

Cancer Prediction with Kernel PLS and Gene Expression Profile

Zhenqiu Liu, Bioinformatic Cell/ TATRC
Decheng Chen, Uniformed Services University
of the Health Sciences
Jaques Reifman, Bioinformatic Cell/ TATRC

August 25, 2004

1. Introduction

A gene expression matrix with M genes and N mRNA samples can be written as

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

where x_{li} is the measurement of the expression level of gene l in mRNA sample i .

The i th column is also denoted by \mathbf{x}_i .

- For gene expression data, M (# genes) far exceeds N (# samples)
- Standard learning methods do not work well when $N < M$
- Development of new methodologies or modification of existing methodologies is needed

In this talk, we propose a novel procedure for classifying the gene expression data.

- dimension reduction via kernel partial least squares (KPLS)
- classification via logistic regression

2. Partial Least Squares (PLS)

- models linear relationship between output variables and input variables
- maps data to a lower dimensional space and then solves a least squares problem
- probably least restrictive among extensions of the multiple linear regression methods

3. Kernel Partial Least Squares (KPLS)

KPLS is a nonlinear version and generalization of PLS.

The procedure is:

- transform the input data from the original input space F_0 into a new feature space F_1
- perform PLS on the feature space F_1

When performing KPLS, a kernel matrix

$$K = [K(\mathbf{x}_i, \mathbf{x}_j)]_{N \times N}$$

is formed using the inner products of new feature vectors.

- Polynomial kernel

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

- Exponential kernel

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

4. Proposed Classification Algorithm

Suppose there is a two-class problem

We are given a training data set $\{\mathbf{x}_i\}_{i=1}^n$ with class labels $\mathbf{y} = \{y_i\}_{i=1}^n$

We are given a test data set $\{\mathbf{x}_t\}_{t=1}^{n_t}$ with labels $\mathbf{y}_t = \{y_t\}_{t=1}^{n_t}$

Step 1.

For the training data, compute the kernel matrix, $K = [K_{ij}]_{n \times n}$, where $K_{ij} = K(\mathbf{x}_i, \mathbf{x}_j)$.

For the test data, compute the kernel matrix, $K_{te} = [K_{ti}]_{n_t \times n}$, where $K_{ti} = K(\mathbf{x}_t, \mathbf{x}_i)$.

Step 2.

Centralize K using

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

Centralize K_{te} using

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

Step 3.

Call a KPLS algorithm to find k component directions $\mathbf{u}_1, \dots, \mathbf{u}_k$.

Set $U = [\mathbf{u}_1, \dots, \mathbf{u}_k]$.

Step 4.

Find the projections $\mathbf{V} = KU$ and $\mathbf{V}_{te} = K_{te}U$ for the training and test data, respectively.

Build a logistic regression model using \mathbf{V} and $\{y_i\}_{i=1}^n$.

Test the model performance using \mathbf{V}_{te} and $\{y_t\}_{t=1}^{n_t}$.

5. Some Notes

- Can show that the above algorithm is a nonlinear version of the logistic regression
- For a c -class problem, we train c two-class classifiers. The decision rules are then coupled by voting, i.e., sending the sample to the class with the largest probability.

6. Feature Selection

Given $X = [x_{li}]_{M \times N}$, calculate, for gene l ,

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

where

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

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

We selected genes with the largest T values.

7. Experiments on 5 Datasets

- LEUKEMIA (Golub et al. 1999)
- OVARIAN (Welsh et al. 2001)
- LUNG CANCER (Garber et al. 2001)
- LYMPHOMA (Alizadeh et al. 2000)
- NCI (Ross et al. 2000).

Results show our algorithm is very promising.

1. LEUKEMIA dataset consists of expression profiles of 7129 genes from 38 training samples and 34 testing samples. Both training and test error are zero with KPLS.

2. 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. \circ test error achieved with leave-one-out method.

3. LUNG CANCER dataset has 918 genes, 73 samples, and 7 classes. A Comparison of the Performance:

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

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

4. LYMPHOMA dataset has 4026 genes, 96 samples, and 9 classes. A Comparison of the Performance:

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

Misclassification of Lymphoma:

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

5. A comparison for NCI data
(9703 genes, 60 samples, 9 classes):

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

Misclassification of NCI:

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

12. Conclusion

- The proposed algorithm involves nonlinear transformation, dimension reduction, and logistic classification.
- Results show that the procedure is able to predict with a high accuracy.