*I.J. Information Technology and Computer Science,* **2011, 4, 1-8**
Published Online August 2011 in MECS (http://www.mecs-press.org/)

# Several cancer classifiers combined with PLS-DR for base on gene expression profile

JianGeng Li, Hui Li
Beijing University of Technology
Building Science 1126,BJUT
Beijing, P.R.China,100124
+86 010-67396155
lijg@bjut.edu.cn
shinedennis@163.com

*Abstract*—**It is known that Logistic Regression coupled with Partial Least Squares dimension reduction (PLSDR-LD) is capable of extracting a great deal of useful information for classification from gene expression profile and getting a rather high classification accuracy rate. In this study, we replace the logistic function of Logistic Regression with several functions which are similar to logistic function in appearance, and apply these functions to the analysis of microarray data sets from two cancer gene expression studies. We compare these newly introduced models with PLSDR-LD proposed in the literature. The most effective models with good prediction precision are lastly provided through analyzing the results of two experiments.**

***Index Terms-Logistic Regression; Partial Least Squares; gene expression profile; PLSDR-LD***

## I. INTRODUCTION

A vast amount of data generated through gene microarray, gives us sufficient biological information. Meanwhile, it brings us great difficulties to manage, integrate and interpret these data sets. Comprehensive approaches are needed to take full advantage of the huge information offered by the data. As a generally used statistical modelling method, PLS is first used in econometric path modeling by Herman Wold and afterwards it is used in chemometric and spectrometric modeling as a multivariate regression too[11-13]. Nguyen and Rocke proposed using PLSDR(PLS based dimension reduction) for dimension reduction as a preliminary step of classification, based either on linear logistic discrimination, linear or quadratic discriminant[9-10]. What attracts our attention is the classification method namely logistic discrimination. LD performs well in classification and has its own specialty when dealing with dummy variable. In this article, the application of sigmoid function to classification attains even better effect and a contrast with other like-logistic model will be given in the METHODS section. In recent years, many researches on PLSDR have been carried on embracing a great many aspects in various domains[7-8], besides, comparison have been conducted in the dimension reduction and discrimination realm[2]. The results of many experiments carried in the paper mentioned above demonstrate that PLSDR is an effective technique for dimension reduction. So PLSDR will be adopted in this paper, and the classical algorithm PCA is given to make a compare. Considering the fact that PLSDR is good at correlation but weak at remove irrelevant features among a set of complex features[14]. It is necessary to combine eliminating noise(irrelevant features) and dimension reduction into the ultimate model. As a result, we propose a model begin with irrelevant genes elimination. When the preliminary gene selection is finished, the process of dimension reduction (PLSDR and PCA) and classification are immediately implemented. These models would be used to dispose gene data sets. Some like-logistic functions are used in discrimination and the relevant results of classification are given in the Experiment part. In order to make a comprehensive comparison with logistic function, similar data processing method are adopted as Nguyen et al 2002b. The relevant indicators such as number of misclassification will be given in order to measure their swords in handling with biological information. Four data sets are used in the Experiment section in order to provide a stable result.

## II. METHODS

### A *Gene Selection*

The original gene data is rich of various kinds of biological information, but only a number of genes are of interest in our experiment. Of course, other genes are not of no use. They are just ‘misplaced resources’ and may play an important role in other practice. In every special experiment, the ‘noise’ may refer to different genes. So the elimination of irrelevant genes is necessary to improve the accuracy of classification and cut down on the calculation time spent on classification. Although the computing capacity of computer continues to expand, reducing the high computing complexity is of great significance because of the enormous volume of gene data sets. So in this article we employ the t-statistic scores to select the important genes. In this article, we just consider of binary classification problem. So the samples belong to two classes and the t-statistic score is given as:

$$t = \frac{\bar{x}_0 - \bar{x}_1}{\sqrt{\mathrm{var}(x_0)/N_0 + \mathrm{var}(x_1)/N_1}} \qquad (1)$$

$\bar{x}_0$, $\bar{x}_1$ is the mean expression value of two classes for a single gene. var( $\bar{x}_0$ ) and var( $\bar{x}_1$ ) are the variance respectively. $N_0, N_1$ is the size of each class. Obviously, the bigger t-statistic score represents the more different expression in the two classes for a single gene. It could mean that a single gene behaves disorder in one type of the sample. So we could suspect that this gene is somehow related to the classification of samples.

### B *Some Classification Function*

We find that some functions which like logistic can be used in classification problems. All of them are strictly monotonically increasing and invertible. The domains of those functions are all specified in $(-\infty, +\infty)$, the value domains are in $[-1, +1]$.

In this paper, three logistic-like functions will be applied to the model of classification. The functional form are denoted in (3)~(5).

$$f(x)=\frac{e^{u_k(x)}-e^{-u_k(x)}}{2(e^{u_k(x)}+e^{-u_k(x)})}+\frac{1}{2} \qquad (3)$$

$$f(x)=\frac{1}{\pi}\arg\tan(u_k(x))+\frac{1}{2} \qquad (4)$$

$$f(x)=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{u_k(x)} e^{-\frac{t^2}{2}}dt \qquad (5)$$

The first one is a transformation of sigmoid function which is often used BP neural network. The second one is also a like-logistic function which is named Arg function for convenience. The last one is Gaussian Error function. The discriminate methods employing these three functions are expressed as SD AD and GD respectively in this paper.

### *C Sigmoid Discriminate*

Let X is the gene expression data set. Before all the treatment on this set, it should be centered to zero mean. The original data X is n*p(i=1,2,3…,n; j=1,2,3…,p), the column of X represents the microarray sample of data, the row represent the expression level for each gene. After preliminary gene screening and dimension reduction, the number of columns of independent variable namely X has reduced from p to k(k<n). In binary classification, for example, $y_j$ =0( $y_j \in$ {0,1}) for a normal sample, while $y_j$ =1 for a cancer sample.

If we use linear probability model to simulate the relationship between X and Y, encountering lots of problems will be inevitable. Firstly, y should not be limited in a range of values, because independent variable X, regression coefficients and residuals are preferable of an arbitrary value. However, in our experiments, y is dummy variables. Secondly, there is different variance for different observations. It is contrary to Gauss - Markov assumption. This is called hereroscedasticity. Thirdly, the relationship between observations and variables is mostly likely nonlinear[3]. So in this section, we introduce a new kind of model which is inspired by logistic regression[7]. In sigmoid regression, the condition class probability, $\pi$ = P(y=1|x) = P(sample is classified to tumor for gene profile x) is modeled using the sigmoid functional form.

Sigmoid function is usually used in BP network. It is denoted as:

$$\pi(k\,|\,x)=\frac{e^{\beta_{i0}+x_{i1}\beta_{i1}+x_{i2}\beta_{i2}+\ldots x_{ip}\beta_{ip}}-e^{-\beta_{i0}-x_{i1}\beta_{i1}-x_{i2}\beta_{i2}-\ldots x_{ip}\beta_{ip}}}{e^{\beta_{i0}+x_{i1}\beta_{i1}+x_{i2}\beta_{i2}+\ldots x_{ip}\beta_{ip}}+e^{-\beta_{i0}-x_{i1}\beta_{i1}-x_{i2}\beta_{i2}-\ldots x_{ip}\beta_{ip}}}$$

k $\in$ {0,1}

(2)

$u_k(x)$ = $\beta_{i0}+x_{i1}\beta_{i1}+x_{i2}\beta_{i2}+\ldots x_{ip}\beta_{ip}$ . So the ultimate conditional class probability is

$$\pi(k\,|\,x)=\frac{e^{u_k(x)}-e^{-u_k(x)}}{2(e^{u_k(x)}+e^{-u_k(x)})}+\frac{1}{2} \quad \text{k} \in \{0,1\} \qquad (3)$$

Estimate of $\beta$ is obtained by maximum likelihood estimation(MLE). The estimated value is denoted as $\beta_E$. With this value, a given sample can be endowed with a sigmoid probability value. With this value the sample can be classified to corresponding category by comparison to the mean of $\pi(k\,|\,x)$ of each sample. For example, k=1 denotes normal, k=0 denotes cancer. If the value of a sample $\pi(k\,|\,x)$ is larger than the mean, the sample should be classified as normal, else as cancer.

Owing to the same principle behind classification method and the lack of space, the detail of Arg discriminate(AD) and Gausian discriminate(GD) are not given in this section.

### *D PLSDR and PCA*

Here we give the process of the algorithm PLSDR,

1) Standardize the data,

$$\tilde{x}_{ij}=\frac{x_{ij}-\overline{x_j}}{s_j} \text{ (i=1,2,….n; j=1,2,……p)} \qquad (4)$$

$$\tilde{y}_i=\frac{y_i-\overline{y}}{s_y} \qquad \text{(i=1,2,…n)} \qquad (5)$$

$\overline{x_j}$ is the mean of $x_j$, $s_j$ is the standard deviation of $x_j$, $\overline{y}$ is the mean of $y_i$, $s_y$ is the standard deviation of $y_i$.We use $E_0$ to denote the standardized form of X, $F_0$ denotes Y.

2) We extract the first component $u_1$ from $F_0$, $u_1 = F_0\ c_1$, $\|c_1\|$=1. And we extract the first component $t_1$ from $E_0$, $\|w_1\|$=1. In binary classification, $u_1 = F_0$, so,

$$E_0^{'}F_0=(E_{01},E_{02},...,E_{0P})^{'}F_0=(r(x_1,y),r(x_2,y),...,r(x_p,y))' \quad (6)$$

$$w_1=\frac{1}{\sqrt{\sum_{j=1}^{p}r^2(x_j,y)}}\begin{bmatrix} r(x_1,y) \\ r(x_2,y). \\ \ldots \\ r(x_p,y) \end{bmatrix} \qquad (7)$$

$$t_1 = E_0\ w_1 =$$

$$\frac{1}{\sqrt{\sum_{j=1}^{p}r^2(x_j,y)}}[r(x_1,y)E_{01}+r(x_2,y)E_{02}+...+r(x_p,y)E_{0P}]($$

8)

3) We can get the residual matrix $E_h$ and $F_h$,

$$E_h=E_{h-1}-t_h p_h{}'=E_{h-1}-t_h\frac{E_{h-1}{}'t_h}{\|t_h\|^2} \qquad (9)$$

$$F_h=F_{h-1}-t_h r_h{}'=F_{h-1}-t_h\frac{F_{h-1}{}'t_h}{\|t_h\|^2} \qquad (10)$$

4) We use $E_h$ to replace $E_0$, $F_h$ to replace $F_0$. Repeat the second step. Lastly, we get all $t_h$ and $u_h$.

The process of PCA is given,

1) Standardize the data,

$$\tilde{x}_{ij} = \frac{x_{ij} - \overline{x_j}}{s_j} \qquad （i=1,2,….n; j=1,2,……p） \qquad (11)$$

$\overline{x_j}$ is the mean of $x_j$, $s_j$ is the standard deviation of $x_j$.

2) Calculate the correlation matrix R (p*p) of X.

3) Calculate the eigenvalues $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_h$ ,and the relevant eigenvectors $a_1, a_2, \ldots, a_h$.

4) Get the hth principal component,

$$F_h = Xa_h = \sum_{j=1}^{p} a_{hj} x_j \qquad (12)$$

$a_{hj}$ is the jth component of $a_h$.

## III. EXPERIMENTS

### A Acute leukemia data

Gloub et al.(1999) published the acute leukemia data set, which is widely used. The original data set consisted of 47 Acute Lymphoblastic Leukemia(ALL) and 25 Acute Myeloid Leukemia(AML). It consisted of 62 bone marrow samples(from adult patients) and 10 peripheral blood specimens(from adults and children) containing probes for 7129 genes.

At first, gene selection comes on stage. After that, PCA or PLSDR is used to decrease the dimension. In every step of PCA and PLSDR iteration, there is an residual. The explanatory power(EP) of principal components and latent variables can be represented by ratio of residual and dependent variable.

Figure1 and Figure2 give the EP for predictor variables and response variables of the top 10 LVs and PCs. From Figure 1, we can conclude that PCA performs better than PLSDR in interpreting response variables. The top ten of PCs can stand for 95.96% of predictor variables while PLSDR is 94.91%. They give the similar results in interpreting the predictor variables. However, the PCA gives a much worse performance in EP for dependent variable. Note that the EP values bounces around the fourth PC and then goes steady. This is typical of PCA because the factors are not determined with regard to response variables. So the sharp difference between PLSDR and PCA in Figure 2 is beyond doubt. From Figure 1 and Figure 2, we can also conclude that the number of PCs or LVs(latent variables) in Nguyen et al[10] make sence. The explanatory power for independent variable increase less than 2% when putting the 4th LV into consideration. Meanwhile, we get the value of $\frac{S_{PRESS,4}}{S_{SS,3}}$ through SIMCA-P, it is 0.9425 > 0.9025. So the 4th LV is of no use in decreasing the predictive error of classification model.

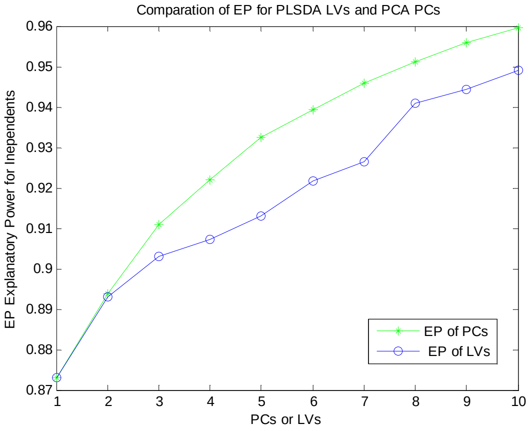

**Fig. 1 Explanatory Power for Independent Variable of PLSDR and PCA**

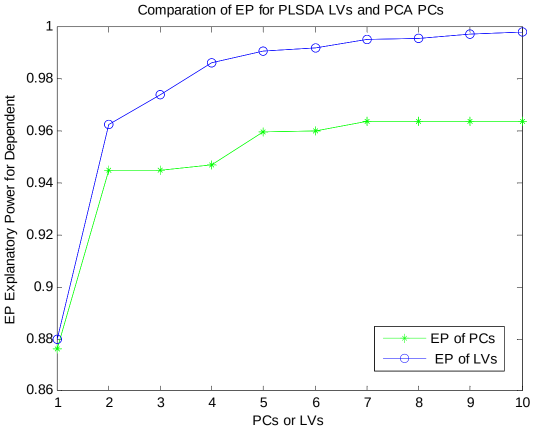

**Fig. 2 Explanatory Power for Dependent Variable for PLSDR and PCA**

We divide the original data set into two parts. One is used as training samples(38 samples, 27 ALL, 11 AML), while the remaining are used as test samples(34 samples, 20 ALL, 14 AML). For the training data, we use leave-one-out Cross-Validation to assess the fitness of LD SD AD and GD. All methods predicted the ALL/AML class correcly 100% for the 38 training samples with all these discriminate methods. The test data set are used to evaluate the performance of models and which provides additional protection against overfitting.

One of the AML samples (#66) is classified into ALL incorrectly by LD SD AD and GD, which is also misclassified by Golub et al 1999 using a weighted voting scheme. In table 2, SD and GD give an equivalent performance in the discrimination procedure. When P*=500 and P*=1000, SD and GD get the best result among the four methods. AD's classification ability is relatively weaker than the other three methods. It makes another error in the test data set (#42) besides the sample #66 based on P*=50.

**Table 1. ALL-AML data classification results of LD SD AD and GD(pca). Given are the number of samples correctly classified in the 38 training set and 34 test set.**

| P* | Training data (leave-out-one CV) | | | | Test data (out-of-sample) | | | |
|---|---|---|---|---|---|---|---|---|
| | LD[1] | SD | AD | GD | LD | SD | AD | GD |
| 50 | 38 | 38 | 38 | 38 | 33 | 33 | 31 | 33 |
| 100 | 38 | 38 | 38 | 38 | 32 | 32 | 30 | 32 |
| 500 | 38 | 38 | 38 | 38 | 31 | 31 | 30 | 31 |
| 1000 | 38 | 38 | 38 | 38 | 31 | 31 | 30 | 31 |
| 1500 | 38 | 38 | 38 | 38 | 30 | 31 | 29 | 31 |

**Table 2. ALL-AML data classification results of LD SD AD and GD(pls). Given are the number of samples correctly classified in the 38 training set and 34 test set.**

| P* | Training data (leave-out-one CV) | | | | Test data (out-of-sample) | | | |
|---|---|---|---|---|---|---|---|---|
| | LD | SD | AD | GD | LD | SD | AD | GD |
| 50 | 38 | 38 | 38 | 38 | 33 | 33 | 32 | 33 |
| 100 | 38 | 38 | 38 | 38 | 32 | 33 | 31 | 33 |
| 500 | 38 | 38 | 38 | 38 | 31 | 33 | 30 | 33 |
| 1000 | 38 | 38 | 38 | 38 | 31 | 31 | 30 | 31 |
| 1500 | 38 | 38 | 38 | 38 | 31 | 31 | 29 | 31 |

**Table 3. ALL-AML data re-randomization(36/36 spliting) classification results of LD SD AD and GD(pca). Given are the correct classification percentage averaged over 100 rerandomizations.**

| P* | Training data (leave-out-one CV) | | | | Test data (out-of-sample) | | | |
|---|---|---|---|---|---|---|---|---|
| | LD | SD | AD | GD | LD | SD | AD | GD |
| 50 | 34.08 | 34.05 | 33.53 | 34.10 | 33.66 | 33.61 | 33.15 | 33.59 |
| 100 | 33.29 | 33.44 | 32.96 | 33.39 | 32.92 | 34.06 | 32.43 | 34.12 |
| 500 | 34.32 | 34.13 | 34.17 | 34.12 | 34.08 | 33.97 | 33.74 | 33.91 |
| 1000 | 32.95 | 33.19 | 33.02 | 33.17 | 32.50 | 32.78 | 32.62 | 32.81 |
| 1500 | 32.51 | 32.65 | 31.98 | 32.77 | 32.11 | 32.21 | 31.29 | 32.24 |

**Table 4. ALL-AML data re-randomization(36/36 spliting) classification results of LD SD AD and GD(pls). Given are the correct classification percentage averaged over 100 rerandomizations.**

| P* | Training data (leave-out-one CV) | | | | Test data (out-of-sample) | | | |
|---|---|---|---|---|---|---|---|---|
| | LD | SD | AD | GD | LD | SD | AD | GD |
| 50 | 36.00 | 36.00 | 36.00 | 36.00 | 34.72 | 35.11 | 34.83 | 35.10 |
| 100 | 35.88 | 36.00 | 35.82 | 36.00 | 34.30 | 34.71 | 34.32 | 34.68 |
| 500 | 36.00 | 36.00 | 36.00 | 36.00 | 34.73 | 34.68 | 34.28 | 34.73 |
| 1000 | 36.00 | 35.98 | 35.97 | 35.98 | 34.82 | 34.85 | 34.61 | 34.86 |
| 1500 | 36.00 | 36.00 | 35.97 | 36.00 | 34.71 | 34.77 | 34.50 | 34.78 |

[1] The number of samples correctly classified by LD is from Nguyen 2002b[10].

The re-randomization study is also carried out on the Leukemia data set to assess the stability of the results shown in Table 1, 2 and make a good compare among the four discrimination methods. The 72 samples are equally random divided in two parts: N1=36 training and N2=36 test samples. This analysis is also repeated for 100 re-randomizations for the purpose of utilizing the result of Nguyen et al. 2002b[10].

The results in the column of Table 4 are better than the corresponding value in Table 3. Careful study of each line of test data in Table 4 showed that the effectiveness of SD as well as GD is the same or even better than LD. AD give a worse performance than the other three discriminating methods. Analyzing the corresponding row of Table 3 and Table 4, we can get that PLS definitely give a perfect result in dimension reduction, whereas, PCA is slightly worse.

**Table 5. Colon data classification results of LD SD AD and GD. Given are the number of samples correctly classified out of the 62 samples(40 tumor 22 normal).**

| P* | LD | | SD | | AD | | GD | |
|---|---|---|---|---|---|---|---|---|
| | PC | PLS | PC | PLS | PC | PLS | PC | PLS |
| 50 | 54 | 58 | 55 | 59 | 53 | 58 | 55 | 59 |
| 100 | 53 | 58 | 54 | 58 | 53 | 56 | 54 | 58 |
| 500 | 53 | 56 | 53 | 56 | 52 | 54 | 53 | 56 |
| 1000 | 52 | 57 | 54 | 56 | 51 | 55 | 54 | 56 |

### *B Colon data*

Alon et al.(1999)[1] used Affymetrix oligonucleotide arrays to monitor expression of over 6500 human genes with samples of 40 tumor and 22 normal colon tissues.

Alon et al. clustered the 62 samples into two clusters. One cluster consisterd of 35 tumor and 3 normal samples(n8, n12, n34). The second cluster contained 19 normal and 5 tumor tissues(T2, T30, T33, T36, T37). Furey et al.[5] did leave-out-one CV prediction of the 62 samples using SVM and six tissues are misclassified, namely(T30, T33, T36) and (n8, n34, n36)[10]. Note form Table 5, SD and GD give the best performacne in classification. The two samples T36 and n36 are correctly classified into the relevant category (with conditional class probabilities 0.89 and 0.12). Because SD and GD are not based on the same proposed hypothesis as described in T.S Furey et al.(2000). However, the SD and GD misclassify the same samples (T2, T11, T33 ) as SVM and cluster. The 'Prediction Strength(PS)'(Golub et al.1999) of these samples are all low(PS<0.30).

Three tumor samples are classified to normal class by all four discrimination methods. SD and GD do not make mistake as LD and AD on sample n36. Besides, the SD and GD give a similar result considering the conditional probability, and the result is slightly better than the other two methods in discrimination. From Figure 3, we can also get that SD and GD conditional probability for tumor samples are all closer to 1 and the value for normal samples are all closer to 0 than the other two. The misclassified samples are marked on the horizonal axis. Almost all the small circle and triangle are above the blue star in the tumor part, and below the blue star in the normal part.

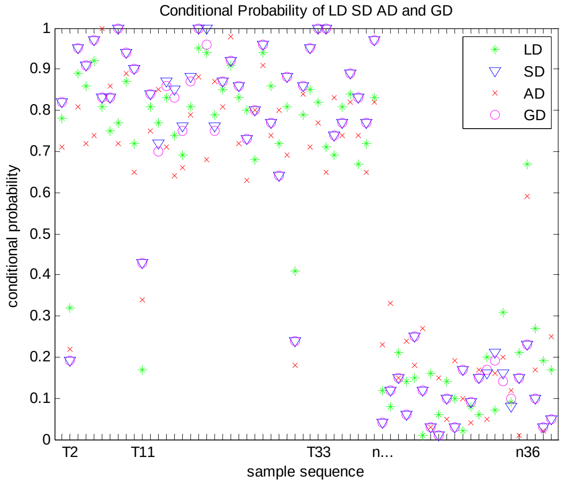

**Fig. 3 Conditional Probability of LD SD AD and GD based on p*=50. Misclassified samples are labeled on horizontal axis.**

Why PCA fails to make a prediction for response variable. The answer can be found in Table 6. Note that four components extracted by PLS base on 50 pre-selected genes can explain 95.00% and 91.16% of predictor and response variability respectively while the variability explained by the components of PCA are 95.46% and 76.45%. PLS and PCA give a similar performance in explaining the variation of predictor.

**Table 6. Variability explained by PLS components and PCs. The number of components base on 50 pre-selected genes is K.**

| K | Predictor | | Response | |
|---|---|---|---|---|
| | Proportion | Cumulative Proportion | Proportion | Cumulative Proportion |
| | | PLS | | |
| 1 | 81.43 | 81.43 | 83.22 | 83.22 |
| 2 | 7.42 | 88.85 | 3.89 | 86.11 |
| 3 | 5.39 | 94.24 | 2.04 | 88.15 |
| 4 | 0.76 | 95.00 | 2.01 | 91.16 |
| | | PC | | |
| 1 | 81.52 | 81.52 | 73.12 | 73.12 |
| 2 | 11.53 | 93.05 | 1.05 | 74.17 |
| 3 | 1.43 | 94.48 | 0.03 | 74.40 |
| 4 | 0.98 | 95.46 | 2.05 | 76.45 |

But PCA interprete only 76.45% of response variability which is much poorer than PLS. The second component of PCA accounts for predictor variability is 11.53% but it accounts for only 1.05% of total response variability. The theoretical reason can be found in section 2.4. Due to these analysis, PCA fail to make a precise prediction as well as PLS is not surprising.

**Table 7. Lymphoma data classification results of LD SD AD and GD. Given are the number of samples correctly classified out of the 74 samples.**

| P* | LD | | SD | | AD | | GD | |
|---|---|---|---|---|---|---|---|---|
| | PC | PLS | PC | PLS | PC | PLS | PC | PLS |
| 50 | 72 | 73 | 72 | 72 | 72 | 71 | 72 | 72 |
| 100 | 72 | 71 | 72 | 71 | 72 | 71 | 72 | 71 |
| 500 | 72 | 71 | 72 | 71 | 71 | 70 | 72 | 71 |
| 1000 | 72 | 70 | 72 | 71 | 71 | 70 | 72 | 71 |

**Table 8. Lymphoma data re-randomization(37/37 spliting ) classification results of LD SD AD and GD(pca). Given are the correct classification percentage averaged over 100 rerandomizations.**

| P* | Training data<br>(leave-out-one CV) | | | | Test data<br>(out-of-sample) | | | |
|---|---|---|---|---|---|---|---|---|
| | LD | SD | AD | GD | LD | SD | AD | GD |
| 50 | 36.3 1 | 36.02 | 35.86 | 36.02 | 36.12 | 35.85 | 35.15 | 35.85 |
| 100 | 36.36 | 36.41 | 36.18 | 36.41 | 36.26 | 36.23 | 35.43 | 36.21 |
| 500 | 35.23 | 35.39 | 35.02 | 35.42 | 35.31 | 35.42 | 34.74 | 35.42 |
| 1000 | 34.98 | 35.11 | 34.90 | 35.11 | 33.94 | 34.01 | 33.42 | 34.01 |

**Table 9. Lymphoma data re-randomization(37/37 spliting) classification results of LD SD AD and GD(pls). Given are the correct classification percentage averaged over 100 rerandomizations.**

| P* | Training data<br>(leave-out-one CV) | | | | Test data<br>(out-of-sample) | | | |
|---|---|---|---|---|---|---|---|---|
| | LD | SD | AD | GD | LD | SD | AD | GD |
| 50 | 36.92 | 36.91 | 36.48 | 36.91 | 35.57 | 35.63 | 34.83 | 35.63 |
| 100 | 36.91 | 36.95 | 36.51 | 36.94 | 35.84 | 36.02 | 34.92 | 36.02 |
| 500 | 36.86 | 36.85 | 36.22 | 36.85 | 35.76 | 35.93 | 34.88 | 35.93 |
| 1000 | 36.84 | 36.88 | 36.08 | 36.88 | 35.61 | 35.49 | 34.61 | 35.49 |

### *C Lymphoma data*

The data set presented by Alizadeh et al. (2000) comprises the expression levels of 4151 genes from 3 different classes: Diffuse Large B-Cell lymphoma (DLBCLL; n1=45), B-Cell Lymphocytic Leukemia (BCLL; n2=29) Follicular(FL; n3=9). In order to test our binary classification method and make a compare with logistic discrimination[10]. We chose the first two class, namely DLBCLL and BCLL.

We got the value of $\frac{S_{PRESS,4}}{S_{SS,3}}$ using SIMCA-P, it is 0.8579 < 0.9025. So the 4th LV should be considered in the subsequent process. Lastly, using LOOCV, each sample is predicted to be DLBCLL or BCLL base on 5 gene components constructed from p*=50,100,500,1000 genes. The two samples(#33 and #51) are consistently misclassified. Table 7 and Table8 provide the detailed consequence of LD SD AD and GD. All these methods can be used as means to predict the samples, because they achieve perfect result in the procedure.

As with the analysis of acute leukemia data, we turned to re-randomization to access the stability of the classification performance. Table 8 and Table 9 are the results using PCA and PLS to reduce dimention respectively. Compare with Table 7 in Nguyen 2002b[10], the accuracy rate of classification base on 5 components is impoved appreciably. All the classification methods displayed rather good ability in discrimination. Especially the LD SD GD, the accuracy rate for training data using LOOCV is nearly 100%. The accuracy is best for p*=50, which denoted that gene select process is of great importance for the entire experiment. Because this process filtered the irrelevant gene which may be interference to the prediction procedure.

### *D Gastric cancer data*

This gastric cancer data contains the probes of over 19900 human genes with samples of diffused ones and intestinal ones(20 diffused and 20 intestinal). All the samples which are microarray data base on specimen of gastric carcinoma and clinical data are from China. They are all supplied by Beijing Cancer Hospital. In this experiment, we employ similar procedure with the above practice to deal with the original data set.

The EP is vary for different preliminary selected gene number. We use the mean of EP which is calculated from each row of Table 2 for the purpose of ruling out of chance. In these two Figures(Figure 4 and Figure 5), a vertical line is drawn at number of LVs or PCs equals 4. The fifth LV can just interpret 1.24% of response variability, so we use top 4 LVs in the classification process because the 5th LV is of little use(less than 2%). This method is proved to be as effective as

$\frac{S_{PRESS,i}}{S_{SS,i-1}}$ in the following experiment. The advantage of PLSDR in interpreting response variablity is much more obvious than PCA. Correspondingly, the classification accuracy is much better for PLSDR than PCA.

In order to further certify the effectiveness of LD SD AD and GD. We compare the four models using this data set. The result is exhibited in Table 10. The difference in explaining the response variability of PCA and PLS makes an interpretation for the gap of classifiction results. The number of misclassified samples is the same for LD SD and GD base on P* = 50, however the samples are disparate. (#11, #18)for LD,(#11, #24)for SD(#11,#18) and GD(#11, #24). AD gives the worst performance no matter how many genes are considered. The re-randomization is no longer provided, because the number of sample is too small.

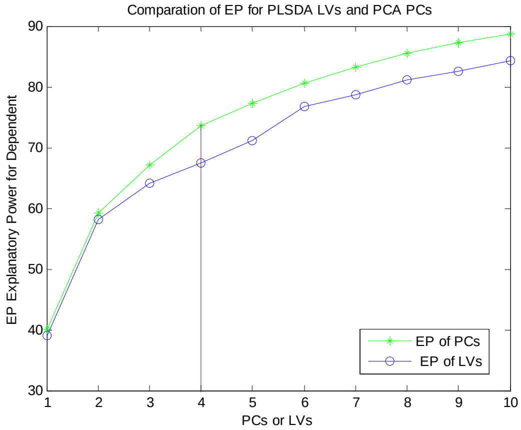

**Fig. 4 Explanatory Power for Predictor Variable of PLSDR and PCA**

Note from Table 5 that with the decrease of gene number, the classification accuracy is increasing. It proves that gene selection is of a significant role in this experiment.

## IV. ANALYSE AND CONCLUTION

Why SD and GD are competitive or even better than LD, AD gives the worst performance. Here we give the compare of these function curves in Figure 6. From it, we can see that there is no distinguishment between them in the outline of curve.

However, we should pay attention to the sharp boosting of sigmoid function. The sharpness gives the opportunity of less iteration in the algorithm of determining the coefficient. As a result, it achieves better results in saving time of step-by-step process. The Logistic function curve overlaps with Gaussian Error function. So they give a similar result in classification procedure.

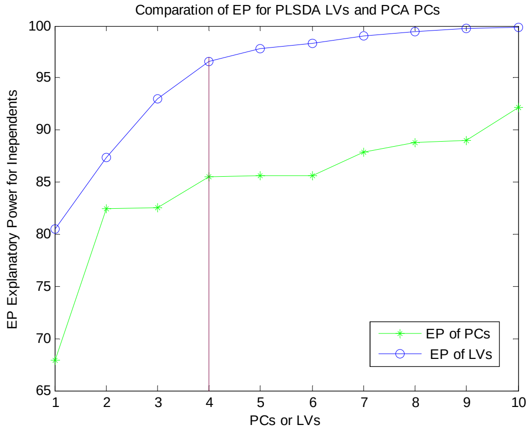

**Fig. 5 Explanatory Power for Response Variable of PLSDR and PCA**

**Table 10. Gastric cancer data classification results of LD SD AD and GD. Given are the numbers of samples correctly classified out of the 40 samples.**

| P* | LD | | SD | | AD | | GD | |
|---|---|---|---|---|---|---|---|---|
| | PC | PLS | PC | PLS | PC | PLS | PC | PLS |
| 50 | 35 | 38 | 35 | 38 | 35 | 38 | 35 | 38 |
| 100 | 35 | 38 | 35 | 38 | 35 | 37 | 35 | 38 |
| 500 | 34 | 38 | 34 | 38 | 33 | 37 | 34 | 38 |
| 1000 | 34 | 37 | 35 | 37 | 33 | 36 | 35 | 37 |
| 2000 | 34 | 37 | 34 | 37 | 33 | 36 | 34 | 37 |

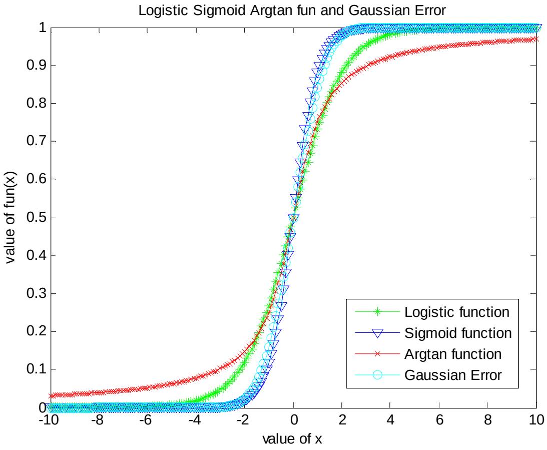

**Fig. 6 Comparison of Logistic function and Transformed Sigmoid function**

The value domain of red curve is a smaller range than the other three functions in Figure 6, which is the reason of poor discrimination ability of AD. In Figure 3, the effectiveness of SD LD AD and GD has been already obtained from one side. And the result is agreed with which concluded from the figure above.

This article gives comparative study of some new models to classify samples base on microarray dataset. The tool presented here can achieve better or competitive result compare with PLS-LD (PLS-Logistic Discriminate).

Our experimental codes which programmed by MATLAB language are all carried out on a PC workstation with Intel Core DuoT2060 (1.6G) and 4GB RAM.

**Acknowledgement** Thanks Professor Lu Youyong from Beijing Cancer Hospital for providing gastric cancer data to us. This research is Supported by Beijing Natural Science Foundation (Grant No.4092021) and Scientific Plan of Beijing Municipal Commission of Education (JC002011200903).

## V. REFERANCE

[1] Alon, U. Barkai,N. Notterman,D.A. Gish,K. Ybar,S. Mack,D. and Levine,A.J. Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. Proc. Natl Acad. Sci. USA,1999;96,6745-6750.

[2] Boulesteix,A.-L. PLS dimension reduction for classification of microarray data. Statistical Applications in Genetics and Molecular Biology. 2004;3(1),33.

[3] Boulesteix A L and Strimmer K, Partial least squares: a versatile tool for the analysis of high-dimensional genomic data[J]. BRIFINGS IN BIOINFORMATICS. VOL 8.NO 1. 32~44, Jan, 2007.

[4] D ai,J.J. Lieu,L. and Rocke,D. Dimension reduction for classification with gene expression data. Statistical Applications in Genetics and Molecular Biology 2006;5(1), Article 6.

[5] Terrence S. Furey, Nello Cristianini , Nigel Duffy, David W. Bednarski , Michèl Schummer and David Haussler. Support vector machine classification and validation of cancer tissue samples using microarray expression data. Bioinformatics 2000; v16, n10, 906-914.

[6] J.G. Liao Khew-Voon Chin. Logistic regression for disease classification using microarray data: model selection in a large p and small n case.Bioinformatics 2007; v23, n15,1945-1951.

[7] Jian-Nan Zhang, A-Li luo, Yong-Heng Zhao, Automated estimation of stellar fundamental parameters from low resolution spectra: the PLS method. Research in Astonomy and Astrophysics.June 2009;V9,n6,712-724.

[8] Menezes,J.c. Felicio,C.C. Bras,L.P. Lopes,J.A. and Cabrita,L. Comparison of PLS algorithms in gasoline and gas oil parameter monitoring with MIR and NIR. Chemometrics and Intelligent Laboratory Systems. 28 July 2005;V78,n1-2,74-80.

[9] Nguyen,D.V. and Rocke,D.M., Tumor classification via partial least squares using microaray gene expession data. Bioinformatics, 2002(a),18(1),30-50.

[10] Nguyen,D.V. and Rocke,D.M, Multi-class cancer classification via partial least squares with gene expression profiles. Bioinformatics, 2002(b),18(9),1216-1226.

[11] Wold,H, Estimation of principal components and related models by iterative least squares. Krishnaiah PR (ed). Multivariate Analysis. New York: Academic Press, 1966; 391-420.

[12] W old,H, Nonlinear Iterative Partial Least Squares(NIPALS) modeling: some current development. Krishnaiah PR (ed). Multivariate Analysis. New York: Academic Press, 1973; 383-407.

[13] Wold,H, Path models with latent varibles: the NIPALS approach. Blalock HM(ed). Quantitative Sociology: Internatioanl Perpectives on Mathematical and Statistical Model Building. New York: Academic Press, 1975.

[14] Xue-Qiang ZengGuo-ZhengLi Geng-FengWu JackY.Yang M.Q. Irrelevant gene elimination for Partial Least Squares based Dimension Reduction by using feature probes. International Journal of Data Mining and Bioinformatics. 2009;v3,n1,85-103.