Gene Expression Data Classification with Revised Kernel Partial Least Squares Algorithm

Zhenqiu Liu[1], **Dechang Chen**[2]
[1]Department of Computer Science
Wayne State University, 110 Market Street, Frederick, MD 21703, USA.
[2]Division of Epidemiology and Biostatistics, PMB
Uniformed Services University of the Health Sciences, Bethesda, MD 20814, USA

Abstract

One important feature of the gene expression data is that the number of genes M far exceeds the number of samples N. Standard statistical methods do not work well when $N < M$. Development of new methodologies or modification of existing methodologies is needed for the analysis of the microarray data. In this paper, we propose a novel analysis procedure for classifying the gene expression data. This procedure involves dimension reduction using kernel partial least squares (KPLS) and classification with logistic regression (discrimination) and other standard machine learning methods. KPLS is a generalization and nonlinear version of partial least squares (PLS). The proposed algorithm was applied to five different gene expression datasets involving human tumor samples. Comparison with other popular classification methods such as support vector machines and neural networks shows that our algorithm is very promising in classifying gene expression data.

Introduction

One important application of gene expression data is classification of samples into different categories, such as the types of tumor. Gene expression data are characterized by many variables on only a few observations. It has been observed that although there are thousands of genes for each observation, a few underlying gene components may account for much of the data variation. PLS provides an efficient way to find these underlying gene components and reduce the input dimensions (Nguyen and Rocke 2002). PLS is a method for modeling a linear relationship between a set of output variables and a set of input variables and has been extensively used in chemometrics. In general, the structure of chemometric data is similar to that of microarray data: small samples and high dimensionality. With this type of inputs, linear least squares regression often fails, but linear PLS excels. Rosipal and Trejo (2001) and Bennett and Embrechts (2003) extended PLS to nonlinear regression using kernel functions, mainly for the purpose of real value predictions. Nguyen and Rocke (2002) applied PLS/PCA, together with logistic discrimination, to classify the tumor data and claimed success of their approach. However, their procedure is linear and limited with the implementation of SAS.

In this paper we propose a novel analysis procedure for classification of tumor samples using gene expression profiles. Our algorithm combines KPLS with logistic regression. Involved in our procedure are three steps: feature space transformation, dimension reduction, and classification. The proposed algorithm has been applied to five different popular gene expression datasets. One is a two-class recognition problem (AML *versus* ALL), and the other four concern multiple classes.

Algorithm

A gene expression dataset with M genes (features) and N mRNA samples (observations) can be conveniently represented by the following gene expression matrix

$$X = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1N} \\ x_{21} & x_{22} & \cdots & x_{2N} \\ \vdots & \vdots & \ddots & \vdots \\ x_{M1} & x_{M2} & \cdots & x_{MN} \end{bmatrix},$$

where x_{li} is the measurement of the expression level of gene l in mRNA sample i. Let $\mathbf{x}_i = (x_{1i}, x_{2i}, \ldots, x_{Mi})'$ denote the ith column (sample) of X, where $'$ represents the transpose operation, and y_i the corresponding class label (e.g., tumor type or clinical outcome).

PLS constructs a mapping of the data to a lower dimensional space and solves a least squares regression problem in a subspace. KPLS is a nonlinear version and generalization of PLS. To perform KPLS, one first transfers the input data from the original input space F_0 into a new feature space F_1 with a nonlinear function ϕ. Then a kernel matrix $K = [K(\mathbf{x}_i, \mathbf{x}_j)]_{N\times N}$ is formed using the inner products of new feature vectors. Denote by Φ the matrix whose i-th row is the vector $\phi(\mathbf{x}_i)'$, so that we have $K = \Phi\Phi'$. Finally, a PLS is performed on the feature space F_1. Such a linear PLS on the feature space F_1 may be viewed as a nonlinear PLS on the original data. This transition is sometimes called "kernel trick" in the literature.

The following are among the popular kernel functions:

- First norm exponential kernel

$$K(\mathbf{x}_i, \mathbf{x}_j) = \exp(-\beta||\mathbf{x}_i - \mathbf{x}_j||)$$

- Radial basis function kernel (RBF)

$$K(\mathbf{x}_i, \mathbf{x}_j) = \exp\Big(-\frac{|\mathbf{x}_i - \mathbf{x}_j|^2}{\sigma^2}\Big)$$

- Power exponential kernel (a generalization of RBF kernel)

$$K(\mathbf{x}_i, \mathbf{x}_j) = \exp\Big[-\Big(\frac{|\mathbf{x}_i - \mathbf{x}_j|^2}{r^2}\Big)^{\beta}\Big]$$

- Sigmoid kernel

$$K(\mathbf{x}_i, \mathbf{x}_j) = \tanh(\beta \mathbf{x}_i' \mathbf{x}_j)$$

- Polynomial kernel

$$K(\mathbf{x}_i, \mathbf{x}_j) = (\mathbf{x}_i' \mathbf{x}_j + p_2)^{p_1}$$

- Linear kernel

$$K(\mathbf{x}_i, \mathbf{x}_j) = \mathbf{x}_i' \mathbf{x}_j$$

KPLS Classification Algorithm

Suppose there is a two-class problem, and we are given a training data set $\{\mathbf{x}_i\}_{i=1}^n$ with class labels $\mathbf{y} = \{y_i\}_{i=1}^n$ and a test data set$\{\mathbf{x}_t\}_{t=1}^{n_t}$ with labels $\mathbf{y_t} = \{y_t\}_{t=1}^{n_t}$

1. Compute the kernel matrix, for the training data, $K = [K_{ij}]_{n\times n}$, where $K_{ij} = K(\mathbf{x}_i, \mathbf{x}_j)$. Compute the kernel matrix, for the test data, $K_{te} = [K_{ti}]_{n_t \times n}$, where $K_{ti} = K(\mathbf{x}_t, \mathbf{x}_i)$.
2. Centralize K and K_{te} using

$$K = \Big(\mathbf{I}_n - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n'\Big)K\Big(\mathbf{I}_n - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n'\Big),$$

and

$$K_{te} = \Big(K_{te} - \frac{1}{n}\mathbf{1}_{n_t}\mathbf{1}_n' K\Big)\Big(\mathbf{I} - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n'\Big).$$

3. Call KPLS algorithm to find k component directions (Rosipal and Trejo 2001):
   (a) for $i = 1, \ldots, k$,
   (b) initialize $\mathbf{u}^i$, $K^1 = K$, and $\mathbf{y}^1 = \mathbf{y}$.
   (c) $\mathbf{t}^i = \Phi\Phi'\mathbf{u}^i = K\mathbf{u}^i$, $\mathbf{t}^i \leftarrow \mathbf{t}^i/||\mathbf{t}^i||$.
   (d) $\mathbf{c}^i = \mathbf{y}^{i\prime}\mathbf{t}^i$
   (e) $\mathbf{u}^i = \mathbf{y}\mathbf{c}^i$, $\mathbf{u}^i \leftarrow \mathbf{u}^i/||\mathbf{u}^i||$
   (f) repeat steps (b) -(e) until converge.
   (g) deflate K^i, $\mathbf{y}^i$ by $K^{i+1} \leftarrow (I - \mathbf{t}^i\mathbf{t}^{i\prime})K^i(I - \mathbf{t}^i\mathbf{t}^{i\prime})$ and $\mathbf{y}^{i+1} \leftarrow \mathbf{y}^i - \mathbf{t}^i\mathbf{t}^{i\prime}\mathbf{y}^i$.
   (h) obtain component matrix $U = [\mathbf{u}^1, \ldots, \mathbf{u}^k]$.
4. Find the projections $\mathbf{V} = KU$ and $\mathbf{V}_{te} = K_{te}U$ for the training and test data, respectively.
5. Build a logistic regression model using $\mathbf{V}$ and $\{y_i\}_{i=1}^n$ and test the model performance using $\mathbf{V}_{te}$ and $\{y_t\}_{t=1}^{n_t}$.

We can show that the above KPLS classification algorithm is a nonlinear version of the logistic regression. In fact, it follows from our KPLS classification algorithm that the probability of the label y given the projection $\mathbf{v}$ is expressed as

$$P(y|\mathbf{w}, \mathbf{v}) = g\Big(b + \sum_{i=1}^{k} w_i v_i\Big), \qquad (1)$$

where the coefficients $\mathbf{w}$ are adjustable parameters and g is the logistic function

$$g(u) = (1 + \exp(-u))^{-1}.$$

Given a data point $\phi(\mathbf{x})$ in the transformed feature space, its projection v_i can be written as

$$v_i = \phi(\mathbf{x})\Phi'\mathbf{u}^i = \sum_{j=1}^{n} u_j^i K(\mathbf{x}_j, \mathbf{x}).$$

Therefore, from equation (1), we have

$$P(y|\mathbf{w}, \mathbf{v}) = g\Big(b + \sum_{j=1}^{n} c_j K(\mathbf{x}_j, \mathbf{x})\Big), \qquad (2)$$

where

$$c_j = \sum_{i=1}^{k} w_i u_j^i, \quad j = 1, \cdots, N.$$

When $K(\mathbf{x}_i, \mathbf{x}_j) = \mathbf{x}_i'\mathbf{x}_j$, equation (2) becomes a logistic regression. Therefore, KPLS classification algorithm is a generalization of logistic regression.

Described in terms of binary classification, KPLS algorithm can be readily employed for multi-class classification tasks. Typically, two-class problems tend to be much easier to learn than multi-class problems. While for two-class problems only one decision boundary must be inferred, the general c-class setting requires us to apply a strategy for coupling decision rules. For a c-class problem, we employ the standard approach where c two-class classifiers are trained in order to separate each of the classes against all others. The decision rules are then coupled by voting, i.e., sending the sample to the class with the largest probability.

Feature Selection

Since many genes show little variation across samples, gene selection is required. We chose the most informative genes with the highest scores, described below. Given a two-class problem with an expression matrix $X = [x_{li}]_{M\times N}$, we have, for each gene l,

$$T(\mathbf{x}_l) = \log\frac{\sigma^2}{\sigma'^2},$$

where

$$\sigma^2 = \sum_{i=1}^{N}(x_{li} - \mu)^2,$$

and

$$\sigma'^2 = \sum_{i\in \textit{class } 0}(x_{ij} - \mu_0)^2 + \sum_{i\in \textit{class } 1}(x_{ij} - \mu_1)^2.$$

Here μ, μ_0 and μ_1 are the corresponding mean values. We selected the most informative genes with the largest T values. This selection procedure is based on the likelihood ratio and was used in our classification.

Results

To illustrate the applications of the algorithm proposed in the previous section, we considered five gene expression datasets: LEUKEMIA (Golub et al. 1999), OVARIAN (Welsh et al. 2001), LUNG CANCER (Garber et al. 2001), LYMPHOMA (Alizadeh et al. 2000), and NCI (Ross et al. 2000).

LEUKEMIA

The LEUKEMIA dataset consists of expression profiles of 7129 genes from 38 training samples (27 ALL and 11 AML) and 34 testing samples (20 ALL and 14 AML). For classification of LEUKEMIA using KPLS algorithm, we chose the simple linear kernel $K(\mathbf{x}_i, \mathbf{x}_j) = \mathbf{x}_i'\mathbf{x}_j$ and 10 component directions. With 3800 or more genes selected, we obtained 0 training error and 0 test error. This performance of KPLS is superior to that of SVM, neural networks, and any other popular methods reported in the literature. A typical plot of the performance of KPLS in this case is given in Figure 1.

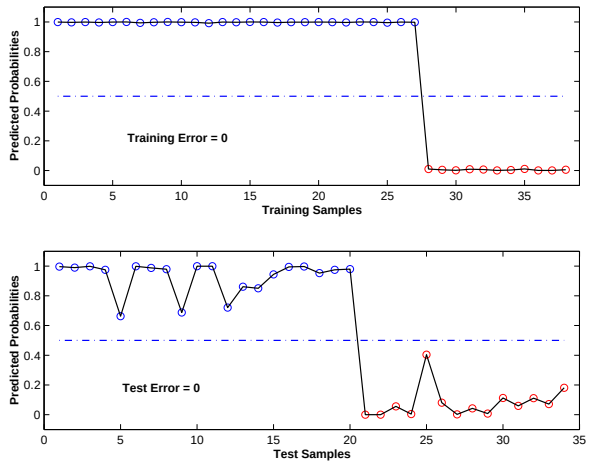

Figure 1: Performance on the training and test data of LEUKEMIA with all genes.

OVARIAN

The OVARIAN dataset contains expression profiles of 7129 genes from 5 normal tissues, 28 benign epithelial ovarian tumor samples, and 6 malignant epithelial ovarian cell lines. This dataset involves three classes. This problem was handled as follows. First, we built three binary classification problems using the well know "one versus all the others" procedure. For each two-class problem, KPLS was then applied to conduct prediction via the leave-one-out cross validation (LOOCV). LOOCV accuracy provides more realistic assessment of classifiers which generalize well to the test data. Finally, each sample was assigned to the class with maximal predicted probability value. With a linear kernel $K = \mathbf{x}_i'\mathbf{x}_j$, we trained the models with 150, 200, 250, 300 informative genes. For such a linear kernel, we also trained the model with all genes involved. In all cases, the test error is 1. A plot of the performance is shown in Figure 2.

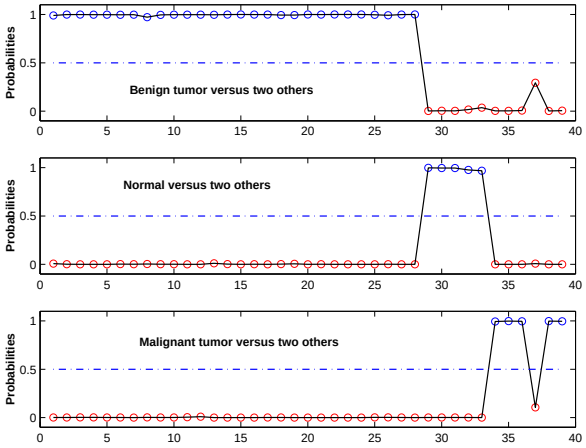

Figure 2: Performance on OVARIAN of KPLS with linear kernel $K = \mathbf{x}_i'\mathbf{x}_j$ and all genes.

The performance of KPLS may be improved by using nonlinear kernels. For example, with nonlinear polynomial kernel $K(\mathbf{x}_i, \mathbf{x}_j) = (\mathbf{x}_i'\mathbf{x}_j + 1)^2$, the test error is 0 for models with 150, 200, 250, and 300 informative genes. A plot of the performance in this scenario is shown in Figure 3. Figure 3 shows the predicted probability from each binary classifier with nonlinear polynomial kernel $K(\mathbf{x}_i, \mathbf{x}_j) = (\mathbf{x}_i'\mathbf{x}_j + 1)^2$. In Figure 3, the probability for benign tumor 14 being a tumor is 0.9999 and the probability for it being normal is 0.73. Based on the voting method, benign tumor 14 is classified correctly as a tumor. KPLS with a nonlinear kernel performs better than that with a linear kernel in this real application.

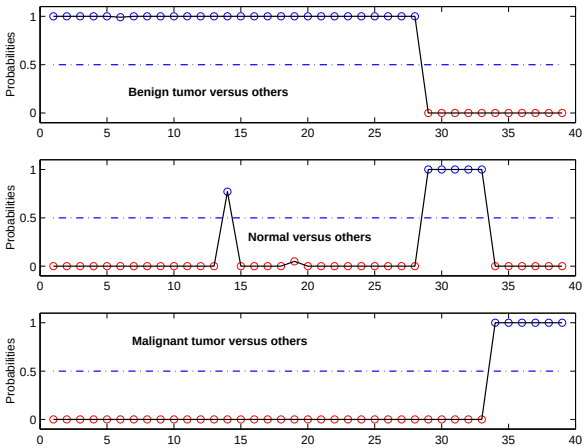

Figure 3: Performance on OVARIAN of KPLS with nonlinear kernel $K(\mathbf{x}_i, \mathbf{x}_j) = (\mathbf{x}_i'\mathbf{x}_j + 1)^2$ and 100 genes.

LUNG CANCER

LUNG CANCER dataset has 918 genes, 73 samples, and 7 classes. 100 most informative genes were selected for

the classification. The computational results of KPLS and other methods are shown in Table 1. The results from SVM for LUNG CANCER, LYMPHOMA, and NCI shown in this paper are those in Ding and Peng (2003). The six misclassifications of KPLS are provided in Table 2.

Table 1: **Comparison for LUNG CANCER**

Methods	Number of Errors
KPLS	6
PLS	7
SVM	7
Logistic Regression	12

Table 2: **Misclassifications of LUNG CANCER**

Sample Number	True Class	Predicted Class
6	6	4
12	6	4
41	6	3
51	3	6
68	1	5
71	4	3

LYMPHOMA

LYMPHOMA dataset has 4026 genes, 96 samples, and 9 classes. 300 most informative genes were selected for the classification. A comprison among KPLS and other methods is shown in Table 3. Misclassifications of LUNG CANCER with KPLS are given in Table (4). We can see there are only 2 misclassifications of class 1 (NSCLC) with our KPLS algorithm.

Table 3: **Comparison for LYMPHOMA**

Methods	Number of Errors
KPLS	2
PLS	5
SVM	2
Logistic Regression	5

NCI

NCI dataset has 9703 genes, 60 samples, and 9 classes. A comparison of computational results is summarized in Table 5 and the details of misclassification are listed in Table 6. KPLS performs extremely well for this particular dataset.

Conclusion

We have introduced a nonlinear method for classifying gene expression data by KPLS. The algorithm involves nonlinear transformation, dimension reduction, and logistic classification. We have illustrated the effectiveness of the algorithm in real life tumor classifications. Computational results show

Table 4: **Misclassifications of LYMPHOMA**

Sample Number	True Class	Predicted Class
64	1	6
96	1	3

Table 5: **Comparison for NCI**

Methods	Number of Errors
KPLS	3
PLS	6
SVM	12
Logistic Regression	6

that the procedure is able to distinguish different classes with a high accuracy. Our future work will focus on providing a rigorous foundation for the algorithm proposed in this paper.

Acknowledgement

The authors express their gratitude to Dr. Jaques Reifman for useful discussions and suggestions. He provided many fresh ideas to make this paper much more readable.

References

Alizadeh, A. A., et al. Distinct types of the diffuse large B-cell lymphoma identified by gene expression profiling. *Nature*, 2000, 403, 503-511.

Bennett, K. P. and Embrechts, M. J. An optimization perspective on partial least square. In Advances in Learning Theory: Methods, Models and Applications, NATO Science Series III: Computer & Systems Sciences, Volume 190, IOS Press, Amsterdam, 2003, 227-250.

Ding, C. and Peng, H. Minimum redundancy feature selection from microarry gene expression data.*CSB* 2003, 523-528.

Garber, M. E., et al. Diversity of gene expression in adenocarcinoma of the lung. *PNAS*, USA. 2001, 98(24), 13784-13789.

Golub, T. R., et al. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, 286, 531-537, 1999.

Nguyen, D. and Rocke, D. M. Partial least squares propor-

Table 6: **Misclassifications of NCI**

Sample Number	True Class	Predicted Class
6	1	9
7	1	4
45	7	9

tional hazard regression for application to DNA microarray data. *Bioinformatics*, 2002, 18, 29-50.

Nguyen, D. and Rocke, D. M. Multi-class cancer classification via partial least squares with gene expression profiles. *Bioinformatics*, 2002, 18, 1216-1226.

Rosipal, R. and Trejo, L. J. Kernel partial least squares regression in RKHS, Thoery and empirical comparison, Technical report, University of Paisley, UK, March 2001.

Ross, D.T., et al. Systematic variation in gene expression pattern in human cancer cell lines. *Nature Genetics*, 2000, 24(3), 227-234.

Welsh, J. B., et al. Analysis of gene expression in normal and neoplastic ovarian tissue samples identifies candidate molecular markers of epithelial ovarian cancer. *PNAS*, USA, 2001, 98, 1176-1181.