Bioinformation by Biomedical Informatics Publishing Group *open access*

www.bioinformation.net **Prediction Model**

Accurate and robust gene selection for disease classification using a simple statistic

Hikaru Mitsubayashi[1], Seiichiro Aso[2], Tomomasa Nagashima[2] and Yoshifumi Okada[2,] *

[1]Division of Production and Information Systems Engineering, Muroran Institute of Technology, [2]Department of Computer Science and Systems Engineering, Muroran Institute of Technology; 27-1, Mizumoto, Muroran 050-8585, Japan;
Yoshifumi Okada* - Email: hikaru@epsilon2.csse.muroran-it.ac.jp; * Corresponding author

received September 05, 2008, accepted September 21, 2008; published October 24, 2008

Abstract:
Discrimination of disease patients based on gene expression data is a crucial problem in clinical area. An important issue to solve this problem is to find a discriminative subset of genes from thousands of genes on a microarray or DNA chip. Aiming at finding informative genes for disease classification on microarray, we present a gene selection method based on the forward variable (gene) selection method (FSM) and show, using typical public microarray datasets, that our method can extract a small set of genes being crucial for discriminating different classes with a very high accuracy almost closed to perfect classification.

Keywords: robust; gene selection; disease; DNA chip; microarray

Background:

On classification in pathological diagnoses from microarray data, feature selection or gene selection is an important step for extracting the genes that contribute to classification. Until now, various gene selection methods have been proposed. They fall into two categories, filter method and wrapper method. Filter method estimates the significance of expression difference between two classes (*e.g.*, normal or cancer sample) gene by gene **[1-5]**. Wrapper method extracts a set of genes being useful for classification by repeating the processes consisting of a combinatorial gene selection and discrimination by classifier such as SVM **[6, 7]**. In general, it is known that wrapper method requires much more computation time than filter method but shows higher classification accuracy **[8]**.

In this paper, we demonstrate that an accurate and robust classification results are achieved by extracting the set of informative genes to classify samples based on very simple statistics (F-value) We apply a forward variable (gene) selection method (FSM) based on Mahalanobis distance for extracting informative genes. FSM uses a very simple algorithm that conducts a greedy search based on a simple statistic, *i.e.*, F-value. While our method applies a recursive process similar to wrapper method, it classifies samples with different classes without repeating the complicated processes of gene selection and classification used in wrapper method. Here, we illustrate a performance of our method by applying to microarray datasets and making a comparison with other gene selection methods.

Methodology:

The calculation method for FSM is briefly described in supplementary material.

Results and discussion:

We apply FSM to five different publicly available microarray datasets, two leukemia **[1, 9]**, breast cancer **[10]**, colon cancer **[11]** and Medulloblastoma **[12]** datasets. Here we show the result on Golub's leukemia dataset (see Supplementary Data for the other datasets). This dataset contains 7,129 genes and 38 samples from 2 classes with 27 ALL (acute lymphoblastic leukemia) and 11 AML (acute myeloid leukemia).

To evaluate the performance of FSM, we conduct leave-one-out cross validation analysis (LOOCV) for the set of genes selected by FSM. In this analysis, each test sample (regarded as a vector consisting of expression values of each gene) is classified into a class to which the Maharanobis distance to the average expression vector is closer, and the number of "error samples" is counted.

The classification accuracy is evaluated by the error rate that is defined by dividing the number of "error samples" by the number of all samples, where error sample is judged using the extracted genes.

We first compare the performance of FSM to those of two filter methods, a ranking by F-value for every single gene (Simple) and a signal to noise metric-based ranking (S2N) **[1]**. It is shown the results of classification for the top 50 genes in Figure 1, where the horizontal axis and vertical axis indicate the cumulative number of genes ordered by F-value and the classification error rate, respectively.

In all methods including FSM, the error rate decreases around the top 10 genes, but it shows substantial fluctuations in the region beyond top 35 genes. This suggests that the gene set includes many uninformative or noisy genes for classification. However obvious differences of the accuracy are observed between FSM and the two filter methods in the region between around the top 10 and top 34 genes. Namely, the error rate of the filter methods grows with increasing the number of genes, and the tendency is very sensitive to the selected genes. In contrast, FSM gives robust and higher accuracy than these filter methods, i.e., all test samples are classified perfectly by the use of only top 8 genes, where the perfect classification continues up to 34 genes.

As seen in Supplementary Data, FSM also provides prominently superior accuracy for the other four datasets, e.g., perfect classifications are achieved on leukemia **[9]** and breast cancer **[10]**. Although the error rates for colon cancer **[11]** and Medulloblastoma **[12]** do not attain 0%, it is expected to improve by taking a larger number of genes for classification. In addition, we give heat maps of the genes extracted by FSM and S2N, which are depicted for the top 10, 30 and 50 genes. Although, from these pictures, it is not easy to find a clear reason why FSM has better performance than S2N, we can see an obvious difference between FSM and S2N. Namely, FSM extracts the genes that are down-regulated in ALL and up-regulated in AML, whereas S2N selects the genes with high expression difference irrespective of up- or down-regulation. In top 30 and 50 genes, several ALL samples in S2N show down-regulation (green or black) over all genes. Such samples are likely to cause a substantial increase of error rate. In contrast, FSM shows a clear contrast in intensities between ALL and AML.

Furthermore, we compare the performance of FSM to the random forest method (RFM) that is one of the salient wrapper approaches based on a bootstrap strategy. As shown in **[13]**, the classification accuracy of RFM is 94.9% for leukemia and 87.3% for colon cancer. In contrast, the accuracy of FSM for these datasets attain100% and 93.5%, respectively which surpass considerably compared to those of RFM.

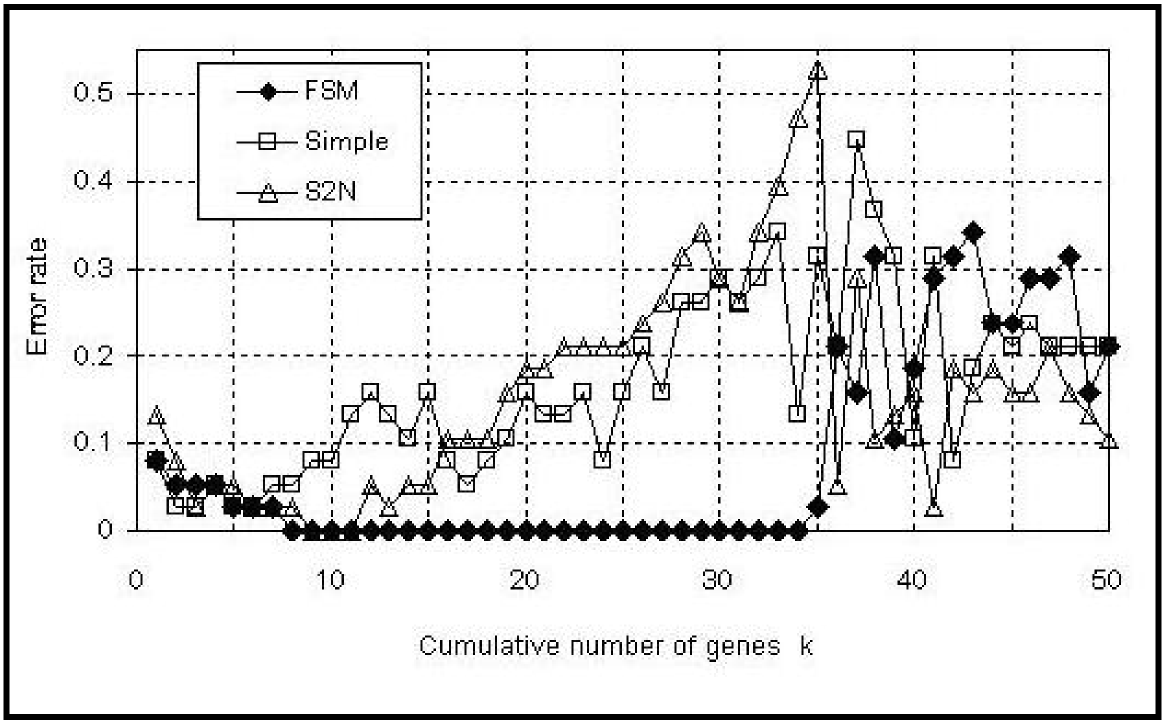

Figure 1: Classification accuracy by FSM, Simple and S2N on Leukemia dataset. The horizontal-axis denotes the cumulative number of genes that are used for classification. For instance, in case of the number of genes k=10, it indicates that the top 10 genes are used for classification. The vertical-axis denotes the error rate which is the ratio of all sample number to error sample number by leave-one-out cross validation (LOOCV).

As described in Methodology, FSM iteratively conducts a greedy, but very simple search, under a restricted number of genes, to find a most informative gene set based on F-value (or Mahalanobis distance) and accomplish a robust and an excellent (or perfect) accuracy of classification. It should be worth to mention here that our results indicate a possibility to extract the minimum gene set which can be applied to design a minimum gene chip for specific diseases. However, for the minimum gene set, it will be necessary to make a statistical test on the interval of the ranking of the genes which have achieved perfect classification.

Conclusion:

In this paper, we have employed the forward variable (gene) selection method based on Mahalanobis distance to search efficiently the combination of informative genes for disease classification. Our method (FSM), which provides a ranking of informative genes, is applied to five microarray datasets and the classification accuracy using the selected genes by our method was compared to traditional filter methods as well as a salient wrapper method. As a result, FSM conspicuously presented robust and high classification accuracy. In particular, it is noteworthy that FSM was able to extract the set of small number of genes having the perfect predictive power, where it is easy to identify the minimum number of genes for classifying samples of different classes. Such minimum set of informative genes must be useful in developing the low-cost and high-reliable microarray (or DNA chip) for specific disease, or designing high-sensitive and non-redundant probes. It might be useful to read that a perfect predictive power of FSM holds over a wide range of gene ranking. Although, we restricted the number of genes to k=50 in this paper to avoid vast combinatorial computation of F-value, the computation can be parallelized by dividing the number of genes k into several groups. It is expected that the classification accuracy would be further improved by adopting a larger number of genes k. In the future, we will extend FSM to a parallel computing or multi-threading program, and verify the performance for large number of k. We will also develop a publicly available web tool or a downloadable standalone tool.

Supplementary data:

See http://www.mmm.muroran-it.ac.jp/~hikaru/article/01/ for additional data.

References:

[01] T. R. Golub *et al., Science*, 286: 531 (1999) [PMID: 10521349]
[02] Ben-Dor *et al., J. Comput. Biol.*, 7: 559 (2000) [PMID: 11108479]
[03] V. G. Tusher *et al., Proc Natl Acad Sci U S A*, 98: 5116 (2001) [PMID: 11309499]
[04] Y. Su *et al., BMC Bioinformatics*, 19: 1578 (2003)
[05] H. Liu *et al, Genome Inform.*, 13: 51 (2002)
[06] R. Kohavi *et al.*, *Artif Intell*, 97: 273 (1997)
[07] I. Guyon *et al., Machine Learning*, 46: 389 (2002)
[08] Inza *et al., Artif Intell Med*, 2: 91 (2004) [PMID: 15219288]
[09] S. A. Armstrong *et al., Nature Genetics*, 30: 41 (2001) [PMID: 11731795]
[10] M. West *et al., Proc Natl Acad Sci U S A*, 98: 11462 (2001) [PMID: 11562467]
[11] U. Alon *et al., Proc Natl Acad Sci U S A*, 96: 6745 (1999) [PMID: 10359783]
[12] S. L. Pomeroy *et al., Nature*, 415: 436 (2002) [PMID: 1807556]
[13] R. Díaz-Uriarte *et al., BMC Bioinformatics*, 7: 3 (2006) [PMID: 16398926]

Edited by P. Kangueane
Citation: Mitsubayashi *et al.*, Bioinformation 3(2): 68-71 (2008)

Supplementary material

(i) The first gene (k=1) is determined by ranking the F-value defined by Eq. (1) below for all genes by putting p=0 and r=1; (ii) For k ($\geq$2) th gene, we pickup a k-th gene from the rest of genes, and add it to the set of k-1 genes. Then, calculate the F-value for the set of total k genes by putting p=k-1 and r=1; (iii) Step (ii) is repeated for all genes in the rest set. k-th gene is determined by choosing the gene with largest F-value; (iv) Step (ii) and (iii) are repeated till the ranking of all genes is accomplished.

Here, F-value is defined by Eq.(1) below. It indicates a statistical measure to evaluate the significance of an increase of squared Mahalanobis distance between the population mean in each classes 1 and 2 when r genes are newly supplied to an original set of p genes.

$$F=\frac{(n^{[1]}+n^{[2]}-p-r-1)n^{[1]}n^{[2]}(D^2_{(p+r)}-D^2_{(p)})}{r\{(n^{[1]}+n^{[2]}-2)(n^{[1]}+n^{[2]})+n^{[1]}n^{[2]}D^2_{(p)}\}} \qquad \textbf{(1)}$$

,where $n^{[c]}$ is the number of samples in class c, p is the number of genes before adding new genes, r is the number of supplied genes, and $D^2_{(p)}$ is the squared Mahalanobis distance between population mean vectors of two classes. It is defined by

$$D^2_{(p)}=(\boldsymbol{\mu}^{[1]}-\boldsymbol{\mu}^{[2]})^T\sum\nolimits^{-1}(\boldsymbol{\mu}^{[1]}-\boldsymbol{\mu}^{[2]}) \qquad \textbf{(2)}$$

,where $\mu^{[c]}$ is the p dimensional population mean vector belonging to class c and □$^{-1}$ is the inverse of population variance-covariance matrix Σ with element σ_{st}, where s, t=1,2,..., p. In our analysis below, instead of $D^2_{(p+r)}$ and $D^2_{(p)}$ which is assumed as population characteristics, we use statistically estimated values of $\hat{D}^2_{(p+r)}$ and $\hat{D}^2_{(p)}$ respectively which are available to estimate experimentally based on gene expression data by microarray. The estimated value $\hat{D}^2_{(p)}$ is given by using the following substitution to (2) respectively:

$$\boldsymbol{\mu}^{[c]}=\left(\mu_1^{[c]},\mu_2^{[c]},\ldots,\mu_p^{[c]}\right)^T \qquad \mu_i^{[c]}\leftarrow\bar{x}_i^{[c]}=\frac{1}{n^{[c]}}\sum_{j=1}^{n^{[c]}}x_{ij}^{[c]} \qquad (c=1,\ 2;\ i=1,2,\ldots,p;\ j=1,\ldots,n^{[c]})$$

$$\sigma_{st}\leftarrow\sum_{c=1}^{2}S_{st}^{[c]}\bigg/\sum_{c=1}^{2}(n^{[c]}-1)\text{, where} \qquad S_{st}^{[c]}=\sum_{j=1}^{n^{[c]}}(x_{sj}^{[c]}-\bar{x}_s^{[c]})(x_{tj}^{[c]}-\bar{x}_t^{[c]})$$

Here, $x_{ij}^{[c]}$ represents the expression value of i-th gene in j-th sample belonging to class c. To avoid a combinatorial explosion in calculating data with colossal dimension, we use, in this paper, the top k genes that are extracted by ranking the F-values for each gene, which is obtainable gene by gene from a simple calculation.