus in classifying human tumors based on microarray information.

The gene expression technology has motivated to develop some statistical methods to solve the classification problem characterized by having a very small sample size. Lee [3] conducted an exhaustive literature review and compare 21 classification algorithms with 7 different data sets and found that the best method to solve the underlying classification problem is the support vector machine. Pang [2] found that the support vector machine is one of the best and performed similar to diagonal linear discriminant analysis (DLDA), diagonal quadratic discriminant analysis (DQDA), and the k nearest neighbor (kNN). The selected classification methods to be compared with our proposed method should exhibit good performance and must be easy to be implemented. Thus, the selected methods are: Fisher discriminate function, logistic regression, and artificial neural networks.

In this research a new technique for analyzing microarrays is introduced, the proposed approach is simple and robust to conduct microarray analyses. The entertained algorithm includes some percentiles to express an approximation of the cumulative probability distribution of the gene expression. The percentiles of a gene from a cancer tissue were also used to create a vector that represents the expression of that particular gene, and similarly the corresponding expression of a normal gene was used to create a second vector. The cancer (or normal) gene expressions will exhibit similar behavior if the angle between two vectors is very small. On the other hand, the cancer (or normal) gene expressions will exhibit different behavior if the angle between those vectors approaches to 90 degrees. Thus, the genes that show the largest projection angles are the genes with the largest expression and were used to perform classification of a tissue from an independent data set.

The main purpose of this study is to use mathematical tools to analyze microarrays for developing a cancer detection algorithm. The specific objectives of this study are: to develop the projection algorithm to identify the genes with the maximum expression; to use the projection algorithm to classify a new tissue as either a cancer or normal tissue; and finally, to compare the performance of the projection algorithm with some of existing methods for cancer detection. The second section of this study shows a detailed description of proposed projection algorithm. The third section describes some of the most prominent classification methods used to detect cancer on microarrays. The fourth section presents some statistics to measure the accuracy of classifications methods. This section also presents the evaluations of the Projection Algorithm and compares the performances of this algorithm with some existing algorithms. The last section presents some conclusions.

2. PROPOSED METHOD

2.1 Projection algorithm

An efficient algorithm is proposed to distinguish between cancer and normal tissues. The algorithm starts by identifying the genes that exhibits the maximum expression, and then uses the genes probability distribution to compute angles between genes and finally classify tissues on healthy and nohealthy. The data set used in this research was originated from microarray technology studies and contains 10,692 genes with 33 tissues, 25 out of 33 are cancerous; and the remaining 8 tissues are normal [4]. The data set was used without applying any mathematical transformation. The microarray data were divided in two parts: data for training, and for validation. The training data were used for designing the centroids and projection vectors, and the validation data were used for testing performance of the algorithm.

The training process requires first identifying the genes with the largest expression and second assigned the selected genes to a specific tissue. The identified genes with the largest expression will be used to create the cancer (or normal) population. The data set for training is used to calculate the central tendency of each population and will be represented by two vectors: the cancer and normal vectors. The application of the projection algorithm consists of selecting a tissue from the validation data to be classified as a cancer or normal tissue. The tissue from validation data will be used to create a third vector, which will be projected onto the cancer and normal vectors. The magnitude of the projection angles and the probability distribution of the gene expressions will be used to classify the tissue. The implementation of the projection algorithm requires performing of three steps:

2.1.1 Identifying the most expressive genes

The original microarray information is organized in two matrices:$T_{cancer}$and T_{normal}; these matrices are of the size $N \times n_1$, and $N \times n_2$, respectively; where N is the number of genes that are included in each tissue and n_1 and n_2 are the number of cancer and normal tissues for training purposes.

Five percentiles for each gene would be a representation of the accumulative probability distribution of each gene and will be computed for both normal and cancer genes, and will be organized in the following vectors:

$$\mathbf{c}_i = [c_{5,i} \quad c_{25,i} \quad c_{50,i} \quad c_{75,i} \quad c_{95,i}] \qquad (1) \qquad \mathbf{h}_i = [h_{5,i} \quad h_{25,i} \quad h_{50,i} \quad h_{75,i} \quad h_{95,i}] \qquad (2)$$

Where $\mathbf{c}_i$ and $\mathbf{h}_i$ are the cancer and normal percentiles vectors for the i^{th} gene, respectively. The $c_{r,i}$ and$h_{r,i}$ are the r^{th} percentile for the i^{th} gene of the cancer and normal tissues, respectively. The percentiles vectors are collected to create the percentiles matrices (**C** and **H**) which will be the basis for calculation the projection angles. The size of the percentile matrices are both $N \times 5$, and can be represented by:

$$\mathbf{C} = \begin{bmatrix} c_{5,1} & c_{25,1} & \cdots & c_{95,1} \\ c_{5,2} & c_{25,2} & \cdots & c_{95,2} \\ \vdots & \vdots & \ddots & \vdots \\ c_{5,N} & c_{25,N} & \cdots & c_{95,N} \end{bmatrix} = [\mathbf{c}_i] \qquad (3) \qquad \mathbf{H} = \begin{bmatrix} h_{5,1} & h_{25,1} & \cdots & h_{95,1} \\ h_{5,2} & h_{25,2} & \cdots & h_{95,2} \\ \vdots & \vdots & \ddots & \vdots \\ h_{5,N} & h_{25,N} & \cdots & h_{95,N} \end{bmatrix} = [\mathbf{h}_i] \qquad (4)$$

The projection angle of the i^{th} gene is the angle between the normal vector, $\mathbf{h_i}$**,** and the cancer vector, $\mathbf{c_i}$. The definition of the inner product is used to calculate the angle between vectors in the n-dimensional space and can be expressed as follows [5] (Strang 1976):

$$\theta_i = cos^{-1}\left(\frac{\mathbf{h_i} \cdot \mathbf{c}_i}{\|\mathbf{h_i}\|\|\mathbf{c_i}\|}\right), \quad i = 1,2,\dots,N \qquad (5)$$

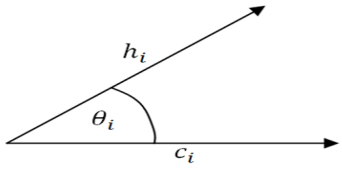

Figure 1. Projection angle

where$\mathbf{h_i} \cdot \mathbf{c_i}$isthe inner product between the $\mathbf{h_i}$and $\mathbf{c}_i$vectors**;** $\|\mathbf{h_i}\|$ is the module (the vector length) of the vector $\mathbf{h_i}$and $\|\mathbf{c}_i\|$ is the module of the vector $\mathbf{c_i}$**.**The geometrical representation of the angle can be given in Figure 1.

The genes with the largest angles are selected as the ones that exhibited the largest expression. The genes are ranked according with the projection angle and the angle that correspond the 90 percentile is selected as the reference angle, ϕ. All the genes that exhibit a projection angle larger than the reference angle are declared as the genes with the largest expression; i.e., the i^{th} gene with $\theta_i > \phi$ is defined as the gene with a large expression.

2.1.2 Develop centroids for normal and cancer populations

Now the most expressive genes will be associated with cervical tissues to develop a classification scheme. Arbitrarily 10 out of the most expressive genes are selected in each training tissue to initialize the algorithm. Thus the initial centroid for cancer and normal tissues would be $\mathbf{B}_k$and $\mathbf{G}_k$, respectively, with dimensions of $k \times m_1$ and $k \times m_2$; where m_1 and m_2 are the number of cancer and normal tissues used for training, respectively and initially *k=10*.

$$\mathbf{B}_k = \begin{bmatrix} b_{1,1} & b_{1,2} & \cdots & b_{1,m_1} \\ b_{2,1} & b_{2,2} & \cdots & b_{2,m_1} \\ \vdots & \vdots & \ddots & \vdots \\ b_{k,1} & b_{k,2} & \cdots & b_{k,m_1} \end{bmatrix} \qquad (6) \qquad \mathbf{G}_k = \begin{bmatrix} g_{1,1} & g_{1,2} & \cdots & g_{1,m_2} \\ g_{2,1} & g_{2,2} & \cdots & g_{2,m_2} \\ \vdots & \vdots & \ddots & \vdots \\ g_{k,1} & g_{k,2} & \cdots & g_{k,m_2} \end{bmatrix} \qquad (7)$$

The percentiles for each tissue are estimated by computing the percentile for each column of matrices (6) and (7). For instance, the percentiles of the j^{th}tissue of the matrix (6) would be:

$$\Longrightarrow \qquad \mathbf{b}_{\mathrm{j}} = \begin{bmatrix} b_{1,j} \\ b_{2,j} \\ \vdots \\ b_{k,j} \end{bmatrix} \mathbf{u}_{j,k} = [u_{5,j,k} \quad u_{25,j,k} \quad u_{50,j,k} \quad u_{75,j,k} \quad u_{95,j,k}] \quad (8)$$

Thus, the matrices of percentiles for cancer and normal tissues ($\mathbf{U}_{b,k}$ and $\mathbf{V}_{g,k}$) have dimensions of $m_1 \times 5$ and $m_2 \times 5$ and are based on *k* genes and can be written as follows:

$$\mathbf{U}_{b,k} = \begin{bmatrix} u_{5,1,k} & u_{25,1,k} & u_{50,1,k} & u_{75,1,k} & u_{95,1,k} \\ u_{5,2,k} & u_{25,2,k} & u_{50,2,k} & u_{75,2,k} & u_{95,2,k} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ u_{5,m_1,k} & u_{25,m_1,k} & u_{50,m_1,k} & u_{75,m_1,k} & u_{95,m_1,k} \end{bmatrix} \quad (9)$$

$$\mathbf{V}_{g,k} = \begin{bmatrix} v_{5,1,k} & v_{25,1,k} & v_{50,1,k} & v_{75,1,k} & v_{95,1,k} \\ v_{5,2,k} & v_{25,2,k} & v_{50,2,k} & v_{75,2,k} & v_{95,2,k} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ v_{5,m_2,k} & v_{25,m_2,k} & v_{50,m_2,k} & v_{75,m_2,k} & v_{95,m_2,k} \end{bmatrix} \quad (10)$$

where $u_{p,j,k}$ is the p^{th} percentile, of the j^{th} cancer tissue including the k^{th} genes with a significant expression; likewise $v_{p,j,k}$ is the p^{th} percentile, of the j^{th} normal tissue including the k^{th} genes with a significant expression.

The average of each column of the matrices (9) and (10) are computed and used as the central tendency for normal and cancer populations. The central tendency is called the centroid of the corresponding population. Thus, the centroids of the cancer and normal,($\mathbf{PB}_k$ and $\mathbf{PG}_k$) populations with only k genes are given as follows:

$$\mathbf{PB}_k = [\bar{u}_{5,k} \quad \bar{u}_{25,k} \quad \bar{u}_{50,k} \quad \bar{u}_{75,k} \quad \bar{u}_{95,k}] \quad (11)$$

$$\mathbf{PG}_k = [\bar{v}_{5,k} \quad \bar{v}_{25,k} \quad \bar{v}_{50,k} \quad \bar{v}_{75,k} \quad \bar{v}_{95,k}] \quad (12)$$

where the averages for cancer and normal percentiles are computed as follows:$\bar{u}_{p,k} = \frac{1}{m_1}\sum_{j=1}^{m_1} u_{p,j,k}$ and $\bar{v}_{p,k} = \frac{1}{m_2}\sum_{j=1}^{m_2} v_{p,j,k}$.

The angle between centroids is computed using the following expression:

$$\delta_k = cos^{-1}\left(\frac{\mathbf{PG}_k \cdot \mathbf{PB}_k}{|\mathbf{PG}_k||\mathbf{PB}_k|}\right) \quad (13)$$

A gene from the list of the most expressive ones is added to the matrices (6) and (7), and the value of *k* is actualized by *k=k+1*. The matrices changes to $\mathbf{B}_{k+1}$and $\mathbf{G}_{k+1}$, and the centroids $\mathbf{PB}_{k+1}$ and $\mathbf{PG}_{k+1}$ are computed again, and the angle δ_{k+1} is also computed. It should be noted that the number of genes may increase; however, the dimensions of the matrices to $\mathbf{B}_{k+1}$ and $\mathbf{G}_{k+1}$ remain the same; i.e., $m_1 \times 5$and $m_2 \times 5$, respectively.

If $\delta_{k+1} > \delta_k$, the selected gene k+1 is added to the list of genes that best contributed to improve the centroids. On the other hand, if the $\delta_{k+1} \leq \delta_k$, the gene is eliminated from the matrices. This process is repeated until the last of the most expressive genes is tested. Thus, the last value of *k* indicates the number of genes that integrate the centroids of the cervical tissues with cancer and normal populations.

2.1.3 Validation

Cervical tissues that were selected for validation will be used to create a validation vector, which will be projected onto the cancer centroid and normal centroid. The tissue under consideration will be classified into one of the two classes depending of the projection angle. Thus, the tissue will be classified into the group G that is associated with the minimum projection angle, where $G$is computed as follows:

$$G = \min\{\theta_l\}, \qquad l = 1, 2 \tag{14}$$

Thus, if *l*=1 indicates a cancer group; otherwise will indicate normal group.

The tissues that were saved to perform validation are used to create the projection vector. The most expressive genes that were identified during the training process are selected and the percentiles are computed based on the k genes for each validation tissue. The projection vector is created in a similar fashion as in the training process. The selected genes are given in column vector$\boldsymbol{T_j}$ and these values are used to calculate the percentiles expressed by the row vector $\boldsymbol{\tau_j}$

$$\Longrightarrow \quad \mathbf{T}_j = \begin{bmatrix} t_{1,j} \\ t_{2,j} \\ \vdots \\ t_{k,j} \end{bmatrix} \boldsymbol{\tau}_j = [\tau_{5,j} \quad \tau_{25,j} \quad \tau_{50,j} \quad \tau_{75,j} \quad \tau_{95,j}] \tag{15}$$

Thus, the $\mathbf{T}_j$ vector is formed by the most expressive genes and identified during the training process; whereas, the $\boldsymbol{\tau_j}$ vector is composed of the percentiles of the genes with the largest expression of the j^{th} tissue, which was selected to perform validation. The $\boldsymbol{\tau}_j$ vector will be called the projection vector of the j^{th} tissue.

3. RESEARCH METHOD

3.1 Comparison with existing classification methods

The performance of the proposed algorithm was compared with existing classification methods by means of analyzing a single microarray data set. The selected methods to perform this comparison are the following: logistic regression, artificial neural networks, and Fisher discriminate function. Independently of the methods involved in the comparison the following tasks were conducted:

- The projection algorithm was used for all classification methods to identify the genes with the largest expressions.
- Two tissues were randomly selected out of 25 cancer tissues, and two normal tissues out of 8 normal tissues were also randomly selected to perform validation. The 29 tissues were used to create the information to train of a classification method. Once the method was trained the 4 validation tissues were used to measure the classification capabilities of the considered method, and this process was repeated one thousand times to measure the accuracy of classification method.

3.1.1. Logistic regression.

The logistic regression is a nonlinear regression technique that successfully has being used to perform classification of information [6], [7], [8]. The logistic regression can be expressed as follows:

$$y_j = \frac{e^{\rho_j}}{1 + e^{\rho_j}} + \varepsilon_j \tag{16}$$

$$\rho_j = \beta_0 + \sum_{i=1}^{m} \beta_i x_{i,j} \tag{17}$$

where y_j is a binary variable, when $y_j = 1$ indicates that the j^{th} tissue is a cervical cancer tissue otherwise is a normal tissue, e is the base of the nature logarithms, $x_{i,j}$ is the measurement expression of the i^{th} gene with the largest expression from the j^{th} tissue, β_i are the regression coefficients that will be estimated during the training process, and ε_j is a random noise with cero mean and constant variance. As mentioned before the genes with the largest expression were obtained by using the most expressive geansfrom the projection algorithm.

To derive consistent and reliable estimates for the β's the multicolinearity test was implemented [9]. If the muticolinearity problem was present, the genes ($x_{i,j}$) that caused that caused the multicollinearity problem were removed before estimating β's. Thus, once the multicolinearity problem was solved the best estimates were obtained by using the maximum likelihood method. Note that m is the total number of genes with the largest expression and without multicollinearity problem. Finally, the estimates of y_j are values between cero and one; therefore, any value larger than 0.5 is considered as one and cero, otherwise. The estimates of y_j were computed as follows:

$$\hat{y}_j = \frac{e^{\hat{\rho}_j}}{1+e^{\hat{\rho}_j}} \quad (18) \qquad \text{and} \qquad \hat{\rho}_j = \hat{\beta}_0 + \sum_{i=1}^{m} \hat{\beta}_i x_{i,j} \quad (19)$$

As mentioned before the training was conducted with the most expressive genes from 29 tissues that do not exhibit muticolinearity problem. Once the regression model is developed the validation of the regression method is conducted with 4 cervical tissues that were reserved for this purpose. To measure the classification accuracy of the logistic regression method the training and the validation processes were repeated one thousand times and results are presented in the section 4.

3.1.2 Artificial neural networks.

Artificial neural networks (ANNs) have been reported to successfully solve classification problems [10], [11]. ANNs are especially useful when data are classified by using nonlinear boundaries. The successful applications of the ANN depend on determining the appropriate identification of the structure of the ANN. The identification of the structure of an ANN consists of determining the number of layers, the number of neurons in the hidden layers, and the type of the transfer function in each layer. A systematic method was developed to identify the neural network structure for modeling the nonlinear behavior of the variables involved in a given system [12], [13]. Neural networks with more than two layers were discarded because, in such cases, the networks over parameterize the learning process causing a reduction of the prediction capabilities. On the other hand, a neural network consisting on a single layer was not used because it does not have the capability of modeling nonlinear classification boundaries. The identified neural network consists of two layers, which has three neurons in the hidden layer and a single neuron in the output layer. A sigmoidal transfer function was implemented in both the hidden and the output layers.

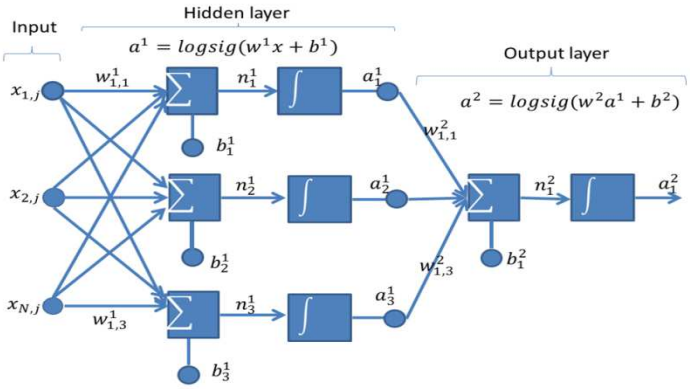

Figure 2.The identified neural network structure.
The hidden layer (left) and the output layer (right) are the main components of the selected neural network. $x_{i,j}$is the input, w's are the weights, b's are the bias, n's are the net input and *a* is the output of the network.

Figure 2 shows the selected structure of the implemented neural network. The input matrix to train the neural network is formed by the most expressive genes. The diagram presented in Figure 2 follows the Hagan's notation [14] and the variables *w's* are the weight matrices, *b's* are the bias vectors, *n's* are the net input of the transfer functions, and *a* is the output of the neural network and is a number between cero and one; thus, if the output is larger than 0.5 the tissue is classified as a cancer tissue; otherwise, the tissue will be classified as a normal tissue.

An optimization algorithm is used to estimate the weights and biases so that the output of the neural network is as close as possible to the target value. In this particular case the target is the number one for a cancer tissue and zero for a normal tissue. This optimization strategy is known as the training process and the neural network will learn the relation between the inputs and the outputs of a dynamic system. Once the neural network is trained, the optimal weights are saved with the purpose of classifying a new tissue. There are various learning algorithms to train neural networks. Some of the most useful are: Perceptron, Hebbian, Widrow-Hoff, and Backpropagation; the latter is a suitable algorithm for this research because it has the capability of training multilayer neural networks with nonlinear classification boundaries [14]. The BackpropagationLevenberg-Marquardt (BLM)algorithm computes automatically the learning rate by introducing a constant that ensures that the Jacobian matrix of weights and biases will be a positive definite

matrix and consequently convergence is accomplished [14], [15]. One of the limitations of this method is the large amount of memory required to perform calculations. However, for the underlying application, the BLM algorithm becomes a suitable technique since only the most expressive genes are the input of the neural network.

A feeforward neural network with the BLM algorithm was used to train the network. The most expressive genes from the 29 tissues were used to create the input matrix to perform the training of the ANN. Once the ANN was trained, the most expressive genes from the 4 validation tissues were used to measure the classification capabilities of the ANN, and this process was repeated one thousand times to measure the accuracy of classification.

3.1.3 The Fisher's discriminant function

The Fisher's discriminant function is a method for classifying information into groups. Fisher's method is based on the idea of transforming a set of multivariate observations into univariate variables, such that the univariate variables are the linear combinations of the multivariate variables and these univariate variables are separated as much as possible between groups and very close within groups. Fisher technique does not assume normality; although, it is assumed equal variances.

Assuming that there are m_1 observations from a multivariate random variables $\mathbf{X_1} = [\mathbf{x_{11}}, \ \mathbf{x_{12}} \ , \cdots, \ \mathbf{x_{1p}}]$ and m_2 observations from $\mathbf{X_2} = [\mathbf{x_{21}}, \ \mathbf{x_{22}} \ , \cdots, \ \mathbf{x_{2p}}]$, where $\mathbf{x_{1j}}$ is a $m_1 \times 1$ vector, from a population 1; and $\mathbf{x_{2j}}$ is a $m_2 \times 1$ vector from population 2, for $j = 1, 2, \dots, p$ and $m_i - 1 > p$ for $i = 1, 2$. In this particular application m_1 and m_2 are the cancer and normal tissues used during the training process and p is the number of genes, and population 1 is associated with cancer tissues whereas population 2 with normal tissues. Since Fisher's method pursuit large differences between groups and small variability within groups, this objective can be accomplished by forming a ratio of two components. The numerator expresses the difference between groups and the denominator variability within groups. Therefore, the Fisher's objective can be accomplished by solving the following optimization problem: find the $\mathbf{a}p \times 1$ vector such that the following ratio is maximized:

$$\max_{\mathbf{a} \neq \mathbf{0}} \frac{(\mathbf{a}'\mathbf{d})^2}{\mathbf{a}'\mathbf{S}\mathbf{a}} \qquad (20)$$

where

$$\mathbf{d} = (\bar{\mathbf{x}}_\mathbf{1} - \bar{\mathbf{x}}_\mathbf{2}) \qquad (21) \qquad \mathbf{S} = \frac{(n_1 - 1)\mathbf{S_1}}{n_1 + n_2} + \frac{(n_2 - 1)\mathbf{S_2}}{n_1 + n_2} \qquad (22)$$

where $\bar{\mathbf{x}}_\mathbf{1}$ and $\bar{\mathbf{x}}_\mathbf{2}$ are average-sample $p \times 1$ vectors from populations 1 and 2, respectively; $\mathbf{S_1}$ and $\mathbf{S_2}$ are the sample variance-covariance $p \times p$ matrices from populations 1 and 2.

The Cauchy-Schwarz [16], [17] inequality can be used to show that the **a**vector that maximizes the ratio of equation (20) is

$$\mathbf{a} = \mathbf{d}'\mathbf{S}^{-1}c \qquad (23)$$

where c is a constant value different from zero. Thus an unknown $p \times 1$ vector $\mathbf{x_0}$ is projected into the **a**vector to classify into one of the underlying populations. Thus, the Fisher's classification rule can be given as follows:

Allocate the $\mathbf{x_0}$ tissue into the cancer population if $y - L \geq 0$; otherwise allocate the tissue into normal population, where

$$y = \mathbf{a}'\mathbf{x_0} \qquad (24) \qquad L = \frac{1}{2}\mathbf{a}(\bar{\mathbf{x}}_\mathbf{1} + \bar{\mathbf{x}}_\mathbf{2}) \qquad (25)$$

To implement the Fisher's Discriminate method requires reducing the number of genes from an extreme large number (available genes) to a number smaller than the number of training tissues. One of the suggested procedures consists of two steps. First, identify and select the most expressive genes using the Projection Algorithm and the second step consist on selecting the genes that do not exhibit the multicolinearity problem. Thus, the test of muticolinearity[9] is performed and the vectors associated with the genes that shows the smallest variance inflation factor are selected, and finally, verifying that the selected genes are less than the number of training tissues (min $\{m_1 - 1, m_2 - 1\}$).

4. RESULTS AND ANALYSIS

4.1 Measuring of accuracy of classification methods

To compare the performance of the classification methods a set of accuracy classification scores is developed in this work. Let $f_1(\mathbf{x})$ and $f_2(\mathbf{x})$ be the probability density function associated with $p \times 1$ vector of random variable $\mathbf{x}$ for the cancer and normal populations, respectively. A cervical tissue $\mathbf{x}$ must be assigned to either a cancer or a normal population. Let Ω be sample space of all possible values of $\mathbf{x}$. Let R_1 be the set of values of $\mathbf{x}$for which the method classified tissue on cancer population, and $R_2 = \Omega - R_1$, the remaining $\mathbf{x}$ values for which the method classify the tissues on normal population.

During the classification process there exist four mutually exclusive events that are associated to the entire probability distribution and can be written as follows:

$$P[(\pi_1 \cap c_1) \cup (\pi_1 \cap c_2) \cup (\pi_2 \cap c_1) \cup (\pi_2 \cap c_2)]$$

$$= P(\pi_1 \cap c_1) + P(\pi_1 \cap c_2) + P(\pi_2 \cap c_1) + P(\pi_2 \cap c_2) = 1 \qquad (26)$$

where

$P(\pi_i \cap c_j)$ is the probability that a tissue belongs to population *i* and has been classified as population*j*, where *i=1, 2* and *j=1, 2,*

π_i is a tissue that belongs to population *i*, and *i=1, 2*

c_i is a tissue that has been classified as population *i*, and *i=1, 2*

It should be noted that in the previous notation the subscript 1 is associated with cancer and 2 with normal population. Thus, the multiplication law can be used to express the joint probability as follows:

$$P(\pi_i \cap c_j) = P(c_j|\pi_i)P(\pi_i) \qquad (27)$$

where

$$P(c_j|\pi_i) = P(x \in R_j|\pi_i) = \int_{R_j} f_i(x)dx \qquad (28)$$

In the classification process there are four possible outcomes two of them correspond to correct decisions and two of them to incorrect decisions, and these four possible outcomes are the following:

A_1 = *A tissue that belongs to cancer population and correctly has been classified as cancer population*
A_2 =*A tissue that belongs to normal population and incorrectly has been classified as cancer population*
A_3 =*A tissue that belongs to cancer population and incorrectly has been classified as normal population*
A_4 =*A tissue that belongs to normal population and correctly has been classified as normal population*

The probability of these events are organized and presented in Table 1. It should be mentioned that these probability statement also correspond to the probability distribution established in eq. (26).

Table 1. Probability distribution of the classification process

		True population	
		π_1	π_2
Classified as:	c_1	$P(A_1) = P(c_1\|\pi_1)P(\pi_1) = \frac{a}{n}$	$P(A_2) = P(c_1\|\pi_2)P(\pi_2) = \frac{b}{n}$
	c_2	$P(A_3) = P(c_2\|\pi_1)P(\pi_1) = \frac{c}{n}$	$P(A_4) = P(c_2\|\pi_2)P(\pi_2) = \frac{d}{n}$

where

$$n = a + b + c + d \qquad (29)$$

a is the number of times that a tissue correctly has been classified on the cancer population and belongs to cancer population.

b is the number of times that a tissue incorrectly has been classifies on the cancer population and belongs to normal population.

c is the number of times that a tissue incorrectly has been classifies on the normal population and belongs to cancer population.

d is the number of times that a tissue correctly has been classifies on the normal population and belongs to normal population.

The hit rate (HR) is the probability that a tissue is correctly classified and is expected that a good classification method will exhibit a HR value close to one; i.e., the desired value of HR is one and the worst case is zero, and it can be estimated as follows:

$$HR = P[(\pi_1 \cap c_1) \cup (\pi_2 \cap c_2)] = P(\pi_1 \cap c_1) + P(\pi_2 \cap c_2)$$

$$= P(c_1|\pi_1)P(\pi_1) + P(c_2|\pi_2)P(\pi_2) = \frac{a+d}{n} \qquad (30)$$

The probability of misclassifying a tissue or the misclassification rate (MR) can be estimated as follows:

$$MR = P[(\pi_1 \cap c_2) \cup (\pi_2 \cap c_1)] = P(\pi_1 \cap c_2) + P(\pi_2 \cap c_1)$$

$$= P(c_2|\pi_1)P(\pi_1) + P(c_1|\pi_2)P(\pi_2) = \frac{b+c}{n} = 1 - HR \qquad (31)$$

The probability of detecting a cancer tissue (POD_1) is the likelihood that a tissue that belongs to cancer population is classified as a cancer tissue. The POD_1 can be estimated as follows:

$$POD_1 = \frac{P(c_1|\pi_1)P(\pi_1)}{P(c_1|\pi_1)P(\pi_1) + P(c_2|\pi_1)P(\pi_1)} = \frac{a}{a+c} \qquad (32)$$

Likewise the probability of detecting a normal tissue can be estimated as follows:

$$POD_2 = \frac{P(c_2|\pi_2)P(\pi_2)}{P(c_1|\pi_2)P(\pi_2) + P(c_2|\pi_2)P(\pi_2)} = \frac{d}{b+d} \qquad (33)$$

The probability of classifying as a cancer tissue given that the tissue belongs to a normal population can be viewed as a false alarm rate (FAR_1) and it can be computed as follows:

$$FAR_1 = \frac{P(c_1|\pi_2)P(\pi_2)}{P(c_1|\pi_1)P(\pi_1) + P(c_1|\pi_2)P(\pi_2)} = \frac{b}{a+b} \qquad (34)$$

Similarly the probability of classifying a normal tissue given that the tissue belongs to a cancer tissue is the false alarm rate for the normal population (FAR_2) and is estimated as follows:

$$FAR_2 = \frac{P(c_2|\pi_1)P(\pi_1)}{P(c_2|\pi_1)P(\pi_1) + P(c_2|\pi_2)P(\pi_2)} = \frac{c}{c+d} \qquad (35)$$

Finally the criterion that can be used to measure the accuracy of classification methods could be the one that maximizes the hit rate or equivalently to minimize the misclassification rate. Other possibility would be to minimize the misclassification index (MI). The MI may be a more robust approach since involves the three classification scores: FAR, POD and HR. The MI is defined as follows:

$$MI_i = \left(\frac{FAR_i - HR_i - POD_i + 2}{3}\right) \; i = 1,2 \qquad (36)$$

It can be noted that the best performance of a classification method will show the following scores: $FAR = 0, HR = 1,$ and $POD = 1,$ and consequently $MI = 0.$ On the other hand the worse performance of a classification method will reveal the following scores: $FAR = 1, HR = 0,$ and $POD = 0,$ which implies $MI = 1.$ Thus, the MI is a number that range between zero and one, and zero is associated the best performance of the classification method, whereas a one is associated to the worse performance.

4.2 Evaluation of classification methods

The performance of the projection algorithm and the comparison with three existing classifications methods are compared through the use of the accuracy scores derived in the previous section. To perform the comparison among the methods a single microarray data set was used. As it was mentioned before the studied microarray data includes 8 normal cervical tissues and 25 cancer cervical tissues. The DNA microarrays contain 10,692 genes and these data were generated based on 25-grossly dissected primary tumors from cervical cancer patients and 8 normal cervical expression profiles from hysterectomy. These data were obtained at the Princes of Wales Hospital of the department of Obstetrician and Genecology at the Chinese University of Hong Kong [4].

Table 2. Statistics for validation

Classification method	Cancer tissues			Normal tissues			Overall performance	
	POD_1	FAR_1	MI_1	POD_2	FAR_2	MI_2	HR	MR
Projection Algorithm	0.95	0.15	0.11	0.82	0.06	0.12	0.89	0.11
Neural Networks	0.83	0.19	0.18	0.81	0.17	0.18	0.82	0.18
Fisher Discriminant	0.79	0.22	0.22	0.77	0.21	0.22	0.78	0.22
Logistic Regression	0.81	0.27	0.23	0.70	0.21	0.25	0.75	0.25

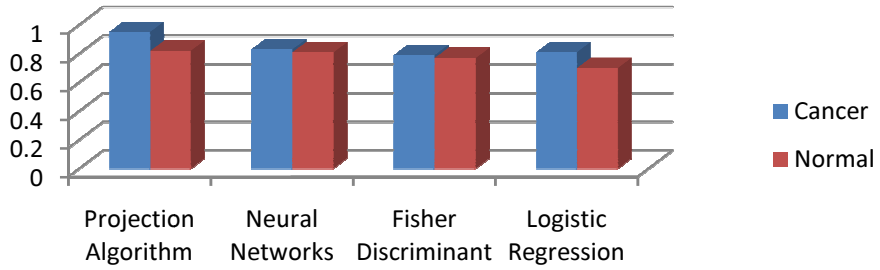

Figure 3. Probability of detection

The validation process consists of randomly selecting two tissues out of the 8 normal tissues and randomly selecting two cancer tissues out of the 25 cancer tissues to perform validation, and the remaining tissues were used for training. Thus, each classification method was trained with 23 cancer tissues and 6 normal tissues. Once a method was trained the 4 validation tissues were used to perform classification and validation. This process was repeated one thousand times and at the end of the classification exercise the measurements of accuracy scores were computed and results are summarized in Table 2. This table shows results for the cancer and normal statistics and also for the overall performance. Figures 3-5 show the comparison of the proposed projection algorithm with three existing classification methods. In general, Figures 3-5 and Table 2 show that the Projection algorithm outperform some of the existing classification methods.

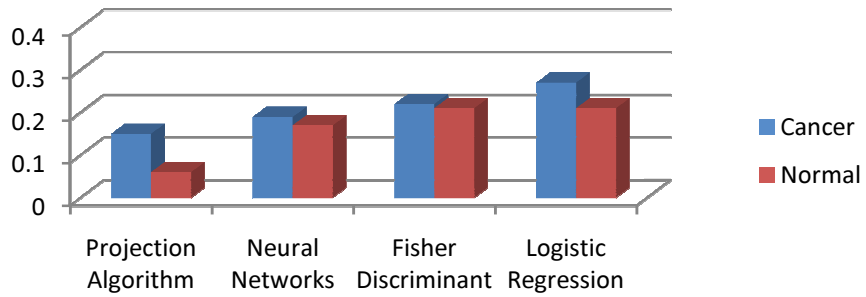

Figure 4. False alarm rate

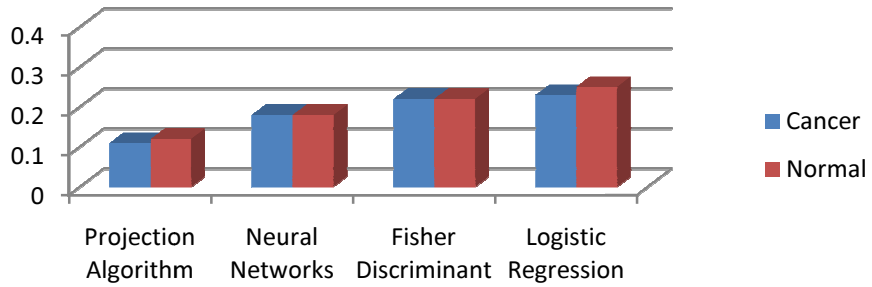

Figure 5. Misspecification Index

5. CONCLUSION

The proposed projection algorithm is a nonparametric statistical tool to perform classification of information into two or more categories. The algorithm uses the empirical probability distribution to create a vector and to represent the stochastic variability of a tissue based on the variability of the genes that exhibits the strongest differences between cancer and normal tissues. The projection algorithm shows that 89% of times classified correctly the cancer and normal tissues, whereas 82%, 78% and 75% of the time the Artificial Neural Network, the Fisher's Discriminant Function, and the Logistic Regression, correctly classify the cancer and normal tissues, respectively. It should be mentioned that smaller misspecification index and misspecification rate are associated with the Projection Algorithm and these results confirms that the Projection Algorithm outperforms the reference methods and therefore the Projection Algorithm can be used as a potential tool to perform tissues classification.

Results show that regardless of the used method the cancer probability of detection is larger than the normal probability of detection. These results indicated that there is more likely to identify a cancer tissue than a normal tissue, and this is good property that exhibits the classification methods.

A systematic procedure to perform compassion of classification methods was introduced in this work. The suggested measurements of classification accuracy are based on the probability distribution of all possible outcomes that are generated during the classification process. The probability statements generate an objective and a quantitative manner to measure the classification performance and therefore it is possible to identify without ambiguities the algorithm that best classifies a given set of information.

ACKNOWLEDGEMENTS

This research has been supported by Bio Science and Engineering Initiative Program of the University of Puerto Rico at Mayagüez. The authors appreciate the technical contribution of the principal investigator of this project UPRM BioSEI, with grant number 33010308030. The authors appreciate and recognize the funding support from this institution, and also recognize the invaluable contribution of the students Leemary Berrios, and Zahira Irizarry.

REFERENCES

[1] Genetic Science Learning Center DNA Microarray Virtual LAB. Learn. Genetics. Retrieved August 14, 2010, from http://learn.genetics.utah.edu/content /labs/microarray/

[2] Pang, H., Tong, T., and Zhao, H. Shrinkage-based Diagonal Discriminant Analysis and Its Applications in High-Dimensional Data. Biometrics, 2009. 1-9. DOI: 10.1111/j.1541-0420.2009.01200.x

[3] Lee, J.W., Lee, J.B., Park, M., and Song, S.H. An extensive comparison of recent classifcationtools applied to microarray data. *Computational Statistics & Data Analysis*, 48 (2005) 869-885.

[4] Wong, Y. F., Z. E. Selvanayagam, N. Wei, J. Porter, R. Vittal, R. Hu, Y. Lin, J. Liao, J. W. Shih, T. H. Cheung, K. W. Kit Lo, S. F. Yim, S. K. Yip, D. T. Ngong, N. Siu, L. K. Ying Chan, C. Sing Chan, T. Kong, E. Kutlina, R. D. McKinnon, D. T. Denhardt, K.V. Chin, and T. K. Hung Chung. Expression Genomics of Cervical Cancer: Molecular Classification and Prediction of Radiotherapy Response by DNA Microarray, Clinical Cancer Research, vol. 9, pp. 5486–5492. 2003. http://www.ncbi.nlm.nih.gov/geo/

[5] Strang, G., Linear Algebra and its Applications, Academic Press, New York, 1976, pp. 418.

[6] Liao, J G G, Khew-Voon, V. Chin Logistic regression for disease classification using microarray data: model selection in a large p and small n case. Bioinformatics. Vol. 23 Issue 15. 2007

[7] Zhoua, X., K-Y. Liua, and S.T.C. Wonga. Cancer classification and prediction using logistic regression with Bayesian gene selection Journal of Biomedical Informatics Volume 37, Issue 4, pp 249-259. 2004.

[8] Xing, E.P., M.I. Jordan, and R.M. Karp. Feature Selection for High-Dimensional Genomic Microarray Data, Proc. 18th Int'l Conf. Machine Learning. 2001.
[9] Montgomery, D.C., Peck, E.A, and Vining,G.G. Introduction to Linear Regression Analysis. Third Ed., John Willey and Sons, Inc. New York, 2001, pp 641.
[10] Kim, K-J., and Cho, S-B. Evolving Artificial Neural Networks for DNA Microarray Analysis. Evolutionary Computation, Vol. 4, pp 2370-2377. 2003.
[11] Linder R. Richards T., and Wagner M. Microarray data classified by artificial neural networks. Methods Mol Biol, 382: pp 345-372. 2007.
[12] Ramírez-Beltran, N.D, Kuligowski, R.J., Castro, J.M.,Cardona-Soto, M, Vasquez, R. A Projection Algorithm for Satellite Rainfall Detection. WSEAS Transaction on Systems. Issue 6, Vol 8, pp 763-772. 2009. http://www.worldses.org/journals/systems/systems-2009.htm
[13] Ramirez-Beltran, N.D. and Montes, J.A. Neural Networks to Model Dynamic Systems with Time Delays. IIE Transactions, Vol. 34, 313-327. 2002
[14] Hagan, M.T., Demuth, H.B., and Beal, M., Neural Network Design, PWS Publishing Company, Boston. 1996.
[15] Hagan, M.T., and Menhaj, M.. "Training Feedforward Networks with Marquardt Algorithm," IEEE Transaction on Neural Networks, Vol. 5, No. 6. 1994.
[16] Seber, G.A.F. Linear Regression Analysis. John Wiley & Sons, New York, p 465. 1977.
[17] Johnson, R.A., and D.W. Wichern. Apply Multivariate Statistical Analysis, Fourth Ed., p 816. 1998.

BIOGRAPHIES OF AUTHORS

Dr. Ramirez-Beltran gotB.S. in Industrial Engineering (in Mexico 1976),he got a M.S. (1983) and a Ph.D. (1988),both degrees in Industrial Engineering at Texas A&M University. Dr. Ramire-Beltran is a professor of the Departement of Industrial Engineering at the University of Puerto Rico.
He has published 79 papers and two book chapters and he is a reviewer of the following journals: Water Resources Research, Soil and Water Conservation, Computer & Chemical Engineering, Geophysical Research Letters, Journal of Applied Meteorology and Climatology, Canadian Journal of Soil Science, and Journal of Phase Equilibria and Difusion.
He has been supported by the following agencies and companies to conduct research: National Aeronautics and Space Administration (NASA), National Oceanic and Atmospheric Administration (NOAA), National Science Foundation (NSF), U.S. Department of Energy (DOE), U.S. Department of the Interior, Industry University Research Center (INDUNIV), Abbott Diagnostics, Inc., Abbott Chemical Inc., and MOVA Pharmaceutical, Co.
From January 2009 to July 2010 he developed a research project entitled Nonparametric Statistical Microarray Analysis for Cancer Research. Funded by R&D Center/BioSEI of University of Puerto Rico at Mayaguez. As a result of this research he developed an algorithm for detecting cancer using microarrays data.
Currently he is conducting a research project supported by NOAA. Since weather numerical prediction models at short term (1-2 hours) exhbit large prediction errors. Dr. Ramirez-Beltran is developing a short-term rainfall prediction algorithm. He is using infrared channels from GOES satalliete,the cloud dropsize distribution, and the cloud motion vector to predict the spatio temporal distution of rainfall rate and reduce the prediction error.

Joan Manuel Castro is a Ph.D. student at the Department of Civil Engineering of University of Puerto Rico. He got (2007) a M.S. in Electrial and Compueter and gineering at the University of Puerto Rico and B.S. (2002) in Electrial and Computer Engineering at University of Puerto Rico.
He was involved in a research project entitled: Improvement of the Hydro-Estimator and NEXRAD Rainfall Estimates over Puerto Rico: The general objective of this research was to develop an algorithm to improve the HE rainfall detection and rain rate estimation over tropical climate conditions by using the full suite of observations available from GOES and from a numerical weather prediction model.
He collaborate with other graduated students to help them to develop their thesis work and dictate seminars to undergraduate students to learn basic software applications on Industrial Engineering Department supported and supervised by Edwin Morales, Dr. William Hernandez, Prof. Mercedes Ferrer, Dr. Nazario Ramirez and Israel Tirado.