Available online at www.sciencedirect.com
ScienceDirect
Progress in Natural Science 19 (2009) 267–272
Progress in Natural Science
www.elsevier.com/locate/pnsc

Short communication

A novel methodology for finding the regulation on gene expression data

Wei Liu [a,*], Bo Wang [a], Jarka Glassey [b], Elaine Martin [b], Jian Zhao [a]

[a] *The Key Laboratory of Biomedical Information Engineering of Ministry of Education, School of Life Science and Technology, Xi'an Jiaotong University, Xi'an 710049, China*

[b] *School of Chemical Engineering and Advanced Materials, University of Newcastle upon Tyne, Merz Court, Newcastle upon Tyne NE1 7RU, UK*

Received 19 April 2008; received in revised form 14 July 2008; accepted 16 July 2008

Abstract

DNA microarray technology is a high throughput and parallel technique for genomic investigation due to its advantages of simultaneously surveying features of large scales complex data in biology. This paper aims to find feature subset to build the classifier for gene expression data analysis. At first, *K*-means clustering algorithm was carried out on the dataset of yeast cell cycle. Based on Rand calculation, a statistical method was used to pick out the data points (genes) for classifier design. Meanwhile, the principal component analysis was applied to help to construct the classifier. For the validation of classifier built and prediction of a target subset of genes, discriminant analysis in terms of partial least square regression and artificial neural network were also performed.

Keywords: Classifier design; Discriminant analysis; Gene expression data; Rand calculation

1. Introduction

Microarray technology is considered to be the latest breakthrough in experimental molecular biology, which allows monitoring gene expression of a large amount of genes in parallel with valuable data. The data contain many variations and co-regulated networks, as well as the non-linear correlations. How to analyze and handle such data is currently one of the major issues in the application of this technique. Usually, there are two types of classifications. One is the clustering analysis, the other is the discriminant analysis. They are all involved in the machine learning technology, in which the most popular computational models, such as neural networks and Bayesian networks, are used [1].

Generally, the clustering analysis is applied in finding genes with similar expression patterns under various conditions for pattern recognition. These genes may participate in the same signal pathway or may be functionally co-regulated. For interpreting the behavior of a large amount of data from cDNA microarray experiments, the clustering analysis plays an important role in revealing the underlying relationships in biological experiments. It is sometimes used as a preprocessing step in inferring regulatory networks [2]. As a typical technique to capture structures of datasets, *K*-means clustering aims to group objects (genes or samples) with similar behavior. It can also be described as a method of dimensionality reduction of systems due to the clusters classified.

On the other hand, the discriminant analysis has been applied to build the classifier to predict an interesting gene to which group it belongs. For example, the partial least square (PLS) has been used to investigate the leukemia microarray dataset [3]; and the threshold has also been introduced in rescaling data. Compared with PLS algorithm, neural networks (NNs) are a common non-linear method for discriminant analysis.

However, there are inevitably some shortcomings even for some typical approaches. The first main issue lies on

* Corresponding author. Tel.: +86 15029487680; fax: +86 29 82660554.
E-mail address: wliu@mail.xjtu.edu.cn (W. Liu).

1002-0071/$ - see front matter © 2008 National Natural Science Foundation of China and Chinese Academy of Sciences. Published by Elsevier Limited and Science in China Press. All rights reserved.
doi:10.1016/j.pnsc.2008.07.003

the lack of systematic research in ways of assessing the measuring quality and comparing data from various technology platforms [4]. Another issue is that the traditional 'gene by gene' method is not sufficient to understand gene regulatory networks consisting of thousands of genes. So a big challenge of the gene expression and data analysis is involved in the modeling of gene regulatory networks [5,6].

In this study, a so-called Rand calculation has been used to find a classifier for gene expression patterns underlying multivariate dataset. Afterwards, the approaches to supervising the classifications, such as PLS and artificial NN (ANN), were used in model validation and prediction for classifier. It means that an expert can both determine what classes an object may be categorized into and provide a set of sample objects with known classes.

2. Methods

For data preparation, the auto-scaling for the dataset of the budding yeast *S. cerevisiae* (383 × 17) with mitotic cell cycle [7] was performed at first, which scaled data to zero mean and unit variance. Then principal component analysis (PCA) was applied in outlier detection and data dimensional reduction at first. The criterion to decide the number of principal components (PCs) is that the number of them can capture above 90% of the variance. PCs are defined as the projected variable which is uncorrelated with the earlier PCs and has maximal variance, with arbitrary sign. PC1 is the first principal component with the projection of the largest variance. PC2 is the second principal component and PC3 is the third principal component. The p PCs have decreasing variances.

Afterwards, data analyses were performed as follows:

(1) Performing K-means clustering on dataset through GeneSpring^{TM} 4.2 software.
(2) Comparing results of clustering methods by Rand calculation.
(3) Classifier designing.
(4) Model validation and prediction by the discriminant analysis of PLS and ANN.

2.1. K-means clustering

K-means clustering can actually be used in any similarity measure, where Euclidean properties of the vector space are essential. The algorithm for K-means clustering can be expressed by minimizing the sum-of-squares criterion as follows:

$$J = \sum_{j=1}^{K} \sum_{n \in S_j} | \boldsymbol{x}_n - \boldsymbol{\mu}_j|^2, \tag{1}$$

where $\boldsymbol{x}_n$ is a vector representing the nth data point, and $\boldsymbol{\mu}_j$ is the geometric centroid of the data points in S_j. The algorithm does not achieve a global minimum of J over the assignments. For the illustration in biological sense, genes are divided into a user-defined number K of equal-sized groups; K-means algorithm examines each vector in the dataset and assigns it to one of the clusters depending on the minimum distance. It then calculates centroids at the average location of each group of genes. For each iteration, genes are reassigned to the group with the closest centroid. After all the genes have been reassigned, the centroids are recalculated and the process is repeated until no genes move between clusters from one iteration to the next.

2.2. Rand calculation and adjusted Rand index

To compare the results of any two different clustering methods, Rand index (RI) as an objective criterion has been used to measure the similarity of the clustering results. It is not a simple count of the points misclassified, but a measure following three basic assumptions. Firstly, a clustering is discrete so that every point is definitely assigned to a specific cluster. Secondly, the number of clusters defined by those points which they do not contain is just the same as that they do contain. Thirdly, all points are of equal importance in the determination of the clustering. Based on the assumptions, the meaning of a basic unit of comparison between two clusters is how many pairs of points are clustered.

A measure of an agreement between two partitions is introduced according to clustering results comparison against external criteria. It was assumed that each gene is assigned to only one class in the external criterion and to only one cluster [8]. Considering a set of n objects, the data matrix $\boldsymbol{S} = \{\boldsymbol{O}_1, \boldsymbol{O}_2, ..., \boldsymbol{O}_n\}$ is partitioned into two sets of clusters $\boldsymbol{U} = \{u_1, u_2, ..., u_r\}$ and $\boldsymbol{V} = \{v_1, v_2, ..., v_c\}$, where $\boldsymbol{U}$ and $\boldsymbol{V}$ represent biological grouping methods of five phase criterion and K-means Spearman correlation, respectively.

Let a be the number of pairs of objects placed in the same cluster in $\boldsymbol{U}$ and in the same cluster in $\boldsymbol{V}$, b be the number of pairs of objects in the same cluster in $\boldsymbol{U}$ but not in the same cluster in $\boldsymbol{V}$, c be the number of pairs of objects in the same cluster in $\boldsymbol{V}$ but not in the same cluster in $\boldsymbol{U}$, and d be the number of pairs of objects in different clusters in both partitions. Therefore, a and d can be taken as agreements, and b and c as disagreements. The RI is given by

$$RI = \frac{a+b}{a+b+c+d} = \left[\binom{n}{2}\left[0.5\left\{\sum_i\left(\sum_j n_{ij}\right)^2 + \sum_j\left(\sum_i n_{ij}\right)\right\}^2 - \sum\sum n_{ij}^2\right]\right] \Big/ \binom{n}{2}, \tag{2}$$

where n_{ij} is the number of objects simultaneously in the ith cluster in $\boldsymbol{U}$ and the jth cluster in $\boldsymbol{V}$, and $\binom{n}{2}$ is the binomial coefficient, which gives the number of distinct pairs found in a set of n objects.

A higher *RI* indicates a higher degree of the similarity of two partitions involved in, and vice versa. Hubert and Arabie [9] have corrected the *RI* for chance, making it a more appropriate descriptive measure by the adjusted Rand index (*ARI*), which includes a chance element. The *ARI* has the general form of

$$\frac{\text{Index} - \text{expected index}}{\text{Maximum index} - \text{expected index}}. \quad (3)$$

Assuming a maximum *RI* of 1, the *ARI* is calculated as follows:

$$\frac{\sum_{ij}\binom{n_{ij}}{2} - \left[\sum_i \binom{n_{i\cdot}}{2}\sum_j \binom{n_{\cdot j}}{2}\right] \Big/ \binom{n}{2}}{0.5\left[\sum_i \binom{n_{i\cdot}}{2} + \sum_j \binom{n_{\cdot j}}{2}\right] - \left[\sum_i \binom{n_{i\cdot}}{2}\sum_j \binom{n_{\cdot j}}{2}\right] \Big/ \binom{n}{2}}, \quad (4)$$

where $n_{i\cdot} = n_{i1} + n_{i2} + \cdots + n_{ic}$ and $n_{\cdot j} = n_{1j} + n_{2j} + \cdots + n_{rj}$.

This adjusted Rand index ranges from 0 (when the index equals its expected value) to 1 (when the two partitions are identical). Additionally, the *RI* is much higher than the *ARI*. It is obvious that the *ARI* can take a wider range of values, which brings the rise of sensitivity of the index. And then, a higher *ARI* will lead to a higher quality clustering result.

2.3. Discriminant analysis

Discriminant analysis called discriminant function analysis is also a technique for classifying a set of observations into predefined classes. The analysis is based on a set of variables known as predictors or input variables, and has been used to predict classifications of cases.

When a model has been constructed and the discriminant functions have been derived, it could be predicted to which group a particular case (a target gene) belongs. The classification scores for each case for each group could be computed by means of each function as follows:

$$S_i = c_i + w_{i1}x_1 + w_{i2}x_2 + \cdots + w_{im}x_m. \quad (5)$$

In this formula, the subscript i denotes the respective group; the subscripts 1, 2, ..., m denote the m variables; c_i is a constant for the ith group, w_{ij} is the weight for the jth variable in the computation of the classification score for the ith group; x_j is the observed value for the respective case for the jth variable, and S_i is the resultant classification score.

In order to perform the discriminant analysis, PLS and ANN were used for the investigation detailed below.

3. Classifier construction

3.1. Properties of data

Originally, the dataset of genes involved in cell cycle was found for 416 of the 6220 monitored transcripts of the budding yeast *S. cerevisiae* by Koch and Nasmyth in 1994 [10]. The time course of yeast cell cycle was divided into five phases. However, there are additionally 33 of the 416 identified genes peaks twice in different phases of cell cycle [11]. Therefore, the actual number of genes for cell cycle of yeast is 383. In this study, the actual number of genes displayed is 379, denoted by 'yeastdata'. Several genes, such as YDR179c and YAL034w in phase 3, YCL012w in phase 4 and YML033w in phase 5, were found lost. It may be caused by the selection of the initial points of genes while analyzing or genes discarded with no data for half of the conditions shown in *K*-means clustering instruction by using the GeneSpring™ 4.2 software. However, the effect of such a small number of genes could be neglected in statistical analysis. As the number of genes in five clusters is changed randomly by *K*-means clustering, it is requested by the experimental design to take samples at 10 random starting points, which could help to get rid of the dead centers.

3.2. Agreement and normalized yeastdata

K-means clustering based on Standard/Pearson/Spearman correlation was performed 10 times each at random initial starting point in order to get 10 random samples separately. After each Rand calculation, the *ARIs* were obtained. The highest average values for Standard correlation, Pearson correlation and Spearman correlation are 0.375, 0.368 and 0.325, respectively; the lowest deviation values for Standard correlation, Pearson correlation and Spearman correlation are 0.010, 0.040 and 0.026, respectively. In comparison with the *ARIs* on the other two kinds of correlations, the *ARI* depending on Standard correlation has the highest average value and the lowest deviation on the average, which leads to the higher quality clustering result. Thereby the Standard correlation was chosen for *K*-means clustering here as shown in Table 1.

From Table 1, the *RI* and *ARI* are calculated as 0.815 and 0.423, respectively.

On each sample basis, a new subset of sample was obtained by picking the genes corresponding to the maximum or two biggest numbers n_{ij} (if their values are closer in the grids) along the rows of Table 1, which shows the good agreement between two partitions of 5-phase criterion and *K*-means clustering based on standard correlation. The bigger the number is in a grid, the higher the statistical

Table 1
Agreement between two partitions of five phase criterion and *K*-means clustering.

Phase	n_{ij}					Total
1	50	5	12	0	0	67
2	20	104	0	1	8	133
3	3	21	0	26	24	74
4	1	0	11	38	1	51
5	1	0	51	2	0	54
Total	75	130	74	67	33	379

probability of the matching pairs of partitions is. Therefore, this is a new way to search for the feature subset of gene expression data. The feature subset with the highest evaluation may be chosen as the final set to build a classifier [12].

Thus, in order to build a classifier for a target subset of genes, ten samples were obtained randomly by *K*-means clustering at first. The numbers of genes corresponding to five phases were obtained according to the method stated above. This action is repeated on 10 samples so as to find the common genes and to form the stable structure of gene expression data. As a result, the data matrix reduced to 139×17, named 'cluster'. Whatever the subset of 'cluster' looks as a classifier, it needs verification.

3.3. Dataset for discriminant analysis

As the subset of Yeast cell cycle, 'cluster' (139×17), was decided as the input variables for discriminant analysis, the output variable (column) denoted by group numbers, such as 1, 2, 3, 4, 5, should be set to the data matrix as an output vector in column 1 of Table 1. Then the datasets for training and validation were constructed on the basis of the 'cluster' matrix, in which 70% of the data of genes of each phase in the 5-phase criterion classification (along the rows) were combined to form the training dataset (95×17). One half of the remaining 30% of the data of genes of each phase were combined to form the validation dataset (22×17).

4. Results and discussion

4.1. K-means clustering results

As shown in Fig. 1, the software GeneSpring was used to compute *K*-means clustering. One sample of *K*-means clustering on the basis of standard correlation was produced by classifying genes into five classes. Next, PCA was used to realize the classifier.

4.2. Pattern recognition by PCA and classifier built

(i) *Pattern recognition:* PCA model developed on 'cluster' (139×17) indicated that the first three PCs capturing 92.30% of the total variance were selected for analysis.

Fig. 2 shows there are some patterns (clusters) displayed by PCA score plots, which can also be calibrated by several straight lines starting from one point. At this stage, the expression profile of the reduced and normalized yeast cell cycle data is very close to a typical linear pattern.

(ii) *Classifier built:* By visualization, the graphs of 'cluster' are nearly linear. Thereby, the linear regression is used for mathematical description of the classifiers of Yeast cell cycle. The typical multi-linear regression model is as follows:

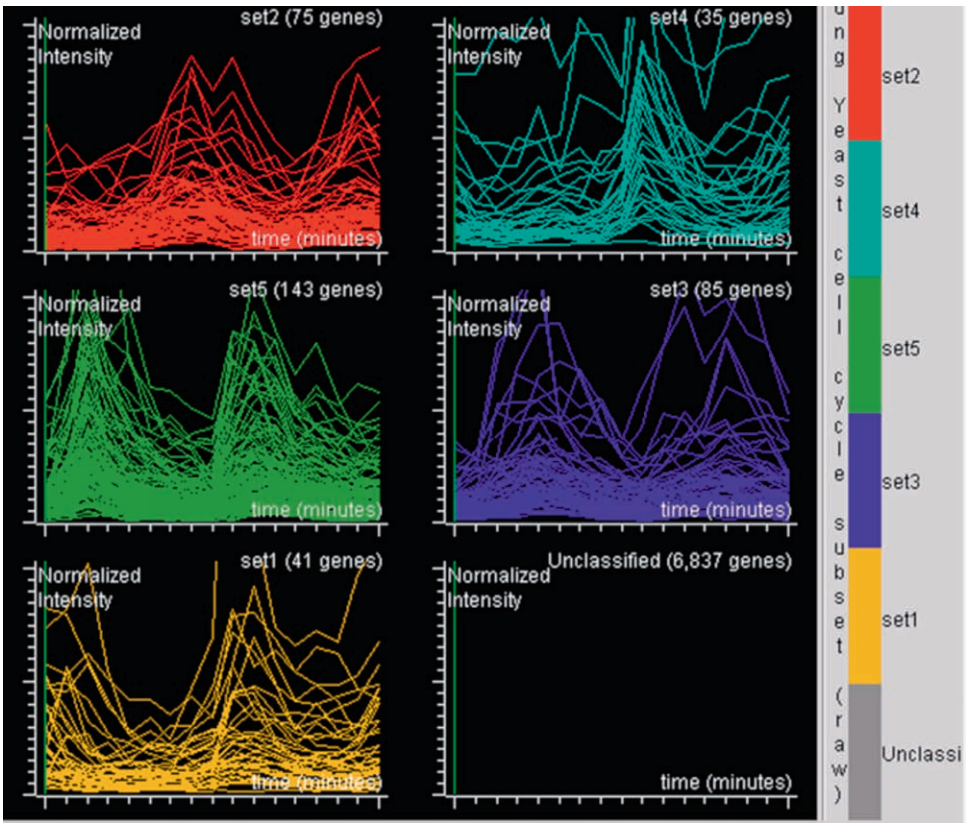

Fig. 1. *K*-means clustering graphs by GeneSpring™.

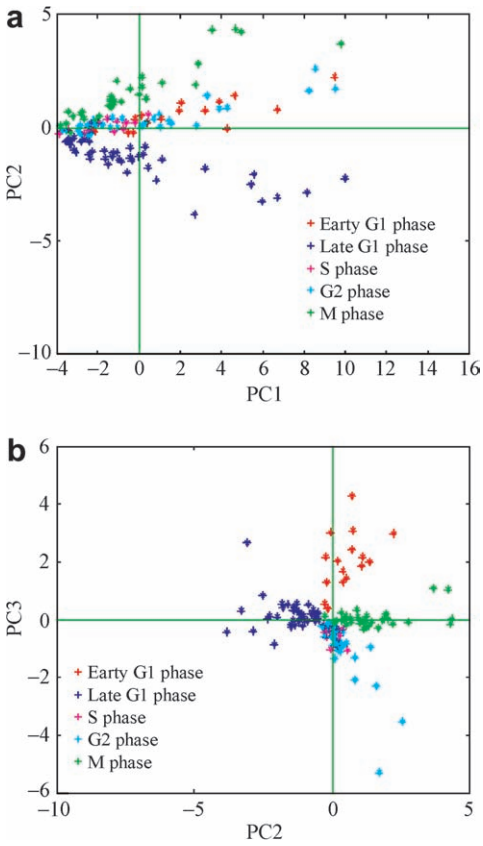

Fig. 2. Score plots of (a) PC1 versus PC2 and (b) PC2 versus PC3 on 'cluster' data.

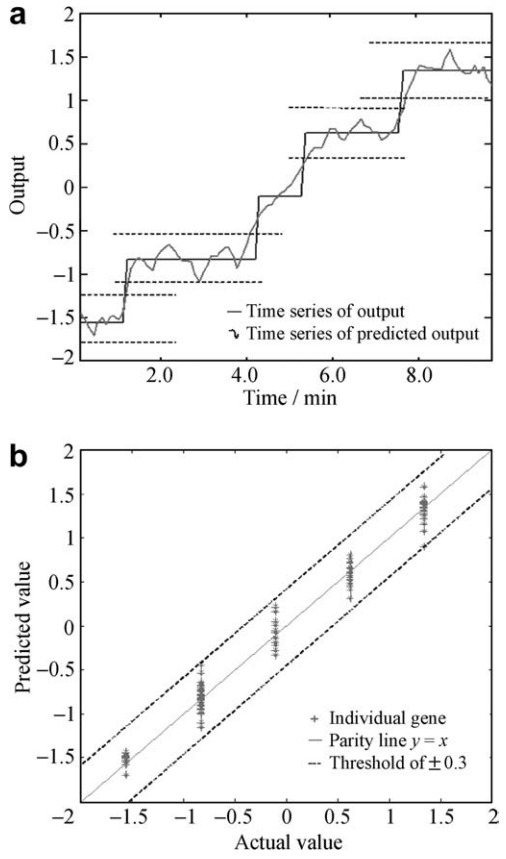

Fig. 3. (a) Training plot (predicted output time series); (b) validation plot (actual value versus predicted value). The RMS of validation is 0.136.

$$\boldsymbol{Y} = \boldsymbol{\beta X} + \boldsymbol{E}, \tag{6}$$

where $\boldsymbol{\beta}$ is the matrix of regression coefficients, $\boldsymbol{X}$ is the matrix of independent variables, and $\boldsymbol{E}$ is the matrix of error. Because the score plots for 5-phase criterion show the trends of linear regression, the model can be expressed as

$$\boldsymbol{Y} = \beta_0 + \beta_1 \boldsymbol{X}_1 + \beta_2 \boldsymbol{X}_2 + \cdots + \beta_5 \boldsymbol{X}_5 + \boldsymbol{E}, \tag{7}$$

where β_0, β_1, β_2, …, β_5 are the regression coefficients. X_1, X_2, …, X_5 are separate basic variables, with $\boldsymbol{Y}$ as the output and $\boldsymbol{E}$ as an error. The results were calibrated by the below mentioned discriminant analysis.

4.3. Discriminant analysis results

4.3.1. PLS modeling

(i) *Model training:* PLS regression using 'Prescreen' software gives the PLS model on 'cluster' as shown in Fig. 3(a).

As shown in Fig. 3(a), the threshold in model training plot refers to an interval with lower control limit (LCL) and upper control limit (UCL), where the distribution of predicted values can be accepted to a certain classification. In this case, the predicted output time series shows that how the predicted output fluctuates around each of the horizontal classification lines is decided by the output variables. It can also be observed that much higher proportion of predicted output is within the threshold of ±0.30. If the threshold is defined too narrow, the limitation is too strict to put into effect, because a large proportion of predicted values will be considered as outliers. On the contrary, if the threshold is defined too wider, e.g. ±0.50, some points may belong to both adjacent groups, which will lose the significance of classification.

(ii) *Model validation:* It was carried out by using 'modlgui' in MATLAB. And the validation plot is produced in Fig. 3(b). The actual values versus the predicted values along the parity line were grouped into five classifications. In addition, the percentage of the genes within the threshold of ±0.30 can be figured out by subtracting the outliers. There is one observation out of the threshold as shown in Fig. 3(b). Since the root mean square error (RMS) is 0.136, the validation value of mean square error (MSE) is $\mathrm{MSE}_{\mathrm{valid1}} = \mathrm{RMS}^2 = 0.018$.

4.3.2. Neural network modeling

(i) *Model training:* The following training plot as shown in Fig. 4(a) indicates that the prediction output fluctuates around the classification horizontal line. This is also a predicted output time series. The threshold was set to be ±0.30 as described above, and the percentage of genes falling in such a range is still very high in spite of two peaks of outliers detected. This prediction model is closer to the real condition, because the raw data were just standardized by

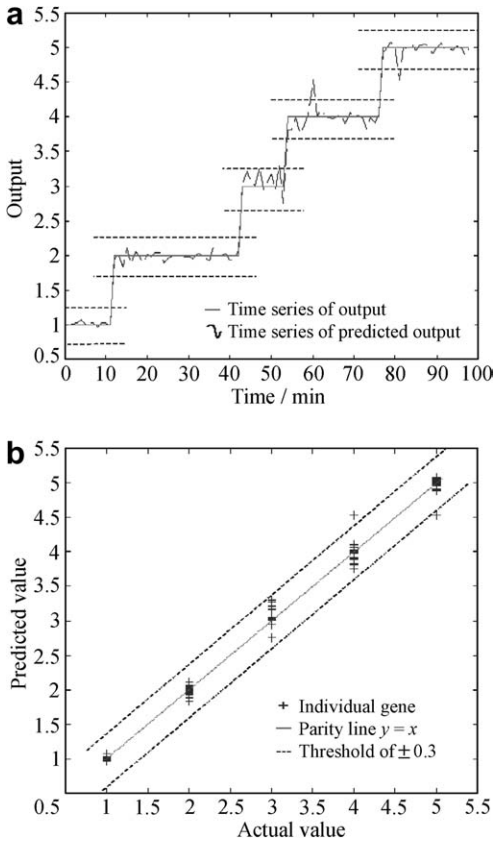

Fig. 4. (a) Training plot (predicted output time series); (b) validation plot (actual versus predicted output). The MSE_{valid2} is 0.012.

scaling and the output keeps in five classes and does not experience parallel movement as that occurred in the PLS modeling plot.

(ii) *Model validation:* The following validation plot shown in Fig. 4(b) gives the evidence of the actual values versus the predicted values on the whole along the parity line. The proportion of genes within the threshold of ±0.30 nears 100%. For this artificial neural network training model, the calibrated mean square error $MSE_{valid2} = 0.012$, which is closer to $MSE_{valid1} = 0.018$ of the PLS training model.

4.4. Comparison of the two models

The results from PLS and ANN models shown in Figs. 3 and 4 can be more easily compared through MSE. The MSE_{valid} for PLS and ANN are 0.018 and 0.012, respectively. Their validation plots are along the parity lines. It indicated that the two models are similar, although there are some differences.

In ANN modeling, as raw data can be input directly, and scaled by default settings, no filtering is needed. Additionally, the model results are close to the actual situation without important information losing. On the contrary, in PLS modeling, raw data should be filtered first. If not, large noise will disturb model training.

5. Conclusions

A novel methodology on building the classifier was found and discussed successively by a case study in this paper. Based on *K*-means clustering and Rand calculation, the classifier for yeast cell cycle of budding yeast *S. cerevisiae* was constructed through multi-linear regression and was validated by discriminant analysis.

During the investigation, two themes were pursued: (i) the combination of the established methods gives a solution for classifier built and systematic research, in a way of assessing the measurement quality and comparing data from various statistical correlations; (ii) the application of discriminant analysis in terms of threshold was introduced to address prediction of target subset of genes by PLS modeling and ANN modeling.

It is clear that the two kinds of classifications were combined effectively in the methodology investigated to establish a framework on finding the regulation for gene expression data.

Acknowledgement

The work was supported by the Doctoral Foundation of Xi'an Jiaotong University (Grant No. DFXJTU 2005-07).

References

[1] Wojna A. Analogy-based reasoning in classifier construction. PhD Dissertation, Warsaw University; 2004.
[2] Chen T, Filkov V, Skiena SS. Identifying gene regulatory networks from experimental data. RECOMB99 1999:94–103.
[3] Golub TR, Slonim DK, Tamayo P, et al. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science 1999;286(5439):531–7.
[4] Brazma A, Vilo J. Gene expression data analysis. FEBS Lett 2000;480(1):17–24.
[5] Akutsu T, Miyano S, Kuhara S. Identification of genetic networks from a small number of gene expression patterns under the boolean network model. In: Proceedings of the fourth Pacific symposium on biocomputing (PSB'99); 1999. p. 17–28.
[6] Liang S, Fuhrman S, Somogyi R. REVEAL, a general reverse. Engineering algorithm for inference of genetic network architectures. In: Proceeding of the third Pacific symposium on biocomputing (PSB'98), vol. 3; 1998. p. 18–29.
[7] Chou RJ, Campbell MJ, Winzeler EA, et al. A genome-wide transcriptional analysis of the mitotic cell cycle. Mol Cell 1998;2(1):65–73.
[8] Rand WM. Objective criteria for the evaluation of clustering methods. J Amer Statist Assoc 1971;66(336):846–50.
[9] Yeung K, Ruzzo W. An empirical study on principal component analysis for clustering gene expression data. Bioinformatics 2001;17(9):763–74.
[10] Hubert LJ, Arabie P. Comparing partitions. J Classif 1985;2:193–218.
[11] Luke BT. Displaying the structure of molecules by multidimensional plots of their torsion angles. J Chem Inf Comput Sci 1993;33:135–42.
[12] Li L, Jiang W, Li X, et al. A robust hybrid between genetic algorithm and support vector machine for extracting an optimal feature gene subset. Genomics 2005;85(1):16–23.