Tclass: tumor classification system based on gene expression profile

Li Wuju\textsuperscript{1} and Xiong Momiao\textsuperscript{2}

\textsuperscript{1}Beijing Institute of Basic Medical Sciences, PO Box 130(3), Beijing 100850, People’s Republic of China and \textsuperscript{2}Human Genetics Center, School of Public Health, University of Texas-Houston Health Science Center, Houston, TX 77225, USA

ABSTRACT

Summary: A method that incorporates feature selection into Fisher's linear discriminant analysis for gene expression based tumor classification and a corresponding program Tclass were developed. The proposed method was applied to a public gene expression data set for colon cancer that consists of 22 normal and 40 tumor colon tissue samples to evaluate its performance for classification. Preliminary results demonstrated that using only a subset of genes ranging from 3 to 10 can achieve high classification accuracy.

Availability: The program is written in Matlab and is being rewritten in the Java language. The source code is available upon request.

Contact: wuju@china.com; mxiong@sph.uth.tmc.edu.

As one of the most important applications of microarrays is tumor classification (Berns, 2000). Although many statistical and computational methods such as cluster analysis (Eisen \textit{et al.}, 1998; Getz \textit{et al.}, 2000), classification (Perou \textit{et al.}, 1999; Golub \textit{et al.}, 1999; Alizadeh \textit{et al.}, 2000) and other methods (Brazma and Vilo, 2000) have been applied to gene expression data analysis; these methods are not very effective for tumor classification. The main advantage of microarray data is that it is able to allow biologists to simultaneously monitor the expression of thousands of genes. However, a large number of genes increases the dimensionality, computational complexity, and cost of data analysis, and introduces some undesired noise. To improve the classification accuracy, an effective tool in machine learning is feature selection (Jain \textit{et al.}, 2000). In this report we introduce a method combining Fisher’s linear discriminant analysis and feature selection based on a stepwise optimization process for tumor classification.

In order to illustrate and evaluate the proposed method, we divided all samples including normal and cancer into training and test sets with different partition ratios 50, 68 and 95%, respectively. The classification accuracy is defined as the percentage of the correctly classified tissue samples in the training set samples. In Tclass system, the standard Fisher’s discriminant method was used. The main calculation processes are as follows: suppose that \( n_N \) normal and \( n_T \) tumor samples are examined. For tissue sample \( i \), we have the vector \( Y_i = (Y_{i1}, Y_{i2}, \ldots, Y_{ik}) \) where ‘\( \cdot \)' denotes the operation of matrix transpose. The \( Y_i \)'s for normal (\( N \)) and tumor (\( T \)) samples constitute the following data matrix,

\[
Y_N = [Y_{N1}, Y_{N2}, \ldots, Y_{NnN}]_{(k \times nN)} \quad (1)
\]

\[
Y_T = [Y_{T1}, Y_{T2}, \ldots, Y_{TnT}]_{(k \times nT)} \quad (2)
\]

Next, we calculate the sample mean vectors, covariance matrices, and middle point as follows:

\[
\bar{Y}_N = \frac{1}{n_N} \sum_{i=1}^{n_N} Y_{Ni},
\]

\[
S_N = \frac{1}{n_N - 1} \sum_{i=1}^{n_N} (Y_{Ni} - \bar{Y}_N)(Y_{Ni} - \bar{Y}_N)'
\]

\[
\bar{Y}_T = \frac{1}{n_T} \sum_{i=1}^{n_T} Y_{Ti},
\]

\[
S_T = \frac{1}{n_T - 1} \sum_{i=1}^{n_T} (Y_{Ti} - \bar{Y}_T)(Y_{Ti} - \bar{Y}_T)'
\]

\[
S = \frac{(n_N - 1)S_N + (n_T - 1)S_T}{n_N + n_T - 2}
\]

\[
\hat{m} = \frac{1}{2} (\bar{Y}_N - \bar{Y}_T)'S^{-1}(\bar{Y}_N + \bar{Y}_T)
\]

where \( S \) stands for the pooled sample covariance matrix of the training sets of the two classes, \( \hat{m} \) stands for the middle point of the tumor and normal samples based on training set, and \( S^{-1} \) denotes the inverse matrix of matrix \( S \).

The classification rule based on Fisher’s linear discriminant function for an unknown sample, \( Y_0 \), is as follows: Assign \( Y_0 \) to the normal group if \( (\bar{Y}_N - \bar{Y}_T)'S^{-1}Y_0 > \hat{m} \), or assign \( Y_0 \) to the tumor group if \( (\bar{Y}_N - \bar{Y}_T)'S^{-1}Y_0 < \hat{m} \).
Since a learning algorithm is employed to evaluate every set of features considered, feature selection is very expensive to run. Although an exhaustive search can find optimal subsets of genes, it requires a large amount of computation. To avoid excessive computation, we adopted a heuristic algorithm sequential forward selection (SFS) (Sahiner et al., 2000). The procedure for SFS is as follows:

1. compute classification accuracy for each of the features. Select the feature with the best value;
2. the best gene selected from step 1 is then combined with each of the other genes (a total of 1999) to form a pair. Next compute the classification accuracy for each pair and select the best one;
3. the best gene pair selected from step 2 is then combined with each of the other genes (a total of 1998) to form a gene set with three genes. Next compute the classification accuracy for each gene set and select the best one. Continue this process until reaching the pre-specified dimension of the feature vector, say l.

The input for the program is the gene expression matrix, the number of tissues in each class, the number of candidate gene sets and the number of genes in each gene set. Then the program will automatically find the subset of genes with the highest classification accuracy for each fixed size of gene sets starting from 1 and ending with the pre-specified number l.

It is known that a subset of features that has high classification accuracy in the training set may not have good generalization properties. In order to evaluate the generalization ability of the optimal subset of genes, the tissue samples were randomly divided into a training set (e.g. 50, 67 or 95% of samples) and a test set. For a given subset of genes that was identified using the above SFS procedure, we computed the number of samples that were correctly classified in both the training and test sets. The average classification rate was taken as the discriminant power. It is noteworthy that for a given subset of genes, the discriminant power may depend on the assignment of samples in the training and test sets. Thus, we randomly shuffled the samples 200 times resulting in 200 different assignments of the ‘training’ and ‘test’ sets. The average classification rate from the 200 assignments seems to be stable and is reported in Figure 1 for the gene expression data from the colon cancer tissue (Alon et al., 1999). We found that five or six genes are enough to classify colon tissue samples. These six genes are R87126, T81492, H40269, M24069, M67454, and H18490. With the increase of the number of genes in the set, the classification accuracy does not increase. The results can be seen clearly from Figure 1. These results are also verified in other data sets (Golub et al., 1999; Perou et al., 1999).

ACKNOWLEDGEMENTS

M.M.Xiong is supported by NIH grants GM56515 and HL 5448. The authors thank John C.Huber Jr, for his helpful comments on this paper, which helped to improve its presentation. The authors also thank two anonymous reviewers for very helpful comments on the manuscript.

REFERENCES