New gene selection method for classification of cancer subtypes considering within-class variation

Ji-Hoon Cho[a], Dongkwon Lee[a], Jin Hyun Park[b], In-Beum Lee[a,*]

[a] *Department of Chemical Engineering, Pohang University of Science and Technology, San 31 Hyoja-Dong, Pohang 790-784, South Korea*
[b] *P&I Consulting Co., Ltd., San 31 Hyoja-Dong, Pohang 790-784, South Korea*

Received 5 May 2003; accepted 13 July 2003

First published online 11 August 2003

Edited by Julio Celis

Abstract In this work we propose a new method for finding gene subsets of microarray data that effectively discriminates subtypes of disease. We developed a new criterion for measuring the relevance of individual genes by using mean and standard deviation of distances from each sample to the class centroid in order to treat the well-known problem of gene selection, large within-class variation. Also this approach has the advantage that it is applicable not only to binary classification but also to multiple classification problems. We demonstrated the performance of the method by applying it to the publicly available microarray datasets, leukemia (two classes) and small round blue cell tumors (four classes). The proposed method provides a very small number of genes compared with the previous methods without loss of discriminating power and thus it can effectively facilitate further biological and clinical researches.

Key words: Gene expression data; Gene selection; Classification; Centroid; Within-class variation; Kernel Fisher's discriminant analysis

1. Introduction

Gene selection (the choice of discriminatory genes) is one of the most challenging issues in the field of microarray data analysis. Gene expression data generally contains a large number of genes (variables) compared to the number of samples, hence conventional data mining techniques cannot be directly applied to the data [1,2,11] due to the singularity problem, the curse of dimensionality and so on. For this reason, the analysis of gene expression data needs performing a dimension reduction technique, gene selection which subtracts the genes most highly correlated to the pattern of each type of disease in order to avoid such problems. Statistical approaches including parametric and non-parametric tests, for example *t*-test and Wilcoxon rank sum test, have been widely used for finding differentially expressed genes since they are easy to understand and implement. However, they have a potential limitation to extend in the case of more than two classes and require time-consuming adjustment to solve the problem of multiple testing [3]. For three or more groups, the Kruskal–Wallis test can be used. However, it may produce biased results because of the dependence on the number of samples, when it is applied to microarray data whose sample sizes are usually unbalanced.

In addition to the statistical methods, extensive researches have focused on gene selection. Golub et al. [4] used signal-to-noise ratio as a criterion for measuring the correlation between a gene and a cancer subtype. Hastie et al. [5] developed the so-called gene shaving method with principal component analysis. Recently, Tibshirani et al. [6] suggested the nearest shrunken centroid method combined with classification and Lee et al. [15] selected the informative gene subset using Bayesian learning. Also, Guyon et al. [7] adopted the support vector machine for recursive feature extraction. These methods can be divided into two categories: individual gene ranking approaches and gene subset ranking approaches [8].

Here, in a supervised manner, we propose a novel gene selection method which belongs to the individual gene ranking approaches, with emphasis on small within-class variation (homogeneity in a certain class) as well as differential expression between groups. Our proposed method consists of two steps, one is a gene ranking and selection step and the other is a validation step. Utilizing sample distances from the centroid of each class, we developed a new metric which reflects the relevance of a gene and ranked genes according to the metric. For validating the classification ability of selected genes, we used the kernel Fisher's discriminant analysis (KFDA) which is a remarkable technique for analyzing gene expression data since it is irrespective of the number of variables (genes) and it generally provides satisfactory results [9]. The main advantage of our method is that it guarantees small within-class variation which has been indicated as a drawback of this kind of method (so-called individual gene ranking approach) [8] and it is simply extended when there are more than two classes. A subset which shows a minimum test error rate is chosen as an optimal candidate set for classification of cancer subtypes. To evaluate the performance of our proposed method, we applied it to publicly available microarray datasets, acute leukemia [4] and small round blue cell tumors (SRBCT) [10].

2. Materials and methods

2.1. Biological data

The leukemia dataset has 7129 probes and 72 train samples and test ones and the SRBCT dataset has 2308 genes (including expressed sequence tags) that were preliminarily chosen by Khan et al. [10] and a total of 83 samples (we removed five non-SRBCT samples). The former dataset has two classes, acute lymphoblast leukemia (ALL) and acute myeloid leukemia (AML), and the latter is classified into four classes, Burkitt's lymphoma (BL), Ewing's sarcoma (EWS), rhabdomyosarcoma (RMS), and neuroblastoma (NB). Both datasets

[*] Corresponing author. Fax: (82)-54-279 3499.
E-mail addresses: cjhjhj@postech.ac.kr (J.-H. Cho), miru@postech.ac.kr (D. Lee), pcanda@postech.ac.kr (J.H. Park), iblee@postech.ac.kr (I.-B. Lee).

0014-5793/03/$22.00 © 2003 Federation of European Biochemical Societies. Published by Elsevier B.V. All rights reserved.
doi:10.1016/S0014-5793(03)00819-6

are publicly available at http://www-genome.wi.mit.edu/cancer and http://research.nhgri.nih.gov/microarray/Supplement/index.html.

2.2. Mathematical formulation of a new metric

The basic idea of this method is to identify genes which have short distances from each class centroid and have simultaneously small variation within the class. Assume that we have two-dimensional data which has two classes as shown in Fig. 1. In the figure, one can intuitively recognize that x_1 is relevant for discriminating two classes since samples of x_1 are not distant from the each class centroid and show little variation within each class. In brief, the distance vector of relevant variable x_1 is composed of small distance values and thus has small variance as well as small mean value.

Suppose that we have $p \times n$ data matrix, $\mathbf{X}$ and x_{ij} be the expression for ith gene and jth sample ($i=1, 2, ..., p$ and $j=1, 2, ..., n$). The data have total K classes and n_k samples in class k. Let C_k be the index of the samples in the kth class. The expression value of each gene is assumed to be auto-scaled.

The ith element of the centroid for class k is obtained as follows.

$$\overline{x}_{ik} = 1/n_k \sum_{j \in C_k} x_{ij} \tag{1}$$

It represents the mean expression value in class k for gene i. We define a distance matrix, $\mathbf{Z}$, of which each element z_{ij} is calculated as follows.

$$z_{ij} = \sqrt{(x_{ij} - \overline{x}_{ik})^2}, \text{ where } j \in C_k \tag{2}$$

We name $\mathbf{z}_i$ ($1 \times n$ row vector) the within-class distance vector of gene i since $\mathbf{z}_i$ consists of n_1 distances from the centroid of class 1, n_2 distances from the centroid of class 2, ..., and n_K distances from the centroid of class K, i.e. total n distances from each class centroid. The schematic diagram of the mathematical procedure is configured in Fig. 2 and conceptual illustration is shown in Fig. 3.

If gene i is suitable for classification (differentially expressed across cancer subtypes and homogeneously dispersed within each class), $\mathbf{z}_i$, the corresponding within-class distance vector, has a small standard deviation with a small mean value. Note that the sample mean and standard deviations depend upon the number of samples, which may distort the statistics especially for the case of multiple classes. Consider that we have three distance vectors, $\mathbf{z}_i$, $\mathbf{z}_j$ and $\mathbf{z}_k$ as follows.

$$\mathbf{z}_i^T = \begin{bmatrix} 1 \\ 1 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \mathbf{z}_j^T = \begin{bmatrix} 0 \\ 0 \\ 1 \\ 1 \\ 1 \\ 1 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \mathbf{z}_k^T = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 1 \\ 1 \\ 1 \end{bmatrix} \begin{matrix} \left.\begin{matrix} \\ \\ \end{matrix}\right\} \text{class 1} \\ \left.\begin{matrix} \\ \\ \\ \\ \end{matrix}\right\} \text{class 2} \\ \left.\begin{matrix} \\ \\ \\ \end{matrix}\right\} \text{class 3} \end{matrix}$$

The sample mean and standard deviation of the vectors are different. However, we want to make the statistic of three vectors be equal and thus eliminate the effect of the number of samples since they have the same property, i.e. a distance value of 1 for a certain class and 0 for other classes. Weighted mean and standard deviation (variance) can achieve our aim and they are formulized as follows.

For gene i,

$$\text{mean}_w(\mathbf{z}_i) = \sum_{j=1}^{n} \frac{w_j}{W} z_{ij} \tag{3}$$

$$\text{std}_w(\mathbf{z}_i) = \sqrt{\frac{\sum_{j=1}^{n} (z_{ij} - \text{mean}_w(\mathbf{z}_i))^2}{(n-1/n)\sum_{j=1}^{n} w_j}} \tag{4}$$

where $W = \sum_{j=1}^{n} w_j$, $w_j = \frac{1}{n_k}(j \in C_k)$

Thus, we develop a new metric which reflects the relevance of gene i as follows.

$$r_i = \text{mean}_w(\mathbf{z}_i) \cdot \text{std}_w(\mathbf{z}_i) \tag{5}$$

The small value of r_i suggests that samples of the ith gene are dispersed near the centroid of each class (mean value of $\mathbf{z}_i$) and are assembled (standard deviation of $\mathbf{z}_i$) simultaneously. And hence we can conclude that the ith gene is subtype-specific and has small within-class variation. It is necessary to consider the mean and variation (standard deviation) together, since it is possible for two distance vectors, $\mathbf{z}_i$ and $\mathbf{z}_j$, to have the same standard deviation although they have different mean values.

However, there is a fatal weak point. If there is a gene which shows no differential expression across classes and nearly uniform expression, it will be chosen as a relevant gene with regard to the r_i value. To prevent such a situation, we included the difference of the centroid of each class and modified r_i as follows.

$$R_i = \frac{\text{mean}_w(\mathbf{z}_i) \cdot \text{std}_w(\mathbf{z}_i)}{\text{std}(\overline{\mathbf{x}}_i)} \tag{6}$$

where $\overline{\mathbf{x}} = [\overline{x}_{i1}, \overline{x}_{i2}, ..., \overline{x}_{iK}]$ denotes the $1 \times K$ centroid vector of gene i and std(·) means standard deviation calculation.

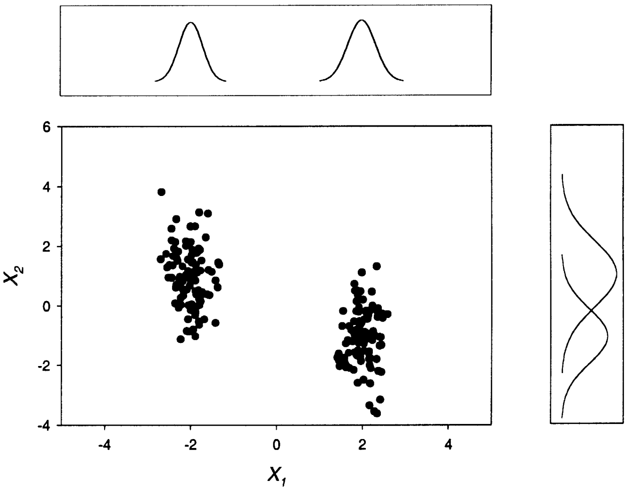

Fig. 1. Plot of simulated dataset. The relevant variable, x_1, shows two distributions that are apart from each other and have small variation (standard deviation).

2.3. Classification strategy

For classification of samples, we chose to use KFDA [9]. While the conventional Fisher's discriminant analysis (FDA) does not work when there is a larger number of variables than samples due to the singularity problem [1,2,11], KFDA is very effective for gene expression data with high dimensionality since the kernel methods use dot products and do not need to perform high dimensional matrix computation [9,12]. Using the kernel trick, we can obtain n (the number of samples which is usually less than 100)-dimensional scatter matrices which are invertible so that they can produce projection weight vector (or matrix). One can find the detailed algorithm of KFDA in the literature [9]. For each selected gene subset, we obtained KFDA scores (generally one less than K, the number of classes) in the same way as conventional FDA and classified them by calculating posterior probabilities.

The classification is achieved by the following procedure. With the projection weight vector, we can obtain the discriminant function (score), $\mathbf{y}_j$ ($K-1 \times 1$ column vector) of input $\mathbf{x}_j$ ($j=1, 2, ..., n$). Then, the chi-square distance of the jth sample from the centroid of each class is computed by

$$\chi^2_{j,k} = (\mathbf{y}_j - \overline{\mathbf{y}}_k)^T \mathbf{D}_k^{-1} (\mathbf{y}_j - \overline{\mathbf{y}}_k) \tag{7}$$

where $\mathbf{D}_k$ is the covariance matrix of $\mathbf{y}$ for class k and

$\bar{\mathbf{y}}_k = 1/n_k \sum_{j \in C_k} \mathbf{y}_j$ denotes a class centroid of discriminant score. The posterior probability is calculated as follows:

$$P(k|\mathbf{x}_j) = \frac{P_k |\mathbf{D}_k|^{-1/2} \exp(-\chi^2_{j,k}/2)}{\sum_{k'} P_{k'} |\mathbf{D}_{k'}|^{-1/2} \exp(-\chi^2_{j,k'}/2)} \tag{8}$$

where P_k is the prior probability for class k. A sample is assigned into the class for which $P(k|\mathbf{x}_j)$ is highest. Note that if the number of samples for a certain class is less than the number of classes, one cannot calculate the Fisher's discriminant function due to rank deficiency of matrix, $\mathbf{D}$.

3. Results

We analyzed the two datasets which represent binary classification and multiple classification problems, respectively.

The procedure is as follows. Removing the proportion, α (in this paper, α = 20%) of the genes having largest value of R_i, we monitored the classification error rate using KFDA and posterior probabilities. We used five-fold cross-validation, i.e. partitioned the set of samples into five approximately equal-sized parts. We trained the classifier with four parts and then predicted the class of the remaining part for evaluating test error rate. The distribution of class labels should be roughly balanced so that each sample is predicted for one time and only one. This procedure should be repeated five times and the mean error on all five times produces the mean cross-validation error. Also, we repeated such cross-validation 20 times and took an overall mean cross-validation error rate, considering the arbitrariness of partitioning (therefore a total

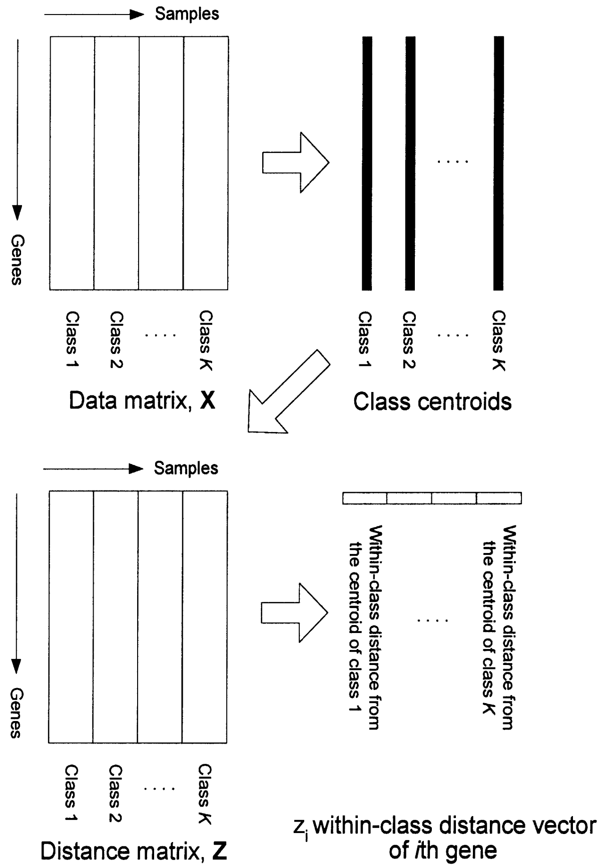

Fig. 2. Schematic diagram of the mathematical formulation procedure.

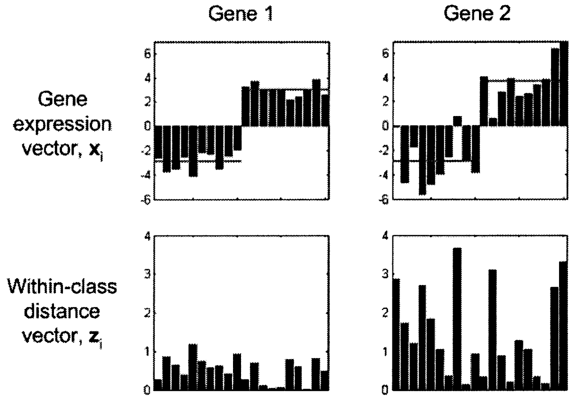

Fig. 3. Conceptual illustration of within-class distance vector, $\mathbf{z}_i$. The bar indicates the gene expression value of each sample and the horizontal line represents the value of each class centroid. Both genes have a similar mean difference between two classes, but gene 1 is more suitable and robust for classification since it has small within-class variation. Consequently, the within-class distance vector of gene 1 shows a small mean and standard deviation.

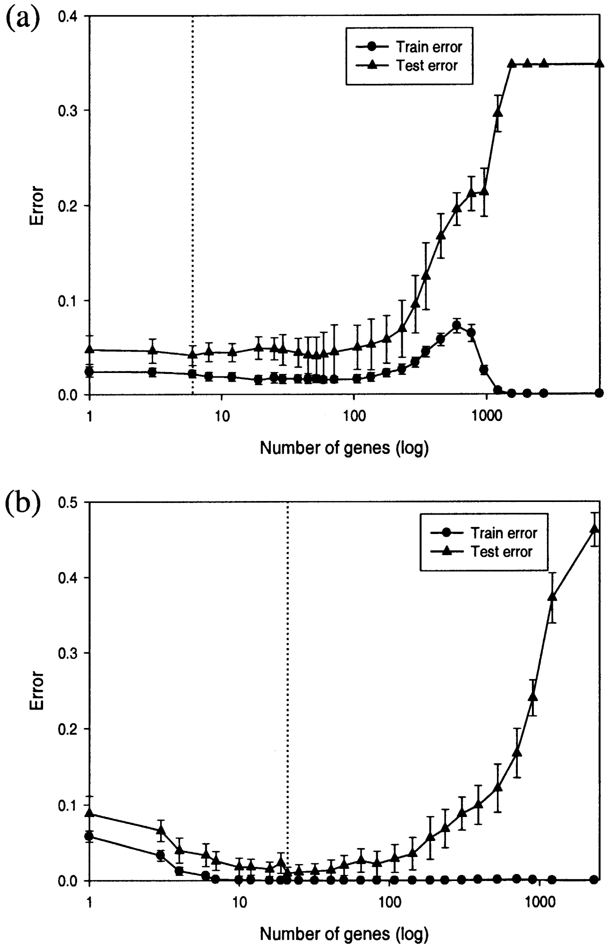

Fig. 4. Classification error rate of two datasets: (a) leukemia and (b) SRBCT. Train and test error are shown as a function of the number of genes always used during total cross-validation. The dashed line indicates the minimum test error point.

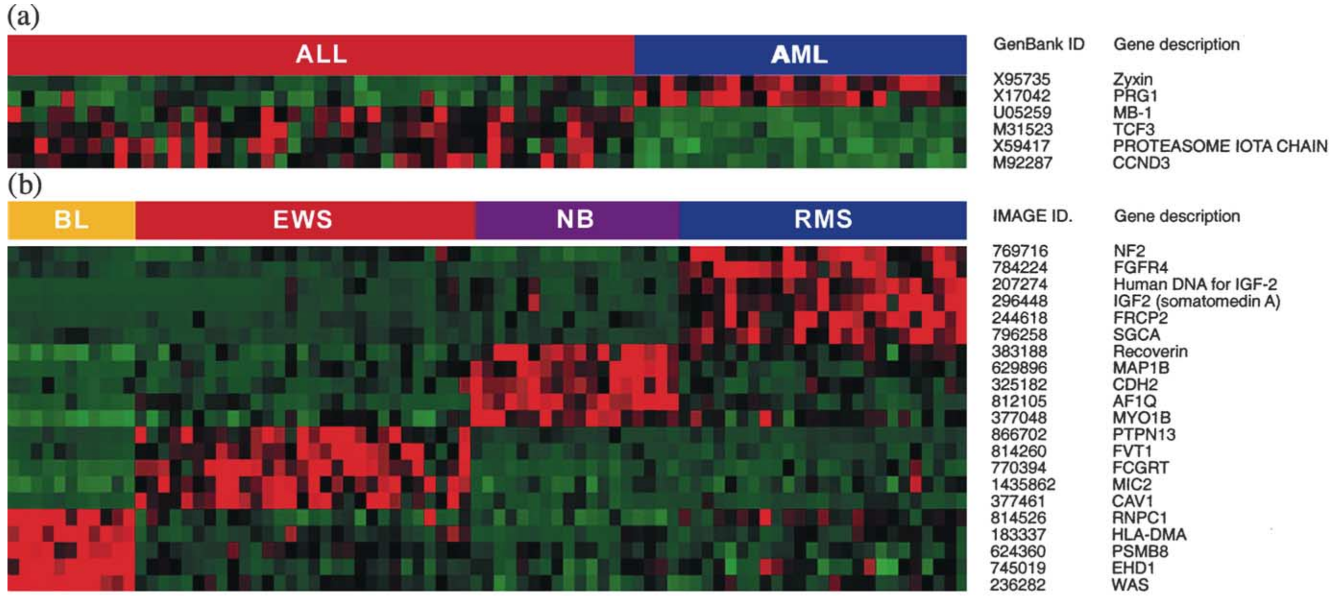

Fig. 5. Gene expression maps of the selected genes for (a) leukemia and (b) SRBCT. Within each of the cancer types, the genes are ordered by hierarchical clustering (average linkage) for clear illustration. We used CLUSTER and TREEVIEW software which are publicly available at http://rana.lbl.gov.

of 100 validations were performed). As a result, we could extract a discriminatory gene subset which showed the minimum cross-validation error rate.

3.1. Classification and gene subset selection

Fig. 4 shows the results of the cross-validation for the two datasets. As expected, the test error tends to decrease with eliminating irrelevant genes. During the cross-validation, the leukemia dataset showed the minimum test error (overall mean cross-validation error, 4.06%) when 6 genes were always used during the total of 100 validation procedures. For the SRBCT data, 21 genes which also always participated in the whole 100 validations gave the minimum overall mean error (0.96%). We considered the above gene subset (6 genes for leukemia and 21 genes for SRBCT) as optimal gene sets for classification of subtypes. For evaluation of optimality, selected gene subsets passed through leave-one-out cross-validation (LOOCV) using conventional FDA and KFDA. Table 1 represents the LOOCV classification results of our selected gene subset and previously published ones for comparison. As one can see, our method markedly reduced the number of genes without loss of separability.

Table 1
LOOCV classification results of our proposed gene subset and previously selected ones: (a) leukemia data and (b) SRBCT data

	Conventional FDA	KFDA	Number of genes
a: Leukemia data			
Golub et al. [4]	9	4	50
Lee et al. [15]	5	3	5
Proposed	3	2	6
b: SRBCT data			
Tibshirani et al. [6]	2	2	43
Proposed	0	0	21

Numerical values indicate the number of misclassifications.

For the leukemia data, recently Lee et al. [15] successfully found the relevance of genes using the Bayesian learning method and made a good classification result, but their gene selection criterion was somewhat arbitrary. As for the SRBCT data, Tibshirani et al. [6] developed a new technique for gene selection and found 43 genes as an optimal set, however, in spite of reducing many genes compared with Khan et al. [10], we concluded that their selection was still relevant according to the comparison results. On the contrary, our method selected the minimal number of genes with explicit selection criteria based on cross-validated classification performance. Note that we provided most plausible classification results since we did not make any assumption (e.g. diagonal within-covariance matrix as [6]) in the comparison study. Fig. 5 effectively illustrates the gene selection performance of our method. From the figure, we can see that selected genes have a significant mean difference between groups (specific subtype and others) and also have small variation within each group, and thus we can consider that they are definitely discriminatory.

3.2. Biological analysis of identified genes

Many studies have been accomplished with the leukemia data [2,4,13–15], and most genes we found were part of previously chosen ones. We also found 21 genes most responsible for SRBCT subtype classification. Previous studies already addressed that they were responsible for subtype classification of SRBCT and suggested plausible conclusions about the biological function of the genes [6,10]. However, we identified some genes not identified in other works for SRBCT. We found the Wiskott–Aldrich syndrome (WAS) gene, overexpressed in Burkitt's lymphoma, known to be associated with Bruton's tyrosine kinase which plays an important role in normal B-cell lymphocyte development [16]. Some publications have reported that the WAS gene is included in lymphoid cell signaling and its expression levels play a substantial role in determining immune outcome [17,18]. This result might reflect the fact that one with immunodeficiency (e.g. WAS) has a high possibility of lymphoid malignancy (non-Hodgkin lymphoma) [19,20].

4. Discussion

Gene selection is a crucial step for analyzing microarray data since there are a large number of genes (irrelevance), which makes it difficult to handle such data. In this paper, we proposed a novel gene selection criterion combined with KFDA. Intuitively, the goal of the proposed method is to find the minimal set of genes that are closely located around the class centroid and thus can effectively discriminate cancer patterns. We adopted a so-called individual gene ranking approach to sort several thousands genes and the KFDA method to evaluate the classification power of a selected subset of genes. In fact, it has been reported that the individual ranking method has the disadvantage that genes found by the method might not have small within-class variation [8]. This problem was solved by considering mean and standard deviation of sample distances from the class centroid. In addition, by considering weighted metrics, we removed the dependence on the sample size of each class. The proposed method gave a satisfactory classification performance with informative genes which are specific to a certain type of cancer. Moreover, it considerably reduced the complexity of the data without loss of class prediction performance even in the multiple classification problems. Although it is hard to assess that only a selected subset of genes is optimal for classifying subtypes, such genes may be strong candidates which represent a certain type of cancer. With the informative subset, one may improve the discriminatory power as much as possible by combining with other supervised pattern recognition techniques. Lastly, this kind of work would be used to find responsive drug targets. If proper gene selection is accompanied by biological and clinical research, our work may speed up and facilitate experimental work and thus be the cornerstone of therapeutic target discovery.

Acknowledgements: This work was supported by the Brain Korea 21 project.

References

[1] Duda, R.O., Hart, P.E. and Stork, D.G. (2001) Pattern Classification, 2nd edn., John Wiley and Sons, New York.

[2] Cho, J.-H., Lee, D., Park, J.H., Kim, K. and Lee, I.-B. (2002) Biotechnol. Prog. 18, 847–854.

[3] Dudoit, S., Yang, Y.H., Speed, T.P. and Callow, M.J. (2002) Stat. Sin. 12, 111–139.

[4] Golub, T.R., Slonim, D.K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J.P., Coller, H., Loh, M.L., Downing, J.R., Caligiuri, M.A., Bloomfield, C.D. and Lander, E.S. (1999) Science 286, 531–537.

[5] Hastie, T., Tibshirani, R., Eisen, M.B., Alizadeh, A. Levy, R. Staudt, L., Chan, W.C., Botstein, D. and Brown, P. (2001) Genome Biol. 1, research0003.1–0003.21.

[6] Tibshirani, R., Hastie, T., Narasimhan, B. and Chu, G. (2002) Proc. Natl. Acad. Sci. USA 99, 6567–6572.

[7] Guyon, I., Weston, J., Barnhill, S. and Vapnik, V. (2002) Mach. Learn. 46, 389–422.

[8] Lu, Y. and Han, J. (2003) Inf. Syst. 28, 243–268.

[9] Mika, S., Rätsch, G., Weston, J, Schölkopf, B. and Müller, K.-R. (1999) Proc. IEEE Neural Networks for Signal Processing Workshop, pp. 41–48.

[10] Khan, J., Wei, J.S., Ringner, M., Saal, L.H., Ladanyi, M., Westermann, F., Berthold, F., Schwab, M., Antonescu, C.R., Peterson, C. and Meltzer, P.S. (2001) Nat. Med. 7, 673–679.

[11] Sharma, S. (1996) Multivarate Techniques, John Wiley and Sons, New York.

[12] Schölkopf, B., Smola, A. and Müller, K.-R. (1998) Neural Comput. 10, 1299–1319.

[13] Bicciato, S., Pandin, M., Didone, G. and Bello, C.D. (2003) Biotechnol. Bioeng. 81, 594–606.

[14] Ben-Dor, A., Bruhn, L., Friedman, N., Nachman, I., Schummer, M. and Yakhini, Z.J. (2000) Comput. Biol. 7, 559–583.

[15] Lee, K.E., Sha, N., Dougherty, E.R., Vannucci, M. and Mallick, B.K. (2003) Bioinformatics 19, 90–97.

[16] Baba, Y., Nonoyama, S., Matsushita, M., Yamadori, T., Hashimoto, S., Imai, K., Arai, S., Kunikata, T., Kurimoto, M., Kurosaki, T., Ochs, H.D., Yata, J., Kishimoto, T. and Tsukada, S. (1999) Blood 93, 2003–2012.

[17] Cory, G.O., MacCarthy-Morrogh, L., Banin, S., Gout, I., Brickell, P.M., Levinsky, R.J., Kinnon, C. and Lovering, R.C. (1996) J. Immunol. 157, 3791–3795.

[18] Shcherbina, A., Rosen, F.S. and Remold-O'Donnell, E. (1999) J. Immunol. 163, 6314–6320.

[19] Cunningham-Rundles, C., Lieberman, P., Hellman, G. and Chaganti, R.S. (1991) Am. J. Hematol. 37, 69–74.

[20] Kersey, J.H., Shapiro, R.S. and Filipovich, A.H. (1988) Pediatr. Infect. Dis. J. 7, S10–S12.