Gene Selection and Classification by Partial Least Squares and Principal component analysis

Hoseok Choi**, Hey-Jin Kim*O, Seungjin Choi*, Sung-Yang Bang*
*Dept. of Computer Science and Engineering
**Dept. of Chemical Engineering, POSTECH, Korea

부분최소자승법과 주성분분석을 이용한 유전자 선택과 분류

최호석**, 김혜진*O, 최승진*, 방승양*
포항공과대학교 *컴퓨터공학과, **화학공학과
{marisan, hohodol, seungjin}@postech.ac.kr

Abstract

DNA chip technology enables us to monitor thousands of gene expressions per sample simultaneously. Typically, DNA microarray data has at least several thousands of variables (genes) with relatively small number of samples. Thus feature (gene) selection by dimensionality reduction is necessary for efficient data analysis. In this paper we employ the partial least squares (PLS) method for gene selection and the principal component analysis (PCA) method for classification. The useful behavior of the PLS is verified by computer simulations.

1. INTRODUCTION

Recent advance in DNA chip technology allows us to monitor thousands of gene expressions per sample simultaneously. One of the challenging problems in DNA microarray data analysis, is how to find underlying gene components which account for the measurement data. In general micro-array data contains the number of variables (genes) far exceeding the number of samples. Typically the number of variables is at least several thousands, whereas the number of samples is several tens [1,2]. In such a case, dimensionality reduction is necessary for efficient data analysis.

The microarray data analysis consists in feature selection (gene selection) and classification. Features can be selected genes or linear combinations of attributes (for example, eigengenes [3]). A major concern of pathologists is to reveal which genes are responsible for a specific disease. Thus a key factor to the successful microarray data analysis would be how to select a few number of genes which explains the measurement data and are also suitable for classification.

Popular dimensionality reduction techniques are principal component analysis (PCA) and partial least squares (PLS). PCA finds a few gene components that explain as much of the observed gene expression variation as possible. However, PCA does not take the prior information on target values (response variables) into account. In contrast to PCA, PLS components are chosen in such a way that the sample covariance between the response and a linear combination of the predictors is maximized. PLS has been used in the field of chemometrics [4]. In this paper we apply the PLS to the task of gene selection and use PCA/Factor analysis for classification.

2. METHODS

The measurement data matrix X (N by p) consists of N samples and p genes. The N-dimensional vector y contains N response variables.

2.1 Partial Least Square (PLS)

PLS components are obtained in such a way that the sample covariance between the response variables and a linear combination of the predictors (genes), are maximized. In other words, the PLS finds a weight vector **w** [6] which satisfies

$$\mathbf{w}_k = \arg\max_{\mathbf{w}'\mathbf{w}=1} \operatorname{cov}^2(\mathbf{Xw}, \mathbf{y})$$

subject to the orthogonal constraint

$$\mathbf{w}'\mathbf{Sw}_j = 0 \text{ for all } 1 \le j \le k$$

where $\mathbf{S}' = \mathbf{X}'\mathbf{X}$. The i-th PLS components are a linear combination of the original predictors ($\mathbf{Xw}_i$).

After the PLS weight vectors are computed, genes are selected via the Variable Importance in the Projection (VIP) [7] which is defined by

$$VIP = \sum_{a} (\mathbf{w}_{ak})^2$$

The VIP is the sum over all model dimensions of the contributions, variable influence (VIN). For a given PLS dimension, $(\mathrm{VIN}_{ak})^2$ is equal to the squared PLS weight $(\mathbf{w}_{ak})^2$. The VIP can be considered as a measurement of how much a certain gene corresponding to the sample state influences. Thus, we select feature gene fifties based on the VIP value according to its descending order and data dimension is reduced.

2.2 Principal component analysis and Factor analysis

We take two algorithms, principal component analysis and factor analysis to separate samples into 2 classes. The response information is used in the step of feature selection and we use unsupervised learning technique for classification in order to avoid over-fitting problem as the input data are too related to the output value [4]. In addition, we represent all results by T2 chart, which provides easier and understandable display.

2.2.1 Principal component analysis

PCA is linear orthogonal transformation to maximize the variance of the linear combination of the predictor variable sequentially,

$$\mathbf{v}_k = \underset{\mathbf{v}'\mathbf{v}=1}{\operatorname{argmax}} \operatorname{var}^2(\mathbf{Xv})$$

subject to the orthogonal constraint

$$\mathbf{v}'\mathbf{S}\mathbf{v}_j = 0, \text{ for all } 1 \le \mathrm{j} \le \mathrm{k}$$

where $\mathbf{S}' = \mathbf{X}'\mathbf{X}$. It shows that PCA is similar to PLS when PLS doesn' t have any information about the response. We can obtain PCs as following equation:

$$\mathbf{X} = \mathbf{TP}' = \sum_{i=1}^{i=k} \mathbf{t}_i\mathbf{p}_i' + \sum_{i=k+1}^{i=p} \mathbf{t}_i\mathbf{p}_i' = \sum_{i=1}^{i=k} \mathbf{t}_i\mathbf{p}_i' + \mathbf{E} = \hat{\mathbf{Z}}_X + \mathbf{E}_X$$

where $\mathbf{t}_i$ is a score vector which contains sample information and $\mathbf{p}_i$ is a loading vector, representing the influence of variables. A score vector is orthogonal and a loading vector is orthonormal. From the two principal components, two score vectors, t[1] and t[2] gives a chart to show identify how data are classified.

2.2.2 Factor analysis

Factor analysis is a statistical approach that can explain interrelationships among a large number of variables and interpret these variables in terms of their common underlying dimensions. Factor analysis is used for an estimate of common variance among the original variables but construct factor scores by means of rotation and interpretation and a certain model base from the initial factor solution. In this paper, we apply Varimax rotation that is an orthogonal rotation criterion that maximize the variance of the squared elements in the columns of a factor matrix. Thus Varimax is the most common rotational criterion.

Factors, unobservable and random quantities are discovered with T2 chart. In other words, we can find not only the prospect of one gene corresponding to one disease but also how much close between genes.

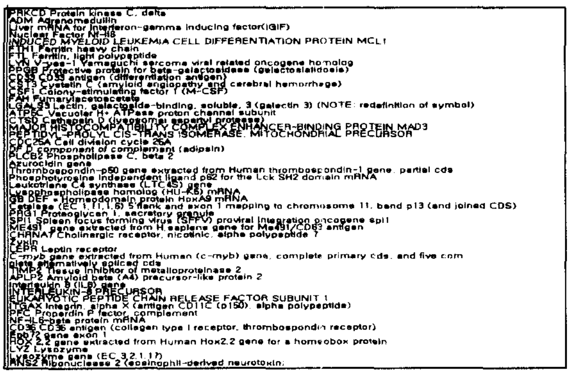
$$\mathbf{y}_i = \mathbf{a}_{i1}\mathbf{f}_1 + \mathbf{a}_{i2}\mathbf{f}_2 + \ldots + \mathbf{a}_{ik}\mathbf{f}_k + \mathbf{e}_i$$

where $\mathbf{y}_i$ is the i-th observed variable and $\mathbf{e}_i$ is the residual of $\mathbf{y}_i$.

3. Experiment and Results

We use the leukemia dataset which is available from the website, http:// www.genome.wi.mit.edu/MPR. The data consist of 47 acute lymphoblastic leukemia (ALL) and 25 acute myeloid leukemia (AML). And ALL data are made of 38 B-cell ALL and 9 T-cell ALL with 7129 genes. The French America British (FAB) system categorized AML into M1, M2, ⋯ M7 based on appearance of leukemia cells under the microscope after routine processing or cytochemical staining, of which classification are associated with certain pathological symptoms such as bleeding, blood clotting and so on. For example, M3 leukemia is apt to response to retinoids, important factor in drug discovery of vitamin A. Another instance is M5. M5 subtype of AML tends to have a worse prognosis and many doctors recommend more intensive chemotherap y for these inheritors.

First of all, we extract chief features. Among 7129 genes, 50 genes are selected on the basis of VIP value by the method of PLS. VIP is squared weighting factor or correlation coefficient. The higher VIP implies, the more donated genes are to cause such response. The number, 50 is not meaningful and just chosen followed by Golub et al (1999). [Table [1]]



PRKCD Protein kinase C, delta
ADM Adrenomedullin
Liver mRNA for interferon-gamma inducing factor(IGIF)
Nuclear Factor Nf-IIB
INDUCED MYELOID LEUKEMIA CELL DIFFERENTIATION PROTEIN MCL1
FTH1 Ferritin heavy chain
FTL Ferritin, light polypeptide
LYN V-yes-1 Yamaguchi sarcoma viral related oncogene homolog
PPGB Protective protein for beta-galactosidase (galactosialidosis)
CD33 CD33 antigen (differentiation antigen)
CST3 Cystatin C (amyloid angiopathy and cerebral hemorrhage)
CSF1 Colony-stimulating factor 1 (M-CSF)
FAH Fumarylacetoacetate
LGALS3 Lectin, galactoside-binding, soluble, 3 (galectin 3) (NOTE: redefinition of symbol)
ATP6C Vacuolar H+ ATPase proton channel subunit
CTSD Cathepsin D (lysosomal aspartyl protease)
MAJOR HISTOCOMPATIBILITY COMPLEX ENHANCER-BINDING PROTEIN MAD3
PEPTIDYL-PROLYL CIS-TRANS ISOMERASE, MITOCHONDRIAL PRECURSOR
CDC25A Cell division cycle 25A
DF D component of complement (adipsin)
PLCB2 Phospholipase C, beta 2
Azurocidin gene
Thrombospondin-p50 gene extracted from Human thrombospondin-1 gene, partial cds
Phosphotyrosine independent ligand p62 for the Lck SH2 domain mRNA
Leukotriene C4 synthase (LTC4S) gene
Lysophospholipase homolog (HU-K5) mRNA
GB DEF = Homeodomain protein HoxA9 mRNA
Catalase (EC 1.11.1.6) 5'flank and exon 1 mapping to chromosome 11, band p13 (and joined CDS)
PRG1 Proteoglycan 1, secretory granule
SPI1 Spleen focus forming virus (SFFV) proviral integration oncogene spi1
ME491 gene extracted from H.sapiens gene for Me491/CD63 antigen
CHRNA7 Cholinergic receptor, nicotinic, alpha polypeptide 7
Zyxin
LEPR Leptin receptor
C-myb gene extracted from Human (c-myb) gene, complete primary cds, and five com
plete alternatively spliced cds
TIMP2 Tissue inhibitor of metalloproteinase 2
APLP2 Amyloid beta (A4) precursor-like protein 2
Interleukin 8 (IL8) gene
INTERLEUKIN-8 PRECURSOR
EUKARYOTIC PEPTIDE CHAIN RELEASE FACTOR SUBUNIT 1
ITGAX Integrin, alpha X (antigen CD11C (p150), alpha polypeptide)
PFC Properdin P factor, complement
NF-IL6-beta protein mRNA
CD36 CD36 antigen (collagen type I receptor, thrombospondin receptor)
Epb72 gene exon 1
HOX 2.2 gene extracted from Human Hox2.2 gene for a homeobox protein
LYZ Lysozyme
Lysozyme gene (EC 3.2.1.17)
RNS2 Ribonuclease 2 (eosinophil-derived neurotoxin)

Table 1. key genes

Secondly, the PCA method applied to classify the dataset. It is well shown at the result that the data matrix, 34 by 50 decomposed into two groups, ALL and AML with the accuracy of 33 samples / 34 samples. [Figure [1], Figure [2]]

ALL can be classified into T-cell lineage and B-cell lineage. In immunologic subtypes, pre B-cell lineage ALL can be cured better than others. Also we classified ALL cases into B-cell lineage. All B and T -cell lineage ALL well separated except 17 samples. In addition, the result figure gave an idea that the M5 subtype of AML tends to have a worse prognosis. It implies that usage of the method of this paper can utilize the decision whether the diagnosis is well or not, comparing with other well-established results. Figure [3]

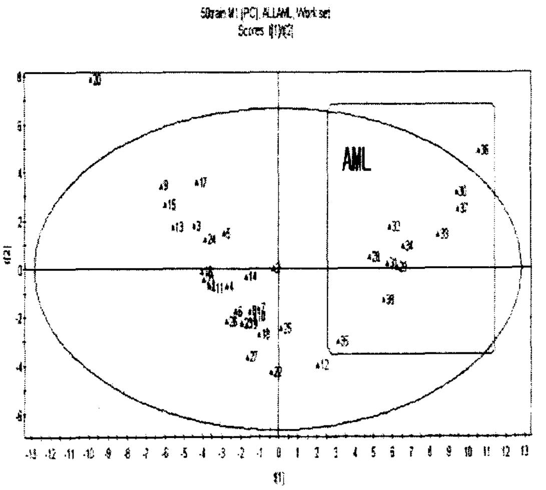


Figure1. Scoring plot for training data set.

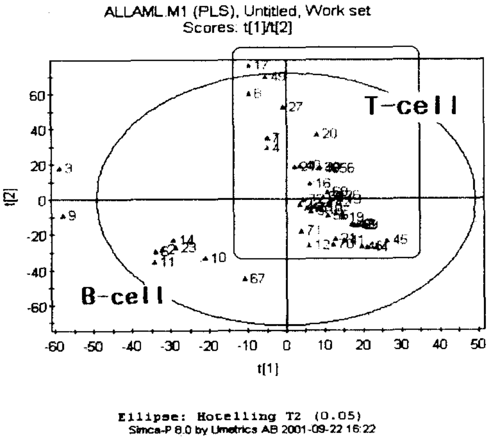


Figure3. Scoring plot of B cell and T cell in ALL

Finally, factor analysis through varimax rotation, we can know what factors are important in some genes. A scree plot, a method to identify how many PCs to compute, indicated 5 factors. 17 genes of 25 genes were much influenced by the first factor corresponding to the subtype acute lymphoblastic leukemia. Instead, 18 genes among 25 genes were much influenced by the second factor, mainly in a subtype AML.

4. Conclusion

Microarray technology has revealed high throughput technology. It has been expected to play a key role of disease discovery and give a novel way of diagnosis in the near future. Essential information from microarray are mainly the gene finding and the functions of genes. The approach of this paper is focused on the purpose.

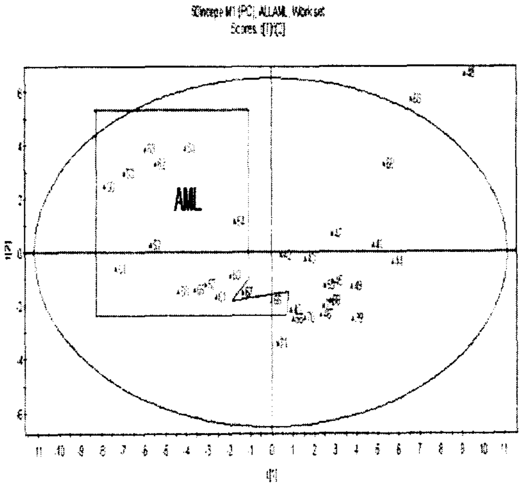


Figure 2. Scoring plot for test data set.

Using VIP from the PLS, we can find essential genes effectively. Moreover, T2 charts reflects classes of samples not only ALL and AML classes but also B cell and T cell classes, which are not well classified in other papers.

Classification methods, PCA and factor analysis have the benefits, which is well dimension reduction a good display as shown above. Based on the result, we assume that each position of sample patients reflect a certain relation between diseases and patients. Although this is just hypothesis and requires much more work to get significant information, this paper showed the possibility.

5. Reference

[1] T. R .Golub, D. K. Slomin, P. Tanmayo, C. Huard.Gaasenbeek, J.P.Mesirov, H.Coller, M.L.Loh, J.R.Downing, M. A.Caligiuri, C.D., " Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring," Science,1999.

[2] Javed Khan, et.al, " Classification and Diagnostic prediction of cancers using gene expression profiling and artificial neural network," Nature Medicine 2001 June

[3] Orly Alter, Patrick Brown, and David Botstein, " Singular value decomposition for genome-wide expression data processing and modeling," PNAS, August 29, 2000 vol.97 no.18

[4] Danh V.Nguyen, David M.Rocke, " Tumor classification by partial least square using microarray gene expression data," University of Califonia, Davis, Davis,CA95616, February 2001

[5] Paul Geladi and Bruce R.Kowalsk," Patial least square regression: a tutorial," Analytica Chimica Acta, 185 (1986) 1-17 Elsevier Science Publishers B.V., Amsterdam

[6] Ka Yee Yeung, Walter L. Ruzzo, " An empirical study on Pricipal Component Analysis for clustering gene expression data," Technical report UW-CSE-2000-11-03

[7]" SIMCA-P for windows: Graphical software for multivariate process modeling; multivariate modeling, analysis and SPC of Process Data," April 19, 1996