# 암 분류 모델을 위한 marker genes 선별 기준의 비교

이민영, 황대희[1], 이인범, 한종훈[*2]
포항공과대학교 화학공학과, [1]The Institute for Systems Biology, [2]서울대학교 응용화학부
(chhan@snu.ac.kr[*])

# Comparative studies of marker genes selection criteria for cancer classification

Min-Young Lee, Daehee Hwang[1], In-Beum Lee, Chonghun Han[*2]
Department of Chemical Engineering, Pohang University of Science and Technology
[1]The Institute for Systems Biology
[2]School of Chemical Engineering, Seoul National University
(chhan@snu.ac.kr[*])

## Introduction

Microarray experiments allow the observation of overall gene expression profiling changes in the cell. These experiments give great insight for the discrimination of cancer subtypes because they have fundamentally different gene regulation patterns although different tumors have similar appearance and clinical behavior. Hence various discriminating methods using data mining techniques were developed for cancer diagnosis.

Most of the class prediction procedures were composed of two steps. First, a subset of marker genes is identified, and then with these selected genes, a discrimination rule is driven. As significant genes for classification, genes highly correlated with each other within the same class or those with statistically significant difference between classes are selected. In order to evaluate the discriminating rule, supervised classification algorithms have been used.

For the correct diagnosis, a selection of marker genes with good discrimination power is crucial before making a classifier. Various selection methods have been proposed, and they have been tested in classification accuracy, computational time, and robustness of noise. On that ground, it is generally believed that there is no universal criterion that is superior over the rest [1-3]. Nevertheless, many researchers are trying to find a powerful selection criterion which can be generally applied in order to save time and cost for making a good classifier.

In this work, we compared recent criteria for marker genes selection. New metric [4], SVMs/GAs [5], and MAximal MArgin Linear Programming (MAMA) method [6] were examined. They have differences compared with conventional methods in their aspects of selection mechanism. They are not based on differences of gene expression level. The details are described below.

## Method

### New metric

Suppose we have P genes, N samples, K classes. For gene i, sample j, class k, the centroid and distance matrix **z**, of which each element $z_{ij}$ are obtained as follows.

$$\bar{x}_{ik} = 1/n_k \sum_{j \in C_k} x_{ij} \quad (1)$$

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

$\mathbf{z}_i$ (1 x n row vector) is the within-class vector of gene i.

For i gene,

$$mean_w(\mathbf{z}_i) = \sum_{j=1}^{n} \frac{w_j}{W} z_{ij} \quad (3)$$

$$std_w = \sqrt{\frac{\sum_{j=1}^{n}(z_{ij} - mean_w(\mathbf{z}_i))^2}{(n-1/n)\sum_{j=1}^{n} w_j}} \quad (4)$$

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

Then, New metric can be formulated like the following. This finds genes which have short distances from each class centroid and have simultaneously small variation within the class.

$$R_i = \frac{mean_w(\mathbf{z}_i)\cdot std_w(\mathbf{z}_i)}{std(\bar{x}_i)} \quad (5)$$

FDA and KFDA were used for classification accuracy test. With the weight vector, the discrimination function (score) and $\mathbf{y}_i$ (K-1 x 1 column vector) of input $\mathbf{x}_j$ (j=1,2, … , n) were calculated. Then, the chi-square distance of the jth sample from the centroid of each class was computed by

$$\chi^2_{j,k} = (\mathbf{y}_j - \bar{\mathbf{y}}_k)^T \mathbf{D}_k^{-1}(\mathbf{y}_j - \bar{\mathbf{y}}_k) \quad (6)$$

where $\mathbf{D}_k$ is the covariance matrix of $\mathbf{y}$ for class k and $\bar{y}_k = 1/n_k \sum_{j\in C_k} \mathbf{y}_j$ denotes a class centroid of discriminant score. The posterior probability is calculated as follows.

$$P(k \mid \mathbf{x}_j) = \frac{P_k \mid \mathbf{D}_k \mid^{-1/2} \exp(-\chi^2_{j,k}/2)}{\sum_{k'} P_{k'} \mid \mathbf{D}_{k'} \mid^{-1/2} \exp(-\chi^2_{j,k'}/2)} \quad (7)$$

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.

### SVMs/GAs

SVMs/GAs algorithm consists of six steps. First, a set of genes differentially expressed across all cancer types is filtered. Then AP/SVM classifier is evaluated for binary tumor classification, and based on that result, a voting scheme is followed to go from binary to multiclass classification. Next, GA feature selection and multiclass classification optimization via LOOCV fitness test are followed. Finally, in order to further eliminate the non-predictive features in the GA-derived gene set, RFE through AP/SVMs and LOOCV are tested.

### MAMA

MAMA algorithm considers simultaneously several gene expression profiles rather than single that. It defines a function that models in mathematical terms biological relationships among genes and thus reflects functional relations among them. This method consists of two stages. First, the initial input data are mapped into feature space F and ideal feature construction are formulated. Then the feature selection procedure that maximizes the margin of an ideal feature is evaluated. The ideal features are constructed in the following form.

$$\mathbf{u}_l = -\beta e + \sum_{i\in J_i} \alpha_i \log(\mathbf{x}_i) \quad (8)$$

where $\mathbf{u}_l$ is the ideal feature vector, Ji is the corresponding gene sets. $\mathbf{x}$ is the initial input data. For selection among probable features, the optimization procedure is constructed as follows. A multiplication of coefficients $\alpha_i$, some constant t provides

$\prod_{i\in J_A}(x_i^k)^{t\alpha_i} = e^{t(z_l+\beta)}$ , for each sample k from class A (9)

$\prod_{i\in J_A}(x_i^k)^{t\alpha_i} = e^{t\beta}$, for each sample n from class B

and $t(z_l+\beta)/t\beta=(z_l+\beta)/\beta$. For this reason one can consider the ratio $(z_l+\beta)/\beta$ as margin between class A and B and prefer features with minimal unity ($\beta\mathbf{e}$) component relative to z. If β is fixed, then this ratio is maximized by maximization of *l z*. This task can be implemented using linear programming (LP). LP practically has no restrictions on the size of the problem that can be solved. Consider the following optimization problem

max $z_l$<br>
α, β, **s**<br>
***subject to:***<br>
$-\beta\mathbf{e} + \Sigma_{i\subseteq Ji}\,\alpha_i\log(\mathbf{x}_i)+\mathbf{s} = \mathbf{u}_{l;}$<br>
$|s|\leq\varepsilon;\ \alpha_i \geq 0;\ \mathbf{s} \subseteq R^K;$

The objective of the above formulation is to find such a gene subspace $J_l$ where the margin between hyperplanes A and B is maximal.

Results

**New metric**

It was tested with two data. For binary classification example, leukemia data was used, and for multiclass example, SRBCT data was used. 6 genes for leukemia and 21 genes for SRBCT were selected at minimum LOOCV test error. Classification results were presented below.

Table1. Classification results of two data for New metric. Numerical values indicate the number of misclassifications.

| | Conventional FDA | KFDA | Number of genes |
|---|---|---|---|
| Leukemia data | | | |
| Golub et al. | 9 | 4 | 50 |
| Lee et al. | 5 | 3 | 5 |
| Proposed | 3 | 2 | 6 |
| SRBCT data | | | |
| Tibshirani et al. | 2 | 2 | 43 |
| Proposed | 0 | 0 | 21 |

**SVMs/GAs**

The followings are comparisons of GA/SVM with some other algorithms. The results were comparable or superior to those previous methods.

Table2. Classification results of two data for SVMs/GAs.

| | NCI60 data | | GCM data | |
|---|---|---|---|---|
| | LOOCV(%) | Number of genes | LOOCV(%) | Number of genes |
| Hierarchical clustering | 81 | 6831 | - | - |
| OVA/SVM | - | - | 78 | 16063 |
| OVA/SVM | - | - | 81.25 | 16063 |
| OVA/KNN | - | - | 72.92 | 16063 |
| GA/MLHD | 85.37 | 13 | 79.33 | 32 |
| GA/SVM/RFE | 87.93 | 27 | 85.19 | 26 |

**MAMA**

MAMA procedure was applied to the dataset on multiple tumor type classification (Ramaswamy et al., 2001) and the dataset on acute leukemia classification. First, the genes with the standard deviation of expression values across the training samples less than SD threshold were filtered. Then the dataset is split into training and test sets and classification is applied. The results are as follows.

Table3. Classification results of multiple tumor dataset (Ramaswamy et al., 2001) for different values of the filter threshold.

| Subset ID | SD threshold | Number of pre-selected genes | misclassifications | | Prediction rate | |
|---|---|---|---|---|---|---|
| | | | LOOCV | Test samples | LOOCV | Test samples |
| 1 | 1300 | 707 | 29 | 16 | 80% | 70% |
| 2 | 1100 | 905 | 28 | 16 | 81% | 70% |
| 3 | 1000 | 1042 | 27 | 14 | 81% | 74% |
| 4 | 900 | 1203 | 26 | 13 | 82% | 76% |
| 5 | 800 | 1445 | 25 | 8 | 83% | 85% |
| 6 | 700 | 1740 | 26 | 10 | 82% | 83% |

Table4. Classification results for the acute leukemia dataset Golub et al. (Golub et al., 1999)

| Subset ID | SD threshold | Number of pre-selected genes | misclassifications | | Prediction rate | |
|---|---|---|---|---|---|---|
| | | | LOOCV | Test samples | LOOCV | Test samples |
| 1 | 3000 | 132 | 2 | 0 | 95% | 100% |
| 2 | 2500 | 185 | 1 | 0 | 98% | 100% |
| 3 | 2000 | 273 | 3 | 0 | 92% | 100% |
| 4 | 1500 | 373 | 2 | 0 | 95% | 100% |
| 5 | 1000 | 549 | 2 | 0 | 95% | 100% |
| 6 | 500 | 1120 | 2 | 3 | 95% | 92% |

## Conclusion

Recent proposed methods are superior as compared with classical methods. New metric is more accurate than conventional methods, and does not depend on the sample size. SVMs/GAs significantly eliminates gene redundancy and yields a more compact and unique gene subset, but the performance is comparable or superior to conventional methods. MAMA predicts a class of samples which did not separate well with prior methods. But they should be tested in more data. They were examined only two data sets. Therefore which method can be used for universal cases is not determined yet. Via further research, a study of their merits and defects is needed.

## References

Szabo, A., Boucher, K., Carroll, W. L., Klebanov, L. B., Tsodikov, A. D., Yakovlev, A. Y., “Variable Selection and Pattern Recognition with Gene Expression Data Generated by the Microarray Technology”, MATH BIOSCI., 71, 176(2001).

Liang, J. and Kachalo, S., “Computational Analysis of Microarray Gene Expression Profiles: Clustering, Classification, and beyond”, CHEMOMETR INTELL LAB., 199, 62(2002).

Lu, Y. and Han, J., “Cancer Classification using Gene Expression Data”, INFORM SYST., 28, 243 (2003).

Cho, J. H., Lee, D. K., Park, J. H., Lee, I. B., “New Gene Selection Method for Classification of Cancer Subtypes Considering Within-Class Variation”, FEBS LETT., 551, 3 (2003).

Peng, S., Xu, Q., Ling, X. B., Peng, X., Du, W., Chen, L., “Molecular Classification of Cancer Types from Microarray Data using the Combination of Genetic Algorithm and Support Vector Machines”, FEBS LETT., 555, 58(2003).

Antonov, A. V., Tetko, I. V., Mader, M. T., Budczies, J., Mewes, H. W., “Optimization Models for Cancer Classification: Extracting Gene Interaction from Microarray Expression Data”, Bioinformatics, (2004). *in press*(available online)