Gene selection and classification from microarray data using kernel machine

Ji-Hoon Cho[a,1], Dongkwon Lee[b,2], Jin Hyun Park[c,1], In-Beum Lee[a,*]

[a] *Department of Chemical Engineering, Pohang University of Science and Technology, San 31 Hyoja-Dong, Pohang 790-784, Republic of Korea*
[b] *Chemicals & Polymer R&D, LG Chem. Ltd., Hwachi-dong, Yeosu 555-280, Republic of Korea*
[c] *Bioinformatics Laboratory, P&I Consulting Co., Ltd., San 31 Hyoja-Dong, Pohang 790-784, Republic of Korea*

Received 5 February 2004; revised 18 May 2004; accepted 18 May 2004

Available online 6 July 2004

Edited by Lukas Huber

Abstract The discrimination of cancer patients (including subtypes) based on gene expression data is a critical problem with clinical ramifications. Central to solving this problem is the issue of how to extract the most relevant genes from the several thousand genes on a typical microarray. Here, we propose a methodology that can effectively select an informative subset of genes and classify the subtypes (or patients) of disease using the selected genes. We employ a kernel machine, kernel Fisher discriminant analysis (KFDA), for discrimination and use the derivatives of the kernel function to perform gene selection. Using a modified form of KFDA in the minimum squared error (MSE) sense and the gradients of the kernel functions, we construct an effective gene selection criterion. We assess the performance of the proposed methodology by applying it to three gene expression datasets: leukemia dataset, breast cancer dataset and colon cancer dataset. Using a few informative genes, the proposed method accurately and reliably classified cancer subtypes (or patients). Also, through a comparison study, we verify the reliability of the gene selection and discrimination results.

Keywords: Gene expression data; Gene selection; Classification; Kernel Fisher discriminant analysis

1. Introduction

The development of microarray technology, which enables simultaneous monitoring of several thousand genes, has revolutionized biological and medical research [1–6]. Microarray data can be used to gain molecular-level insight into phenomena in the human body such as the mechanism of cancer progression. However, appropriate data analysis techniques are needed if we are to extract useful information from such large-scale gene expression measurements. For example, unsupervised learning methods have been developed for exploratory subtype discovery of cancer, and supervised methods can be used to find cancer-specific genes that may be candidates for drug targeting, to develop a diagnostic system for cancer classification, and so on. However, the high dimensionality of microarray data can lead to problems such as the curse of dimensionality and singularity problems in matrix computations, making data analysis difficult [7–10]. Furthermore, even if it were possible to handle such huge data sets, the problem remains of extracting valuable information from data on thousands of genes. Therefore, there is a pressing need for techniques capable of selecting the subset of genes relevant to a particular problem from among the entire set of microarray data. Recently, the issue of gene selection has become a central challenge in the field of microarray data analysis and has been the subject of numerous studies [11–15]. Gene selection methods generally fall into one of the two categories: filter and wrapper approaches. Filter methods rank genes according to some pre-defined criterion; for example, statistical tests such as the T-test, F-test, and Wilcoxon's ranksum test are typical filter techniques. These techniques have been widely used because they are easy to understand and implement [11,16,17]. The wrapper approach, in contrast, is more complex because it requires a trained learning machine that can evaluate the relevance of a selected subset of genes. The wrapper approach finds a subset of genes and estimates its relevance using a machine like classifier. According to a criterion such as a cross-validation error rate, the wrapper updates the subset of genes iteratively. Support vector machine (SVM)-RFE [13] is a good example of a wrapper method for gene selection. It is generally accepted that wrapper methods are usually superior to filter methods because they can consider inter-correlation of individual variables (genes) in a multivariate manner and, moreover, they can determine the optimal number of variables for a particular machine [9,18].

In the present study, we propose a gene selection procedure for classification that uses a kernel Fisher discriminant analysis (KFDA) which showed outstanding performance [19] as the wrapper and a criterion for the machine. Previously, we suggested a classification and gene selection method that uses KFDA [15]; however, it may be regarded as a compromise between the filter and wrapper approaches. In this work, we aimed to develop a superior method that would be purely based on the wrapper approach, and that would therefore have all of the advantages inherent to that approach. Kernel machines have been used in classification including gene selection. They have the advantage that they work well regardless of the data dimension (i.e., the number of genes) and hence are well suited to the analysis of high-dimensional data. However, previous applications of kernel machines and research into variable

* Corresponding author. Fax: +82-54-279-3499.
E-mail address: iblee@postech.ac.kr (I.-B. Lee).

[1] Fax: +82-54-279-3499.
[2] Fax: +82-61-680-6015.

0014-5793/$22.00 © 2004 Published by Elsevier B.V. on behalf of the Federation of European Biochemical Societies.
doi:10.1016/j.febslet.2004.05.087

selection have been mostly confined to conventional SVM [13,20]. We use KFDA as a wrapper because it not only has the aforementioned advantages of kernel machine but also is simpler than SVM because of not solving a constrained optimization problem. Here, we develop a criterion for gene selection using a KFDA classifier and test the performance of the proposed methodology by applying it to three microarray datasets.

In the next section, we present a brief description of the KFDA algorithm and a minimum squared error (MSE)-based framework for KFDA. We define a new gene selection criterion based on the existing formalism of KMSE framework for KFDA and the gradient method proposed by Rakotomamonjy [21]. In addition, we improve the robustness of the criterion by using ranking information based on the expression values of each gene.

2. Materials and methods

Note that KFDA is a non-linear classifier that can outperform when the linear methods fail because of the non-linear properties in the data structure or sample distributions. Here, we described the algorithm of KFDA briefly. Further detailed information about KFDA (e.g., advantages and disadvantages) can be found in Mika et al. [19].

2.1. MSE approach in kernel Fisher discriminant analysis

KFDA was originally suggested by Mika et al. [19]. The derivation of KFDA is similar to that of conventional Fisher discriminant analysis (FDA), except that the mathematical operations are performed in a different space. We consider a binary classification. Let $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_{l_1}, \mathbf{x}_{l_1+1}, \mathbf{x}_{l_1+2}, \ldots, \mathbf{x}_{l_1+l_2}\}$ be the given data matrix, where $\mathbf{x}_j \in R^N$ denotes the jth sample vector, l_1 and l_2 ($l_1 + l_2 = l$) are the number of samples in class 1 and 2, respectively. The complete KFDA classifier is expressed as:

$$f(\mathbf{x}) = \mathbf{K}^{\mathrm{T}}\boldsymbol{\alpha} + b \tag{1}$$

where $\mathbf{K}_{jk} = k(\mathbf{x}_j, \mathbf{x}_k) = \Phi(\mathbf{x}_j) \cdot \Phi(\mathbf{x}_k)$, $\boldsymbol{\alpha}$ is a coefficient vector and b is a bias term. Notice that $\Phi(\mathbf{x})$ is the implicit non-linear mapping function used in kernel machines.

Xu et al. [22] showed that the KFDA classifier can be re-written in the MSE sense as the following set of linear equations:

$$\begin{bmatrix} \mathbf{K}\mathbf{K}^{\mathrm{T}} + \mu\mathbf{I} & \mathbf{K}\mathbf{U} \\ (\mathbf{K}\mathbf{U})^{\mathrm{T}} & l \end{bmatrix} \begin{bmatrix} \boldsymbol{\alpha} \\ b \end{bmatrix} = \begin{bmatrix} \mathbf{K}\mathbf{y} \\ \mathbf{U}^{\mathrm{T}}\mathbf{y} \end{bmatrix} \tag{2}$$

where j and $k = 1, \ldots, l$, $\mathbf{U}$ is a column vector of l ones, μ is a regularization parameter, and $\mathbf{y}$ is a coded output. Obviously, the MSE solution depends on the output coding schemes and there are many possible choices of coding output [22]. We used the same output coding as Xu et al. for binary classification. That is,

$$y_j = \begin{cases} l/l_1 & \text{if the } j\text{th sample belongs to class 1} \\ -l/l_2 & \text{if the } j\text{th sample belongs to class 2} \end{cases} \tag{3}$$

In this case, the complete KFDA classifier solution can be obtained as follows:

$$\hat{\mathbf{y}} = \mathbf{K}^{\mathrm{T}}\boldsymbol{\alpha} + \mathbf{U}b \tag{4}$$

where $\boldsymbol{\alpha} = (\mathbf{K}\mathbf{K}^{\mathrm{T}} + \mu\mathbf{I} - l^{-1}\mathbf{K}\mathbf{U}\mathbf{U}^{\mathrm{T}}\mathbf{K}^{\mathrm{T}})^{-1}(\mathbf{K}\mathbf{y} - l^{-1}\mathbf{K}\mathbf{U}\mathbf{U}^{\mathrm{T}}\mathbf{y})$ and $b = l^{-1}(\mathbf{U}^{\mathrm{T}}\mathbf{y} - (\mathbf{K}\mathbf{U})^{\mathrm{T}}\boldsymbol{\alpha})$.

Notice that the above solution minimizes the squared error, E, between the coded output and estimated output:

$$E = \frac{1}{2}\left(\mathbf{y} - \hat{\mathbf{y}}\right)^{\mathrm{T}}\left(\mathbf{y} - \hat{\mathbf{y}}\right) = \frac{1}{2}\left(\mathbf{y} - \mathbf{K}^{\mathrm{T}}\boldsymbol{\alpha} - \mathbf{U}b\right)^{\mathrm{T}}\left(\mathbf{y} - \mathbf{K}^{\mathrm{T}}\boldsymbol{\alpha} - \mathbf{U}b\right) \tag{5}$$

The proof that the above classifier is equivalent to the conventional KFDA is provided in Xu et al. [22].

2.2. Gene selection criterion

Recently, Rakotomamonjy [21] proposed a new variable selection method for a support vector machine (SVM) classifier using a gradient of the kernel function. This method has the advantage that, compared to SVM-RFE [13], it is significantly less complex computationally because it uses derivatives of a kernel function and hence calculates the Gram matrix $\mathbf{K}$ only once during the evaluation of the relevance of each variable. This reduction in computational complexity facilitates the analysis of high-dimensional data. Rakotomamonjy showed that the derivative of the Gram matrix ($\mathbf{K}$) can be directly connected with the gradient of the weight vector in the SVM classifier, and thus that these derivates can be used to identify the most relevant variable, that is, the variable that maximizes the variation of the squared norm of the weight vector. Here we adopt a similar approach to develop a new gene selection criterion for the KFDA classifier. Contrary to the previous works [13,21], the key idea of our criterion is that the variable (gene) which has a large influence on the output error (not the weight vector of classifier) is relevant to the classification. Since the linear set of equations for KFDA came from the objective function in the MSE sense, we created a criterion using the squared error. The relevance index for the ith gene, R_i, is defined as

$$R_i = \left|\frac{\partial E}{\partial v_i}\right| = \frac{1}{2}\left|\frac{\partial\left(\mathbf{y} - \mathbf{K}^{\mathrm{T}}\boldsymbol{\alpha} - \mathbf{U}b\right)^{\mathrm{T}}\left(\mathbf{y} - \mathbf{K}^{\mathrm{T}}\boldsymbol{\alpha} - \mathbf{U}b\right)}{\partial v_i}\right| \tag{6}$$

where v_i is an indicative variable that represents the ith gene.

If gene i is highly relevant to the classification, it will significantly affect the error gradient and thus R_i will have a large value. In practice, the relevance index R_i can be computed by the following equation:

$$R_i = \frac{1}{2}\left|\boldsymbol{\alpha}^{\mathrm{T}}\left\{(\mathbf{D}_i \otimes \mathbf{K})\mathbf{K}^{\mathrm{T}} + \mathbf{K}(\mathbf{D}_i \otimes \mathbf{K})^{\mathrm{T}}\right\}\boldsymbol{\alpha} - 2\boldsymbol{\alpha}^{\mathrm{T}}(\mathbf{D}_i \otimes \mathbf{K})\mathbf{Y} + 2\boldsymbol{\alpha}^{\mathrm{T}}(\mathbf{D}_i \otimes \mathbf{K})\mathbf{U}b\right| \tag{7}$$

where $\otimes$ is a component-wise product operation and $\mathbf{D}_i$ is a sample distance matrix for the ith gene, defined as:

$$\mathbf{D}_i = \begin{bmatrix} (x_{i1} - x_{i1})^2 & (x_{i1} - x_{i2})^2 & \cdots & (x_{i1} - x_{il})^2 \\ \vdots & \vdots & \ddots & \vdots \\ (x_{il} - x_{i1})^2 & (x_{il} - x_{i1})^2 & \cdots & (x_{il} - x_{il})^2 \end{bmatrix} (l \times l \text{ matrix}) \tag{8}$$

The detailed derivation of relevance index R_i is provided as supplementary information on our website (http://home.postech.ac.kr/~cjhjhj/supp_main.htm). The proposed method has an advantage that it can significantly reduce the computational load because it does not require recalculation of the Gram matrix $\mathbf{K}$ and the solution of the classifier when we evaluate the sensitivity of each gene by removing it from a certain gene set. Instead of the complex calculation of $\mathbf{K}$, we need to only compute the sample distance matrix $\mathbf{D}_i$ of each gene whenever a gene is removed. So far, we have developed a new gene selection criterion for the KFDA classifier based on the derivatives of the kernel function. However, there is a problem related to the sample distance matrix $\mathbf{D}_i$. Gene expression datasets typically contain numerous extreme values caused by artifacts such as systematic noise in the laboratory. These extreme values can distort the gene selection procedure. Specifically, a gene with no distinct pattern across classes, and thus a small value of the error gradient, $|\partial E/\partial v_i|$, may be assigned an unrealistically high value of R_i if the value in the matrix $\mathbf{D}_i$ is extremely large. To prevent this situation, we constructed the sample distance matrix $\mathbf{D}_i$ using the ranking information instead of the real expression values because the ranking information provides a robust similarity/dissimilarity measure, as shown in non-parametric statistical methods such as the Wilcoxon's ranksum test and Kruskal–Wallis test.

2.3. Biological data

We used three publicly available microarray datasets. The leukemia dataset, which was produced for the classification of acute lymphoblastic leukemia (ALL) and acute myeloid leukemia (AML), consists of 7129 probes and 72 samples [11]. The breast cancer dataset for the discrimination between BRCA1 mutation and others (seven BRCA1 mutation samples and 15 BRCA2 mutation and sporadic samples) has 3226 probes and 22 samples [23]. The colon cancer dataset for the diagnosis of cancer patients consists of 2000 probes and 62 samples (40 cancer tissues and 22 normal tissues) [1]. Note that all datasets address binary classification problems. The reason we only consider for the binary classification problem will be described in Section 4.

3. Results

We used the proposed method to analyze the leukemia, breast cancer and colon cancer datasets. We monitored the

squared error between the coded output and actual output of the KFDA classifier and iteratively removed a proportion, δ $(0 \leqslant \delta \leqslant 1)$ (in this paper, $\delta = 0.2$), of the genes with the smallest values of R_i until 100 genes remained. After the pool of relevant genes had been reduced to 100, genes were removed one at a time. The training and test datasets were obtained by splitting the total dataset into two parts containing 80% and 20% of the total samples (similar to fivefold cross-validation), respectively, while ensuring that the proportion of classes was balanced in the two sets. Considering the arbitrariness of this partitioning, we repeated the above cross-validation procedure 100 times, as in our previous work [15]. By this approach, we obtained the point that represents the minimum of the mean test error (mean value of 100 test errors) and indicates the optimal number of genes, m (i.e., the number that produces the minimum error). Once the minimum error point was obtained, we identified the m genes that were most frequently used during 100 cross-validations at that point. For the KFDA classifier, it is necessary to determine the parameters for the kernel function and regularization of inverse operation. The choice of optimal parameters for kernel machines has been extensively researched in the field of machine learning. Various methods (e.g., L-curve and generalized cross-validation) have been suggested [24,25], but a detailed discussion of these methods is beyond the scope of the present study. In the present work, through cross-validation, we empirically determined the values of σ of the kernel function and μ, the regularization parameter that produced the smallest cross-validation error. This approach yielded values of σ which equals five times the number of genes used in the classifier at each repetition, and μ equals unity.

3.1. Gene selection and cross-validation results

The leukemia dataset consists of training and test sets comprising 38 and 34 samples, respectively. In this study, we combined these two datasets into a single set of 72 samples; thus, about 58 randomly selected samples were used for training in each cross-validation step. The cross-validation result is illustrated in Fig. 1(a). Note that the error values do not represent a misclassification error rate but rather the root mean squared error between $\mathbf{Y}$ and $\hat{\mathbf{Y}}$. From Fig. 1(a), we recognize that the optimal number of genes, which produces the minimum test error, is 17. As stated before, we can obtain an optimal subset of genes by finding 17 genes that are most frequently used over 100 cross-validations at the minimum error point. In fact, the selected genes are not always used together during the cross-validations and thus it is hard to evaluate the misclassification rate within the gene selection procedure. Therefore, we regard the selected gene subset as an independent dataset and verify the performance of it in the context of classification. After the identification of 17 genes, we newly make a KFDA classifier and evaluate the classification ability using leave-one-out cross-validation (LOOCV). Such a validation scheme has been already used in the previous work [26]. The assignment of sample to the corresponding class is achieved by logistic regression [27]. Our selected genes for the leukemia dataset produce one misclassification. For breast cancer dataset, the cross-validation result for the gene selection is depicted in Fig. 1(b). Following an approach similar to that used for the leukemia dataset, we extract 21 genes that achieve the minimum CV test error. In the validation procedure (LOOCV-classification), the selected genes produce no misclassification. The cross-validation result for the colon cancer dataset is illustrated in Fig. 1(c). We identified 20 genes to discern cancer tissues from normal ones. The selected subset of genes produced six misclassifications during LOOCV procedure.

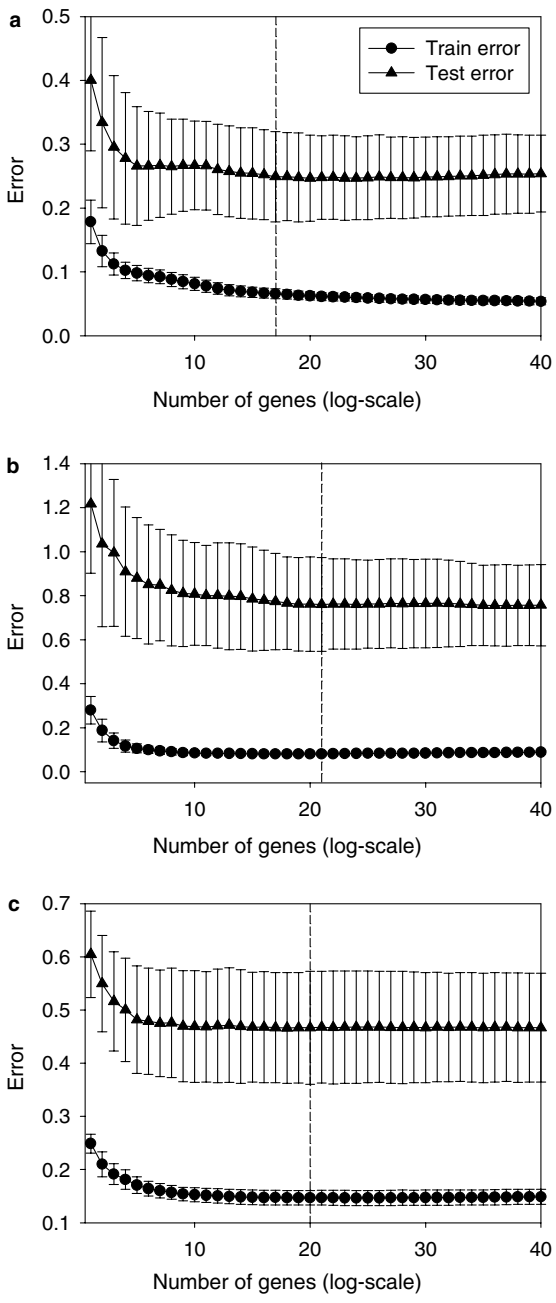

Fig. 1. Cross-validation results of three datasets. (a) leukemia dataset: at the minimum test error point (indicated by dashed line), 17 genes are used for classifier (b) breast cancer dataset: 21 genes and (c) colon cancer dataset: 20 genes.

3.2. Comparison study

To evaluate the performance of the proposed gene selection and classification method, we compared our method with previously developed methods. Although it is somewhat difficult to directly compare these methods because they each use a different criterion, we performed a comparison study by

reproducing other methods. Recent studies related to leukemia (Guyon et al. [13]), breast cancer (Lee et al. [26]), colon cancer (Li et al. [14]) and our previous work (Cho et al. [15]) were chosen as targets for comparison. In studies using the same leukemia dataset as that used here, Guyon et al. extracted four genes and Cho et al. six genes for subtype classification. First, we constructed a support vector classifier with exactly the same conditions as those described by Guyon et al. for their gene subset. Since Guyon et al. [15] found some informative genes using 38 training samples, we newly selected gene subsets and constructed KFDA classifiers by applying our previous method and the proposed one to the 38 training samples. Due to the change of training dataset, the selected genes are slightly different from the ones configured in the previous work and Section 3.1. Note that the classifier used in our previous research was a conventional KFDA algorithm; thus, it is conceptually different from the classifier proposed in the present work, which is modified in the MSE sense. For comparison, with respect to each method, we obtained the score values of 34 test samples and the estimated probabilities (i.e., posterior probabilities) computed by logistic regression.

The breast cancer dataset was previously analyzed by Lee et al. [26] using Bayesian approach. They demonstrated the validity of their subset of genes by LOOCV after the completion of gene selection procedure. Therefore, we merely compared our LOOCV result with theirs. For our previous method, a gene subset is selected using all samples and the posterior probabilities are obtained by LOOCV and logistic regression. Although this validation, in fact, is not appropriate since information of the "left out" feature in the LOOCV has already been used to select the optimal number of genes [28], we used it only for comparison.

For the colon cancer dataset, Li et al. [14] reported the average performance over 100 random partitions into 50 training and 12 test samples. We exactly followed their validation procedure and obtained average performances of our previous method and the proposed one.

Table 1 shows the comparison results. For the leukemia dataset, the method of Guyon et al. produced only no misclassification, while the proposed method and our previous method produced some misclassifications. At first glance, the lower number of misclassifications produced by the method of Guyon et al. [29,30] would seem to suggest that this method is superior to the other methods considered; however, it is well known that the leukemia dataset contains at least one sample that is mislabeled and that may influence the error rate. Thus, the present findings indicate that the gene subset of Guyon et al. is slightly lacking in the ability to detect intrinsic data faults (i.e., mislabeled samples) and will therefore produce incorrect results due to excessive reduction of the number of genes. Our previous method shows a poor classification result as shown in Table 1(a), since it finds only one gene because of its stringent criterion (select genes always included in classifier during 100 cross-validations) [15] and thus the classifier becomes sensitive to small perturbation in test data. Contrary to the above two methods, the proposed method not only identifies the incorrect sample but also gives estimated probabilities of close to unity, the value for a correctly classified sample. The present results therefore indicate that, compared to previous analysis techniques, the proposed method more effectively finds the intrinsic property of the data and provides a clearer classification result.

Table 1
Comparison results

Methods	Misclassifications	Size of subset
(a) Leukemia dataset		
Guyon et al. [13]	0	4
Cho et al. [15]	9	1
Proposed	2	17
(b) Breast cancer dataset		
Lee et al. [26]	0	27, 17, 10
Cho et al. [15]	5	3
Proposed	0	21
(c) Colon cancer dataset		
Li et al. [14], algorithm 1	2.04 ± 0.14	15.13 ± 0.31
Li et al. [14], algorithm 2	2.90 ± 0.13	8.55 ± 0.13
Cho et al. [15]	2.57 ± 1.76	– (29)
Proposed	2.15 ± 1.2	10

For the breast cancer dataset, the proposed method shows a satisfactory classification result as shown in Table 1(b). Like the proposed method, the method of Lee et al. also produces zero or one misclassification over three models that have 27, 17 and 10 genes, respectively. It is hard to address which method is superior to the other in this case, however, we can see that the proposed method, which is obviously simpler than Lee et al. (using Bayesian mixtures and Markov Chain Monte Carlo computation), keeps abreast of the highly sophisticated one.

For the colon cancer dataset, the previous research reported average performance over 100 random partitions into 50 training and 12 test samples. Thus, we exactly follow such a validation procedure when we implement our previous method and the proposed one for comparison. Table 1(c) shows the average number of misclassifications and feature set size. Our previous method tells that the minimum average test error is obtained when 29 genes are used for classifier, however, there is no gene which always participates in the classifier construction. This fact means again that our previous criterion is too stringent. Moreover, it reflects that the filter approach is inappropriate for analyzing the colon cancer dataset, since there are few genes that consistently appear to be relevant according to the variation of training dataset. Considering the classification ability and the size of subset, the proposed method can bear comparison with the two algorithms of previous work [14].

3.3. Biological analysis of the selected genes

Many studies have been carried out on the leukemia dataset considered here [11–13,31,32] and most of the genes we found are part of previously chosen ones. We did, however, find an interesting gene, nucleoside-diphosphate kinase (NM23-H4, Y07604) which had been selected in our previous work [12]. Based on Arthur and Bloomfield [33], other researches [34,35] and the location of NM23-H4 (16p13), we carefully supposed that it might be a strong candidate for the AML diagnostic marker. For breast cancer dataset, we found some genes, keratin 8 (IMAGE ID: 897781) and transducer of ERBB2, 1 (IMAGE ID: 823940), of which the importance had been already revealed in Lee et al. [26]. Based on the fact that more than 80% of the genes selected in the present work are the same as those selected in the work of Hedenfalk et al. [23] and Lee et al. [26], we can see that our method reliably extracts informative genes in consistent with the previous researches. From the gene selection result of colon cancer dataset, we found

vascular endothelial growth factor (VEGF, IMAGE ID: 47326). Evidence from preclinical and clinical studies indicates that vascular endothelial growth factor (VEGF) is the predominant angiogenic factor in human colorectal cancer and is associated with the formation of metastases and poor prognosis [36]. Although it is clinically important and the univariate expression pattern of it is discriminative, it has not been selected as a relevant gene in previous researches [13,14]. The list and heat map of the selected subsets are available on our website (http://home.postech.ac.kr/~cjhjhj/supp_main.htm).

4. Discussion

In this paper, we have proposed a gene selection and classification method that utilizes a KFDA classifier and the derivative of the kernel function. We used a MSE framework of KFDA and constructed a gene selection criterion based on that MSE scheme. Since the MSE framework of KFDA can be regarded as a regression of typical KFDA scores to the coded output, and we focus on the error, our method can produce discriminative scores that are pushed away to fit the class-indicative **y**. Most previous studies of gene selection and subtype classification have adopted a criterion based on misclassification rates, with little consideration given to the reliability of classification, i.e., the extent to which scores (or posterior probabilities) are separated between classes. It should be borne in mind, however, that, even if a method extracts a small subset of genes without any misclassification, its outputs may lie in the marginal region. Moreover, it is possible that the small subset of selected genes will contain no genes of biological (clinical) importance because the gene selection process is a purely data-driven procedure that does not take into account prior biological knowledge. In fact, the number of genes selected by the proposed method is larger than previous ones, but it provided more accurate and reliable classification results. In addition, the proposed method identified some interesting genes unnoticed in previous studies.

The characteristics of our method can be adjusted through the selection of the output coding scheme [22]. If the objective of gene selection is to drastically reduce the number of genes, for example in a commercial diagnosis system, we can achieve this by basing the criterion (and output coding) on the misclassification rate. On the other hand, if the objective is to support further biological and medical research by selecting a subset of genes that has considerable relevance, we could use the error between the outputs to ensure that possible candidates are not missed.

In this paper, we have considered a binary classification problem only. Practically, we have more chance to encounter multiple classification problems and thus it is necessary to extend such a method to the multi-class case. However, it is difficult to obtain a general and non-trivial solution of multiple discriminant analysis (MDA) without solving an eigenvalue problem. To use the proposed gene selection criterion, we need an objective function (to be differentiated, E in our study) which should be directly connected with raw data, i.e., $\mathbf{x}_i$, however we cannot trace the relationship with the raw data from eigenvectors. The extension of our method into MDA solution is in progress and will be our next research topic.

The ultimate goal of this work is to identify a set of candidate genes that are worthy of analysis for purposes such as elucidating the mechanism of a disease, constructing a diagnosis system, and developing drugs. The benefits of employing the present method to study a disease would be enhanced by using it in conjunction with biological (clinical) experiments related to the disease.

Acknowledgements: This work was supported by Brain Korea 21 project. We thank Prof. Alain Rakotomamonjy (INSA de Rouen, France) for the fruitful help in developing the proposed KFDA algorithm. We also thank three anonymous referees for instructive comments.

References

[1] Alon, U., Barkai, N., Notterman, D.A., Gish, K., Barra, Y., Mack, D. and Levine, A.J. (1999) Proc. Natl. Acad. Sci. USA 96, 6745–6750.
[2] Chu, S., DeRisi, J., Eisen, M., Mulholland, J., Botstein, D., Brown, P.O. and Herskowitz, I. (1998) Science 282, 699–705.
[3] Eisen, M.B., Spellman, P.T., Brown, P.O. and Botstein, D. (1998) Proc. Natl. Acad. Sci. USA 95, 14863–14868.
[4] Iyer, V.R., Eisen, M.B., Ross, D.T., Schuler, G., Moore, T., Lee, J.C.F., Trent, J.M., Staudt, L.M., Hudson Jr., J., Boguski, M.S., Lashkari, D., Shalon, D., Botstein, D. and Brown, P.O. (1999) Science 283, 83–87.
[5] Lockhart, D.J., Dong, H., Byrne, M.C., Follettie, M.T., Gallo, M.V., Chee, M.S., Mittmann, M., Wang, C., Kobayashi, M., Horton, H. and Brown, E.L. (1996) Nat. Biotechnol. 14, 1675–1680.
[6] Celis, J.E., Kruhøffer, M., Gromova, I., Frederiksen, C., Østergaard, M., Thykjaer, T., Gromov, P., Yu, J., Pálsdóttir, H., Magnusson, N. and Ørntoft, T.F. (2000) FEBS Lett. 480, 2–16.
[7] Brazma, A. and Vilo, J. (2000) FEBS Lett. 480, 17–24.
[8] Duda, R.O., Hart, P.E. and Stork, D.G. (2001) Pattern Classification, second ed. Wiley, New York.
[9] Lu, Y. and Han, J. (2003) Inf. Syst. 28, 243–268.
[10] Sherlock, G. (2000) Curr. Opin. Immunol. 12, 201–205.
[11] Golub, T.R., Slonim, D., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J., Coller, H., Loh, M., Downing, J., Caligiuri, M., Bloomfield, C. and Lander, E. (1999) Science 286, 531–537.
[12] Cho, J.-H., Lee, D., Park, J.H., Kim, K. and Lee, I.-B. (2002) Biotechnol. Prog. 18, 847–854.
[13] Guyon, I., Weston, J., Barnhill, S. and Vapnik, V. (2002) Mach. Learn. 46, 389–422.
[14] Li, Y., Campbell, C. and Tipping, M. (2002) Bioinformatics 18, 1332–1339.
[15] Cho, J.-H., Lee, D., Park, J.H. and Lee, I.-B. (2003) FEBS Lett. 551, 3–7.
[16] Dudoit, S., Yang, Y.H., Speed, T.P. and Callow, M.J. (2002) Stat. Sinica. 12, 111–139.
[17] Keller, A.D., Schummer, M., Hood, L., and Ruzzo, W.L. (2000) Technical Report University of Washington UW-CSE-2000-08-01.
[18] Kohavi, R. and John, G. (1997) Artif. Intell. 97, 273–324.
[19] Mika, S., Rätsch, G., Weston, J., Schölkopf, B. and Müller, K. (1999) Proceedings of the IEEE Neural networks for signal processing workshop, pp. 41–48.
[20] Furey, T.S., Cristianini, N., Duffy, N., Bednarski, D.W., Schummer, M. and Haussler, D. (2000) Bioinformatics 16, 906–914.
[21] Rakotomamonjy, A. (2003) J. Mach. Learn. Res. 3, 1357–1370.
[22] Xu, J., Zhang, X. and Li, Y. (2001) Proc. of IJCNN-2001, pp. 1486–1491.
[23] Hedenfalk, I., Duggan, D., Chen, Y., Radmacher, M., Bittner, M., Simon, R., Meltzer, P., Gusterson, B., Esteller, M., Kallioniemi, O.-P., Wilfond, B., Borg, A. and Trent, J. (2001) New Engl. J. Med. 344, 539–548.
[24] Chapelle, O., Vapnik, V., Bousquet, O. and Mukherjee, S. (2002) Mach. Learn. 46, 131–159.
[25] Vapnik, V. (1998) Statistical Learning Theory. Wiley, New York.
[26] Lee, K.E., Sha, N., Dougherty, E.R., Vannucci, M. and Mallick, B.K. (2003) Bioinformatics 19, 90–97.

[27] Hosmer, D.W. and Lemeshow, S. (2000) Applied Logistic Regression, second ed. John Wiley and Sons, New York.
[28] Ambroise, C. and McLachlan, G.J. (2002) Proc. Natl. Acad. Sci. USA 99, 6562–6566.
[29] Fridlyand, J., Dudoit, S. and Speed, T.P. (2002) J. Am. Stat. Assoc. 97, 77–87.
[30] Chow, M.L., Moler, E.J. and Mian, I.S. (2001) Physiol. Genomics 5, 99–111.
[31] Nguyen, D.V. and Rocke, D.M. (2002) Bioinformatics 18, 39–50.
[32] Ben-Dor, A., Bruhn, L., Friedman, N., Nachman, I., Schummer, M. and Yakhini, Z. (2000) J. Comput. Biol. 7, 559–583.
[33] Arthur, D. and Bloomfield, C.D. (1983) Blood 61, 994–998.
[34] Magyarosy, E., Sebestyen, A. and Timar, J. (2001) Anticancer Res. 21, 819–823.
[35] Niitsu, N., Okabe-Kado, J., Nakayama, M., Wakimoto, N., Sakashita, A., Maseki, N., Motoyoshi, K., Umeda, M. and Honma, Y. (2000) Blood 96, 1080–1086.
[36] Guba, M., Seeliger, H., Kleespies, A., Jauch, K.W. and Bruns, C. (2004) Int. J. Colorectal. Dis. (in press).