*I.J. Information Engineering and Electronic Business,* 2012, 2, 43-50
Published Online April 2012 in MECS (http://www.mecs-press.org/)
DOI: 10.5815/ijieeb.2012.02.07

# A Hybrid Data Mining Technique for Improving the Classification Accuracy of Microarray Data Set

Sujata Dash,
*KMBB College & Technology of Engineering, Bhubaneswar, India*
*E-mail:* sujata_dash@yahoo.com

Bichitrananda Patra
*KMBB College & Technology of Engineering , Bhubaneswar, India*
*E-mail:* bnpatra@gmail.com

B.K. Tripathy
*VIT-University, Vellore, Tamilnadu, India*
*E-mail:* tripathybk@rediffmail.com

***Abstract*****—**A major challenge in biomedical studies in recent years has been the classification of gene expression profiles into categories, such as cases and controls. This is done by first training a classifier by using a labeled training set containing labeled samples from the two populations, and then using that classifier to predict the labels of new samples. Such predictions have recently been shown to improve the diagnosis and treatment selection practices for several diseases. This procedure is complicated, however, by the high dimensionality of the data. While microarrays can measure the levels of thousands of genes per sample, case-control microarray studies usually involve no more than several dozen samples. Standard classifiers do not work well in these situations where the number of features (gene expression levels measured in these microarrays) far exceeds the number of samples. Selecting only the features that are most relevant for discriminating between the two categories can help construct better classifiers, in terms of both accuracy and efficiency. This paper provides a comparison between dimension reduction technique, namely Partial Least Squares (PLS)method and a hybrid feature selection scheme, and evaluates the relative performance of four different supervised classification procedures such as Radial Basis Function Network (RBFN), Multilayer Perceptron Network (MLP), Support Vector Machine using Polynomial kernel function(Polynomial- SVM) and Support Vector Machine using RBF kernel function (RBF-SVM) incorporating those methods. Experimental results show that the Partial Least-Squares(PLS) regression method is an appropriate feature selection method and a combined use of different classification and feature selection approaches makes it possible to construct high performance classification models for microarray data.

***Index Terms*****—**partial least square, feature reduction, feature selection, microarrays, gene expression.

## 1. Introduction

Classification of patient samples presented as gene expression profiles has become the subject of extensive study in biomedical research in recent years. One of the most common approaches is binary classification, which distinguishes between two types of samples: positive, or case samples (taken from individuals that carry some illness), and negative, or control samples (taken from healthy individuals). Supervised learning offers an effective means to differentiate positive from negative samples: a collection of samples with known type labels is used to train a classifier that is then used to classify new samples. Microarrays allow simultaneous measurement of tens of thousands of gene expression levels per sample. Because typical microarray studies usually contain less than one hundred samples, the number of features (genes) in the data far exceeds the number of samples. This asymmetry of the data poses a serious challenge for standard learning algorithms–that can be overcome by selecting a subset of the features and using only them in the classification. This feature selection step offers several advantages such as improved performance of classification algorithms, improved generalization ability of the classifier to avoid over-fitting, fewer features, making classifiers more efficient in time and space and more focused analysis of the relationship between a modest number of genes and the disease in question.

Many feature selection techniques have been proposed. One of the most basic and popular methods involves filters [26], which select the subset of features as a pre-processing step, independent of the chosen

Copyright © 2012 MECS

classifier. Being computationally simple and fast, they can handle extremely large-scale datasets. Furthermore, feature selection needs to be performed only once, after which different classifiers can be evaluated [26]. Most filters are univariate, considering each feature independently of other features–a drawback that can be eliminated by multivariate techniques. As such many proposed classification algorithms for microarray data have adopted various hybrid schemes. In these algorithms, the classification process usually has two steps, which we now outline.

In the first step, the original gene expression data is fed into a dimensionality reduction algorithm, which reduces the number of input variables or building a small number of linear or nonlinear combinations from the original set of input variables. The former approach is often known as variable selection while the latter is often known as feature selection. In the second step , classification models are trained on the data set with a reduced number of input attributes(created in the previous step) using an ordinary supervised classification algorithm.

In principle, many dimensionality reduction algorithms for supervised learning can be applied to the classification of gene expression data. Various two-step schemes have been presented and all of them reported improved classification accuracy. There is no conclusion from previous studies so far which confirms superiority of any particular scheme for microarray data classification.

In this study we developed a novel feature selection technique based on the Partial Least Squares (PLS) algorithm [30–32], which we call SIMPLS. PLS aims to obtain a low dimensional approximation of a matrix that is,'as close as possible' to a given vector. SIMPLS is a multivariate feature selection method based on PLS that incorporates feature dependencies. In the first step, we implemented two different dimensionality reduction schemes: (i) SIMPLS as the dimensionality reduction algorithm and (ii) an alternative and novel hybrid feature selection scheme which consecutively applied correlation based feature selector method [27] on the original data sets followed by the SIMPLS regression algorithm. Then in the second step, the two sets of filtered data with new features resulting from the two feature selection schemes described in the first step were separately fed into four supervised classification algorithms namely, Support Vector Machine using Polynomial kernel function, Support Vector Machine using RBF kernel function, Multilayer Perceptron and Radial Basis Function Network(RBFN). Three different expression profile datasets comprising a total of 215 samples were collected and used for training and testing. We then used these two schemes Our results show that the use of some SIMPLS variants leads to significantly better classification than that obtained with standard filters.

The use of PLS for classification is not new. In [21] the authors designed a procedure that entailed dimension reduction by PLS, followed by classification using the components constructed by PLS as the new extracted features; only a small subset of the total pool of genes was used for the construction of the components, selected by t-test. In [22] the authors extended this two-step procedure to support multiclass classification. Huang and Pan [12] used PLS and penalized regression for binary classification. First, q PLS components were constructed and a linear regression model was built using the components. Then using a penalizing procedure, only genes with coefficients larger than some threshold λ were kept. Both q and λ were determined by cross validation. The classification itself is obtained by the penalized linear regression model. A similar procedure was employed in [11] in order to combine information from two different datasets of gene expression. Quite recently, Cao et al. [2] used PLS-SVD (a variant of PLS that uses singular value decomposition) together with Lasso Penalty in order to integrate data coming from different sources for classification. The combination of PLS and linear regression techniques was further studied in [4]. Fort and Lambert-Lacroix [6] described a classification using PLS with penalized logistic regression; like [21], this study ran the t-test filter before applying PLS. The discriminating abilities of PLS were studied in [1], where the connection between PLS and Linear Discriminant Analysis is shown.

All the above studies used PLS for classification, and when feature selection was involved, it was implicitly used. For example, in [12], where a penalizing process was applied to reduce the number of genes, the threshold parameter λ, which implicitly determines the number of features, was found using cross validation. The SIMPLS method is unique in that it focuses solely on feature selection; it does not propose a new classification procedure. As a result, it can be used as a pre-processing stage with different classifiers. Thus, we evaluated the performance of SIMPLS with different classifiers, and compared it with a hybrid feature selector method and not to the PLS-based classification methods mentioned above. The rest of this paper is organized as follows.

We begin with a brief overview of the PLS and dimension reduction in the classification framework in Section 2 and in Section 3, classification algorithms are introduced. The experimental framework and results are described in Section 4. Finally, the conclusion and future work are presented in Section 5.

## 2. PLS and dimension reduction in the classification framework

The method denoted as Partial Least Squares (PLS) was originally developed as a multivariate regression tool in the context of chemometrics. An overview of the history of PLS regression is given in [18]. PLS regression is especially appropriated to predict a univariate or multivariate continuous response using a large number of continuous predictors. Suppose we have a n x p data matrix X. The centered data matrix $X_C$ is obtained by centering each column to zero mean. *Y* denotes a univariate continuous response variable and **Y** the n x 1 vector containing the realizations of Y for the n observations. The centered vector $Y_C$ is obtained by subtracting the empirical mean of *Y* from **Y**. From now on, Y denotes a categorical variable taking values 1 to K, with K ≥ 2. Y1…….. Yn denote the n realizations of Y . In this framework, PLS can be seen as a dimension reduction method: t1……$t_n$ € $R^n$ represent the observed m new components. Although the algorithm with orthogonal components has been designed for continuous responses, it is known to lead to good classification accuracy when it is applied to a binary response (K = 2), especially for high-dimensional data as microarray data [14] [16].The same can be said for the SIMPLS algorithm: a binary response can be treated as a continuous response, since no distributional assumption is necessary to use the SIMPLS algorithm. If the response is multi-categorical (K > 2), it can not be treated as a continuous variable. The problem can be circumvented by dummy coding. The multi-categorical random variable Y is transformed into a K-dimensional random vector y € $\{0,1\}^k$ as follows.

$$\begin{aligned} y_{i1} &= 1 \quad \text{if } Y_i = k, \\ y_{ik} &= 0 \quad \text{else,} \end{aligned} \qquad (1)$$

where yi = $(y_{i1}, y_{i2,\ldots\ldots,}\ y_{iK})^T$ denotes the ith realization of y. Y denotes the n x K matrix containing $y_i$ in its i-th row, for i = 1, ….., n. In the following, Y denotes the n x 1 vector Y = $(Y1,\ldots.., Yn)^T$, if Y is binary (K = 2) or the n x K matrix as defined above if Y is multi-categorical (K > 2). In both cases, the SIMPLS algorithm outputs a p x m transformation matrix A containing the $a_1$,……, $a_m$ € $R^P$ in its columns. The n x m matrix T containing the values of the new components for the n observations is computed as

$$\mathbf{T} = \mathbf{X}_C\mathbf{A}.$$

These new components can be used as predictors for classification. Whereas Huang and Pan [10] build a classical linear model to predict the class y, Nguyen and Rocke [23] use logistic regression and linear discriminant analysis. See [13] for an overview of classification methods. In this paper, we attempt to improve predictive accuracy by building a hybrid classification scheme for microarray data sets. In the first step, we implement Partial Least-Squares (PLS) regression [26, 30] as the dimensionality reduction algorithm, on the original data sets. Then in the second step, the filtered data with new features resulting from the feature reduction scheme described in the first step is fed into supervised classification algorithms such as Polynomial Support Vector Machine (SVM) [19], radial SVM [19], Multilayer Perceptron [19] and Radial Basis Function Network (RBFN) [19] to compare the results of the classifiers.

## 3. Related Algorithms

### 3.1. Partial Least Squares Regression

Partial least squares (PLS) regression aims to reduce the data dimensionality with a similar motivation, but differs from PCA by adopting a different objective function to obtain PLS components. Whereas PCA maximizes the variance of each coordinate and whereas both PCA and latent factor analysis will not take into account the values of the target (dependent) attribute, the PLS regression model attempts to find a small number of linear combinations of the original independent variables which maximize the covariance between the dependent variable and the PLS components. (PLS uses the entire data set: input and target attributes.) So the $i^{th}$ PLS component is given by

$$w_i = \underset{w^T w = 1}{\operatorname{argmax}}\, cov\{w^T x, y\}, \qquad (2)$$

Subject to

$$t_i^T t_j = 0, \text{ where } i \neq j, t_k = w_k^T x.$$

The PLS method can be illustrated by examining the following relations. Assuming X is an n x m matrix representing a data set of n instances with p independent variables, then if the number of PLS components is K, then the matrix X can be written as the summation of K matrices generated by outer products between vector ti (which is often known as the score vector) and $p_i^T$ (which is often called the load vector). The optimal number of PLS components, K, is usually determined by applying cross-validation methods on training data.

$$X = TP^T + E = \sum_{i=1}^{K} t_i p_i^T + E \qquad (3)$$

In effect, the relation in the PLS model projects the data vectors X from the original p-dimensional space into a (much lower than p) K-dimensional space. In the same way, when PLS components are used in the regression, the relation between dependent variable y and PLS component ti can be written as

$$Y = TBQ + F \qquad (4)$$

Where T is PLS components matrix, B is the coefficients vector so that TB is orthogonal, Q is the regression coefficients matrix, F is the residual matrix and $|F|$ is to be minimized. Partial least squares regression can be regarded as an extension of the multiple linear regression model. It has the advantage of being more robust, and therefore it provides a good alternative to the traditional multiple linear regression and principal component methods. The original PLS method was proposed by Wold [30] in the late 1960s and initially applied in the field of econometrics. Since then the method had been adopted in other research disciplines and been widely applied in many scientific analyses. SIMPLS [3] is an algorithm for partial least squares regression proposed by de Jong [3]. Compared to conventional nonlinear iterative partial least squares (NIPALS)-PLS, SIMPLS runs faster and is easier to interpret. In SIMPLS, the PLS components are calculated directly as linear combinations of the original variables, which avoids the construction of deflated data matrices. In this paper, we use the SIMPLS algorithm by de Jong [3], which can be seen as a generalization for multi-categorical response variables of the algorithm used by Nguyen and Rocke [23].

### 3.2. Correlation-based feature selection

CFS evaluates a subset of features by considering the individual predictive ability of each feature along with the degree of redundancy between them [14].

$$\mathrm{CFS}_S = \frac{k\overline{r}_{\mathrm{cf}}}{\sqrt{k + k(k-1)\overline{r}_{\mathrm{ff}}}} \qquad (5)$$

where CFS $_S$ is the score of a feature subset *S* containing *k* features, ¯*r*cf is the average feature to class correlation (f ∈ *S*), and ¯*r*ff is the average feature to feature correlation. The distinction between normal filter algorithms and CFS is that while normal filters provide scores for each feature independently, CFS presents a heuristic "merit" of a feature subset and reports the best subset it finds.

### 3.3. Radial Basis Function (RBF) Networks

RBF networks have 2 steps of processing. First, input is mapped in the hidden layer. The output layer is then a linear combination of hidden layer values representing mean predicted output. This output layer value is the same as a regression model in statistics [20]. The output layer, in classification problems, is usually a sigmoid function of a linear combination of hidden layer values. Performance in both cases is often improved by shrinkage techniques, also known as ridge regression in classical statistics and therefore smooth output functions in a Bayesian network. Moody and Darken [17] have proposed a multi-phase approach to RBFNs. This multi-phase approach is straightforward and is often reported to be much faster than, e.g., the back propagation training of MLP. A possible problem of the approach is that the RBF uses clustering method (e.g., k-means) to define a number of centers in input space and the clustering method is completely unsupervised and does not take the given output information into account. Clustering methods usually try to minimize the mean distance between the centers they distribute and the given data which is only the input part of the training data. Therefore, the resulting distribution of RBF centers may be poor for the classification or regression problem.

### 3.4. Support Vector Machines (SVM)

Support Vector Machines (SVMs) have been widely used in the recent years in the field of computational biology due to their high accuracy and their flexibility in modeling diverse sources of data. They are mainly used in binary classification and regression. They are very suitable for classifying microarray gene expression data [7]. Given a training set of instance-label pairs ($x_I$, $y_i$), i = -1,…, l where $x_i$ €$R^n$ and y €$\{1, -1\}^l$ , the support vector machines require the solution of the following optimization problem:

$$\begin{aligned}&\min_{\omega,b,\xi} \frac{1}{2}\omega^T\omega + C\sum_{i=1}^{l}\xi_i \\ &y_i(\omega^T\phi(x_i) + b \geq 1 - \xi_i \\ &\xi_i \geq 0\end{aligned} \qquad (6)$$

SVM finds a linear separating hyperplane with the maximal margin in this higher dimensional space. C > 0 is the penalty parameter of the error term.

$$K(x_i, x_j) = \Phi(x_i)^T\Phi(x_j) \qquad (7)$$

is called the kernel function [29].Here there are four basic kernels: linear, polynomial, radial basis function (RBF), and sigmoid:

$$\begin{aligned}&\text{Linear: } K(x_i, x_j) = x_i^T x_j \\ &\text{Polynomial: } K(x_i, x_j) = (x_i, x_j)^d \\ &\text{RBF: } K(x_i, x_j) = \exp\left(-\frac{\|x_i - x_j\|^2}{2\sigma^2}\right) \\ &\text{Sigmoid: } K(x_i, x_j) = \tanh(k(x_i x_j) + \vartheta)\end{aligned} \qquad (8)$$

SVMs use a kernel function to implicitly map data to a high dimensional space. Then, they construct the maximum margin hyperplane by solving an optimization problem on the training data. Sequential minimal optimization (SMO) [25] is used in this paper to train an SVM. SVMs have been shown to work well for high dimensional microarray data sets [8].

### 3.5. Multi-Layer Perceptron (MLP)

Error back propagation neural network is a feed forward multilayer perceptron (MLP) that is applied in many fields due to its powerful and stable learning algorithm [16]. The neural network learns the training examples by adjusting the synaptic weight according to the error occurred on the output layer. The back-propagation algorithm has two main advantages: local for updating the synaptic weights and biases, and efficient for computing all the partial derivatives of the cost function with respect to these free parameters. A perceptron is a simple pattern classifier. The weight-update rule in back propagation algorithm is defined as follows:

$$\Delta w_{ji}(n) = \alpha\Delta w_{ji}(n-1) + \eta\delta_j(n)y_i(n) \qquad (9)$$

where *w* is the weight update performed during the *n*th iteration through the main loop of the algorithm, η is a positive constant called the learning rate, δ is the error term associated with j, and 0≤ α <1 is a constant called the momentum [20][9,24].

## 4. Experiments

### 4.1. Datasets

Three widely used microarray gene expression data sets are chosen for our experiments: ALL-AML leukemia, lung cancer, and colon tumor. The data is taken from http://sdmc.lit.org.sg/GEDatasets/Datasets.html. Table 1 summarizes these datasets. We conducted the experiments on these three data sets by applying Partial Least Square (PLS) method for feature reduction and Polynomial Support Vector Machine (SVM) , Radial SVM , Multilayer Perceptron and Radial Basis Function Network(RBFN) for classification of the reduced datasets. We used Weka, a well known comprehensive toolset for machine learning and data mining [15] as our main experimental platform. We evaluated the performance of feature reduction in Weka environment with four classifiers, using 10-fold Cross Validation. We performed 10-fold Cross Validation on both the feature reduction process and the classification step.

Table 1. Three microarray datasets

| Dataset | # of genes | # of Instances | # of positive samples | # of negative samples |
|---|---|---|---|---|
| Leukemia | 7129 | 72 | 47(ALL) | 25(AML) |
| Colon Cancer | 2000 | 62 | 22 | 40 |
| Lung Cancer | 12533 | 181 | 31(MPM) | 150(ADCA) |

### 4.2. Methodology

In this study, the dimensionality reduction scheme is implemented as follows. Each column of the training set is normalized, so that each column has a mean of zero and variance of one. The values of the binary target attribute are set to either 0 or 1. Specifying the number of components for the Partial Least Square Regression, then a PLS model for a training data set is built by feeding the original training set into the SIMPLS algorithm. The output scores of the PLS algorithm are regarded as the values of input variables and forms the training set for the classification algorithms.

**Determining the optimal number of PLS components:**

Biologists often want statisticians to answer questions like ’which genes can be used for tumor diagnosis’? Thus, gene selection remains an important issue and should not be neglected. Dimension reduction is sometimes wrongly described as a black box which looses the information about single genes. In the following, we will see that PLS performs gene selection intrinsically. In this section, only binary responses are considered: Y can take values 1 and 2. We denote as $Y_C = (Y_{C1},\ldots\ldots, Y_{Cn})^T$ the vector obtained by centering $Y = (Y1,\ldots.., Yn)^T$ to zero mean:

$$\begin{aligned} Y_{Ci} &= -n_2/n \quad \text{if } Y_i = 1, \\ &= \quad n_1/n \quad \text{if } Y_i = 2, \end{aligned} \qquad (10)$$

where $n_1$ $n_2$ are the numbers of observations.

To perform PLS dimension reduction, it is not necessary to scale each column of the data matrix X to unit variance. However, the first PLS component satisfies an interesting property with respect to gene selection if X is scaled. In this section, the columns of the data matrix X are supposed to be have been scaled to unit variance and, as usual in the PLS framework, centered to zero mean. $a = (a_{1,} \ldots.., a_p)^T$, denotes the p $x_1$ vector defining the first PLS component as calculated by the SIMPLS algorithm.

A classical gene selection scheme consists of ordering the p genes according to BSSj / WSSj and selecting the top-ranking genes. For data sets with binary responses, we argue that $a_j^2$ can also be seen as a scoring criterion for gene j and we prove that the ordering of the genes obtained using BSSj / WSSj is the same as the ordering obtained using $a_j^2$. As a consequence, the first PLS component calculated by the SIMPLS algorithm can be used to order and select genes and the ordering is the same as the ordering produced by one of the most widely accepted selection criteria. Up to a constant, the BSS / WSS-statistic equals the F-statistic which is used to test the equality of the means within different groups. Since BSS / WSS is obtained by a strictly monotonic transformation of $a_j^{2'}$, $a_j^2$ can be seen as a test statistic itself. This PLS-based procedure for gene selection is much faster than the computation of BSS / WSS for each gene.

**Ten-fold Cross-Validation**: For each original data set, 100 pairs of training and test data sets are generated by repeating the 10-fold cross-validation method ten times. Then these 100 pairs of data sets are pre-processed by using procedures described at the beginning of this section. Then for each of 100 pairs of training and test sets which resulted from the above process, classification models were built and tested by using the four classification algorithms Support Vector Machine using Polynomial kernel function , Support Vector Machine using RBF kernel function , Multilayer Perceptron and Radial Basis Function Network(RBFN)) described in section 3.

Table 2. The optimal number of PLS components

| Dataset | RBFN | Polynomial SVM | RBF SVM | MLP |
|---|---|---|---|---|
| Leukemia | 04 | 50 | 8 | 20 |
| Colon Cancer | 08 | 40 | 20 | 40 |
| Lung Cancer | 20 | 50 | 50 | 50 |

Table 3. Predictive error (%) of classification algorithms, using SIMPLS Dimensionality Reduction scheme

| Dataset | RBFN | Polynomial SVM | RBF SVM | MLP |
|---|---|---|---|---|
| Leukemia | **0** | 0.45 | 28.22 | 0.41 |
| Colon Cancer | 10.95 | **0** | 23.33 | 0.31 |
| Lung Cancer | 11.55 | **0** | 16 | 0.95 |

Table 4. Predictive error (%) of classification algorithms, using a Hybrid Dimensionality Reduction scheme

| Dataset | RBFN | Polynomial SVM | RBF SVM | MLP |
|---|---|---|---|---|
| Leukemia | **2.86** | 3.88 | 31.11 | 4.75 |
| Colon Cancer | 32.46 | **17.13** | 33.89 | 22.53 |
| Lung Cancer | 8.65 | **1.91** | 10.95 | 0.75 |

### 4.3. Results and Discussions

Table 3 shows the classification performances of the four classification algorithms on three microarray data sets, with the lowest classification errors for each data set highlighted, SVM-Polynomial kernel achieves the lowest classification error in Colon and Lung data sets. The SVM-Polynomial returns excellent accuracy on noisy data such as microarray data. On the other hand, SVM with RBF kernel did not perform well on such noisy data sets. In general, SVM with Polynomial kernel and Multilayer Perceptron achieve higher predictive accuracies than the other two model.

Table 2 shows optimal number of components selected by SIMPLS algorithm. To determine the optimal number of PLS components, a simple cross-validation procedure is proposed. The reliability of this procedure is quite good, although not perfect [27].

When applying the SIMPLS method directly on the whole gene set from the original data, our tests returned improved classification accuracies on two (Colon and Leukemia) data sets and those reported in Tan and Gilbert’s paper [27]. The classification error rate of all the three dataset indicate that all dataset responded favorably to variable pre-selection for all the classifiers except few exceptions and the predictive accuracy is extremely high something like 100% for SIMPLS-SVM-Polynomial model for Colon data set, SIMPLS-RBFN model for Leukemia data set and SIMPLS-SVM-Polynomial model for Lung data set. The Leukemia and Colon cancer datasets indicate they were not largely affected by variable pre-selection for SIMPLS-SVM-RBF model and Colon and Lung data sets for SIMPLS-RBFN model and achieve predictive accuracy of approximately72% and 88% respectively.

In two-stage dimensionality reduction scheme, irrelevant genes were filtered out by correlation based feature selector method(CFS) [31] in the first step and in the second step, dimension of the data is further reduced by applying SIMPLS , a variant of PLS method .We processed the data using the above scheme, then applied the learning algorithms. These experimental results showed that, in, going from the SIMPLS scheme in Table 3 to the hybrid scheme in Table 4, only a marginal increase in classification accuracy of Lung cancer data set has been obtained.

SIMPLS a variant of PLS is a supervised procedure which uses the information about the class of the observations to construct the new components. Unlike sufficient dimension reduction, PLS can handle all the genes simultaneously and performs gene selection intrinsically. In other word, PLS is a very fast and competitive tool for classification problems with high-dimensional microarray data as regards to prediction accuracy. In future work, one could examine the theoretic connection between PLS and the four classification methods. Since the best classification accuracy is often reached with more than one PLS component, the subsequent PLS components could also be used to perform a refined gene selection. One could also try to improve the procedure to choose the number of components. It seems that cross-validation is appropriate, but a more sophisticated cross-validation scheme may improve the classification performance of our PLS-based approach.

## 5. Conclusion and Future Research

We conducted an extensive survey in the area of building classification models from microarray data with various supervised classification algorithms. Experimental results show that in most cases, the learning algorithms delivered classification accuracies equivalent to or better than those on the same data sets reported by other studies. Combined with the Partial Least-Squares (PLS) regression method, which is proved to be an appropriate feature selection method, the learning algorithms are capable of building classification models with high predictive accuracies from microarray data. As the study shows that our feature reduction scheme improves classification accuracies, one question immediately arises: will there be better hybrid schemes for the feature selection process for building supervised classification models? Since the number of instances in the studied microarray data is small and the performances of many classification algorithms are sensitive to the number of training data, another interesting question is raised: when comparing predictive performances of various classification algorithms on microarray data, what is the impact of adopting different methodologies such as ten-fold cross-validation, leave-one-out cross-validation and bootstrap [5]?

## References

[1] Barker M, Rayens W, Partial least squares for discrimination. journal of chemometrics, 2003, 17: 166–173.

[2] Cao K-AL, Roussouw D, Robert-Granie C, Besse P , A Sparse PLS for Variable Selection when Integrating Omics Data. Statistical Applications in Genetics and Molecular Biology, 2008, 7: Article 35.

[3] De Jong, S.: SIMPLS: an alternative approach to partial least squares regression. Chemometrics and Intelligent Laboratory Systems 2(4),1993, 251–263.

[4] Ding B, Gentleman R, Classification Using Generalized Partial Least Squares, 2004, Bioconductor Project.

[5] Efron, B.: Bootstrap methods: Another look at the jackknife. The Annals of Statistics, 1979, 7(1) 1–26.

[6] Fort G, Lambert-Lacroix S, Classification using partial least squares with penalized logistic regression, 2005, Bioinformatics 21: 1104–1111.

[7] Frank E, Hall M, Trigg L, Holmes G, Witten IH: Data mining in bioinformatics using Weka. *Bioinformatic* , 2004, 20(15):2479-2481.

[8] Furey, T.S., Cristianini, N., Duffy, N., Bednarski, D.W., Schummer, M., Haussler, D.: Support vector machine classification and validation of cancer tissue samples using microarray expression data. Bioinformatics, 2000, 16, 906–914.

[9] Greer BT, Khan J: Diagnostic classification of cancer using DNA microarrays and artificial intelligence. *Ann N Y Acad Sci , 2004,* 1020:49-66.

[10] Huang X, Pan W , Linear regression and two-class classification with gene expression data. Bioinformatics, 2003, 19: 2072–2078.

[11] Huang X, Pan W, Han X, Chen Y, Miller LW, et al. Borrowing information from relevant microarray studies for sample classification using weighted partial least squares. Comput Biol Chem, 2005, 29: 204–211.

[12] Huang, X., Pan, W., . Linear regression and two-class classification with gene expression data. Bioinformatics, 2003, 19, 2072–2078.

[13] Hastie, T., Tibshirani, R., Friedman, J. H., The elements of statistical learning. Springer-Verlag, 2001, New York.

[14] Hall, M.A., Correlation-based feature selection for machine learning. Ph.D. Thesis., 1999, Department of Computer Science University of Waikato.

[15] Ian H. Witten and Eibe Frank. *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann, San Francisco, 2 edition, 2005.

[16] Lippmann R.P., Moody J.E., Touretzky D.S., Neural Information Processing Systems.1991, *Morgan Kauffman*.

[17] Moody J.E., Darken C., Fast learning in networks of locally tuned processing units. *Neural Computation*,1989, 1:281-294.

[18] Martens, H., Reliable and relevant modelling of real world data: a personal account of the development of pls regression. Chemometrics and Intelligent Laboratory Systems, 2001, 58, 85–95.

[19] Mehdi, P., Jack Y. Y., Mary, Q. Y., Youping, D., A comparative study of different machine learning methods on microarray gene expression data, BMC Genomics,2008, 9(Suppl I):S13.

[20] Mitchell Tom M: Machine Learning. *McGraw-Hill*; 1997.

[21] Nguyen DV, Rocke DM, Tumor classification by partial least squares using microarray gene expression data. Bioinformatics, 2002, 18: 39–50.

[22] Nguyen DV, Rocke DM, Multi-class cancer classification via partial least squares with gene expression profiles. Bioinformatics, 2002, 18: 1216–1226.

[23] Nguyen, D., Rocke, D. M., Tumor classification by partial least squares using microarray gene expressio data. Bioinformatics, 2002, 18, 39–50.

[24] Narayanan A, Keedwell EC, Olsson B. (2002): Artificial intelligence techniques for bioinformatics. *Appl Bioinformatics*, 1(4):191-222.

[25] Platt, J., Fast training of support vector machines using sequential minimal optimization. Advances in Kernel Methods–Support Vector Learning. 1998, MIT Press.

[26] Saeys Y, Inza I, Larranaga P, A review of feature selection techniques in bioinformatics. Bioinformatics, 2007, 23: 2507–2517.

[27] Tan, A.C., Gilbert, D., Ensemble machine learning on gene expression data for cancer classification. Applied Bioinformatics, 2003, 2, S75-S83.

[28] V. Bolon-Canedo, A. Alonso-Betanzos,N. Sanchez-Marono , An ensemble of filters and classifiers for microarray Data classification, Pattern Recognition,2012, volume 45, Issue 1.

[29] Vapnik VN: Statistical Learning Theory, Adaptive and Learning Systems for Signal Processing, Communications And Control,1998, *Wiley New York.*

[30] Wold H , Soft modeling: the basic design and some extensions. Systems Under Indirect Observation,1982, 2: 1–53.

[31] Wold H, Partial least squares., Encyclopedia of the Statistical Sciences, , 1985, 6:581–591.

[32] Wold S, Ruhe H, Wold H, Dunn WJ III, The collinearity problem in linear regression. The partial least squares (PLS) approach to generalized inverse. SIAM Journal of Scientific and Statistical Computations, 1984, 5: 735 -743.

**Sujata Dash:** received her Ph.D. degree in Computational Modeling and Simulation from Berhampur University, Orissa, India in 1995. She is a Professor in Computer Science at KMBB College of Engineering and Technology, Biju Pattnaik University of Technology, has published more than 50 technical papers in international journals / Proceedings of international conferences / book chapters of reputed publications. Her current research interests includes Data Warehouse and Data Mining, Bioinformatics, Intelligent Agent, Web Data Mining and Wireless Technology.

**Bichitrananda Patra:** is an Assosiate Professor at the Department of Computer Science Engineering, at KMBB College of Engineering and Technology, Biju Patnaik University of Technology, Orissa, India, He received his master degree in Physics and Computer Science from the Utkal University, Bhubaneswar, Orissa, India. He is currently pursuing his Ph.D. in Computer Science at Berhampur University, Orissa, India. He has published research papers in international and natioanl journals and conferences and also having membership for different professional bodies like ISTE, CSI etc.

**B.K Tripathy:** is a senior professor in the school of computing sciences and engineering, VIT University, at

Copyright © 2012 MECS

Vellore, India, has published more than 140 technical papers in international journals/ proceedings of international conferences/ edited book chapters of reputed publications like Springer and guided 12 students for PhD. so far. He is having more than 30 years of teaching experience. He is a member of international professional associations like IEEE, ACM, IRSS, CSI, IMS, OITS, OMS, IACSIT, IST and is a reviewer of around 20 international journals which include World Scientific, Springer and Science Direct publications. Also, he is in the editorial board of at least 8 international journals. His current research interest includes Fuzzy sets and systems, Rough sets and knowledge engineering, Granular computing, soft computing, bag theory, list theory and social network analysis.