

DNA Microarray Data Analysis for Cancer Classification Based on Stepwise Discriminant Analysis and Bayesian Decision Theory

Sungwoo Kwon¹ Young-Hwan Chu¹
Sw74@postch.ac.kr dosa@postech.ac.kr

Heui-Seok Yi² Chonghun Han^{2,3}
yihs@postech.ac.kr chan@postech.ac.kr

- ¹ Department of Chemical Engineering, Pohang University of Science and Technology, Pohang, Kyungbuk, 790-784, Korea
² Automation Research Center, Department of Chemical Engineering, Pohang University of Science and Technology and ISYSTECH Inc., Pohang, Kyungbuk, 790-784, Korea
³ School of Environmental Science and Engineering, Pohang University of Science and Technology and ISYSTECH Inc., Pohang, Kyungbuk, 790-784, Korea

Keywords: DNA chip, cancer classification, bioinformatics, Bayesian classifier

1 Introduction

DNA microarray are providing an unprecedented amount of information about the genetic changes. From the analysis of the data, we can get insights into various diseases such as cancer whose information is difficult to be obtained. Golub *et al.* showed the possibility of cancer classification based on gene expression profiles [1]. They clearly showed that the detection of wholesale changes in gene expression patterns is possible. However, researchers could not identify the genes whose activity is turned up or down. They also could not find which changes of genes are important for cancer development and progression—the causes and not just the effects. Therefore, the useful methods to analyze DNA microarray data have not been proposed.

In this paper, we find the causal relationship between several tumors and the gene-expression data by sequentially using the stepwise discriminant analysis method (SDA) and Bayesian decision theory (BDT). Eighty-five samples containing four tumor classes are used in this study. The classes are neuroblastoma (NB), rhabdomyosarcoma (RMS), non-Hodgkin lymphoma (BL) and the Ewing family of tumor (EWS). These data are open on world wide web [2]. SDA is used to select critical genes for accurate classification of 4 tumors from original 2308 genes. With the selected genes, Bayesian classifier is made, which minimizes the misclassification rate. As a result, the classification performance increases to 100%, and 9 new genes that have relation with the development of the tumors is found additionally.

2 Methods and Results

2.1 Stepwise Discriminant Analysis (SDA)

SDA is the feature selection method that repeats the addition and removal of a feature at each step. This process allows us to find the best subset with which satisfactory discrimination performance can be obtained. By using only small feature set, we can significantly reduce the unnecessary efforts and

costs. Only one feature among the all of possible ones is added if the Wilk's lambda decreases the most apparently by including the feature. After that, only one feature is removed from the selected feature set if exclusion of the feature increases the Wilk's lambda the most insignificantly. This process is repeated until the Wilk's lambda does not change any more in a statistical aspect.

SDA is applied to the data containing 2308 genes and 72 genes are classified. Among the 72 genes, we again chose 15 genes because classification power slowed down after selection of 15 genes. The selected genes is compared with the ranked genes of original paper [2]. Among the 15 genes, 6 genes coincided with the ranked genes of original paper, but 9 genes were new ones. Table 1 shows these genes. From Figure 1, we can see that these genes were highly expressed in some tumors, but not expressed in other tumors.

2.2 Bayesian Decision Theory (BDT)

Assuming that the prior probabilities of each class and the class-specific density function are known, we can estimate the probability with which \mathbf{x} belongs to class \mathbf{t} , $\mathbf{P}(\mathbf{t}|\mathbf{x})$, by applying BDT. If a test sample is to be classified, $\mathbf{P}(\mathbf{t}|\mathbf{x})$ is calculated for each class, and the sample is classified into the class which has the largest value of $\mathbf{P}(\mathbf{t}|\mathbf{x})$ since this minimizes the misclassification rate. We assumed that data in each class follows a multivariate normal distribution, and made class-conditional density functions on the assumption. We also used generalized squared distance instead of Mahalanobis distance as exponential term of the normal density function, such as equation (1). The generalized squared distance considers the prior probabilities and misclassification cost.

$$P(x|t) = \frac{1}{(2\pi)^{\frac{d}{2}} |\Sigma|^{\frac{1}{2}}} e^{[-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu) + g_1 + g_2]} \quad (1)$$

We used 67 samples as training data (EWS: 25, NB: 14, BL: 7, RMS: 21) and 16 samples (EWS: 4, NB: 4, BL: 4, RMS: 4) as test data. 16 test samples were respectively classified into one of the 4 classes by the Bayesian classifier containing 15 genes. As a result, all of the test data were correctly classified, and therefore, error rate were 0%.

Table 1: Newly discovered genes and their functions.

Image ID	Associated cancer	Gene function
771323	RMS	procollagen-lysine, 2-oxoglutarate 5-dioxygenase
469345	RMS,EWS	kinase insert domain receptor (a type III receptor tyrosine kinase)
142134	RMS	ESTs
770394	EWS	Fc fragment of IgG, receptor, transporter, alpha
796258	RMS	sarcoglycan, alpha (50kD dystrophin-associated glycoprotein)
183337	BL	major histocompatibility complex, class II, DM alpha
812105	NB	transmembrane protein
1434905	EWS,RMS	homeo box B7
236282	BL	Wiskott-Aldrich syndrome (eczema-thrombocytopenia)

3 Conclusion

In this paper, we applied SDA and BDT with generalized squared distance to classify tumor samples with their gene-expression data. Misclassification rate for test data was 0%. In addition, we also discovered 9 new genes related with the tumors. Also, to reduce fabrication cost of DNA chip for diagnostic use, we established that the 15 genes reduced the misclassification to zero.

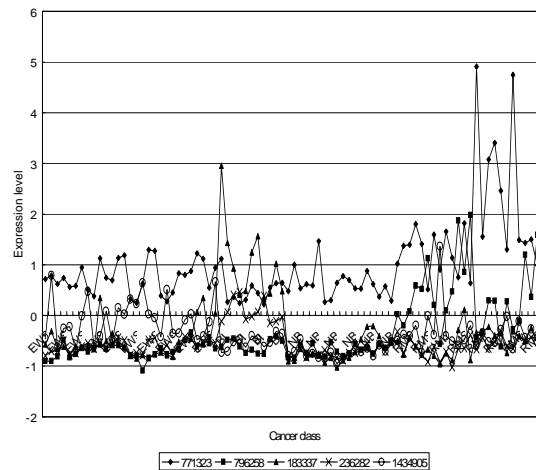


Figure 1: New discovered gene expression profile.

Our results demonstrate that SDA is useful method to discover key genes that have strong relationship with each tumor, and Bayesian classifier can be effectively utilized to the classification of tumor cases by their gene-expression data on the appropriate assumption.

References

- [1] Golub, T.R., Slonim, D.K., Tamayo, P., and Retief, J., Molecular classification of cancer: class discovery and class prediction by gene expression monitoring, *Science*, 286(5439):531–537, 1999.
- [2] Khan, J., Jun, S., Wei, M., Lao, R., and Saal, H., Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks, *Nature Medicine*, 7(6):673–679, 2001.