Estimating Error-Dimensionality Relationship for Gene Expression Based Cancer Classification

Feng Chu and Lipo Wang
School of Electrical and Electronic Engineering
Nanyang Technological University
Singapore 639798
Email: elpwang@ntu.edu.sg

Abstract

Accurate estimating error-dimensionality relationship is important for gene-expression-based cancer classification. Here, we introduce an effective method to deal with this problem. To obtain a good performance of the estimator, we also propose a novel feature (gene) selection approach that aims to minimize the estimated error with respect to dimensionality. The experimental results for a gene expression data set, i.e. the liver cancer data set, substantiate the effectiveness of the estimator and the feature selection method as well.

I. Introduction

Microarrays are also called gene chips or DNA chips that provide researchers a powerful tool to monitor expression profiles of thousands of genes simultaneously [1], [2]. Such expression profiles are able to reveal molecular differences among cells. Therefore, microarrays have been intensively applied to classifying cancers, e.g., lymphoma [3], leukemia [4], liver cancer [5], etc., in recent years.

A typical gene-expression-based cancer classification task usually can be divided into two parts. The first one is selecting a subset of genes with good discriminant power (gene/feature selection); and the second one is designing a good classifier (algorithm design). In literature, a variety of methods have been reported to deal with each them [6], [7], [8]. However, how to estimate the relationship between dimensionality (number of genes) and classification errors is still a muddy problem that needs further investigation. This error-dimensionality relationship is important because of the following points.

1) It makes it possible to estimate how large an optimal gene subset (i.e., the gene subset that leads to the smallest classification error) is, which is very important to design a classifier economically.
2) It can be used as a criteria to evaluate data quality of data sets.
3) It is a guide to efficient feature extraction/selection.

Here, we introduce an accurate method to estimate error-dimensionality relationship for cancer classification systems based on gene expression profiles. In addition, a novel feature selection method that aims to minimize estimated classification error is also proposed. Experimental results indicated that our method is very accurate in estimating error-dimensionality relationship.

This paper is organized as follows. In section II, we introduce the Fisher linear discriminant (FLD) (see e.g. [9]) and the method to estimate the error-dimensionality relationship for FLDs. In section III we propose a novel feature (gene) selection method. In section IV, we test our methods in a gene expression data set. In the last section, we draw our conclusions and recommend the future work.

II. Estimating Error-Dimensionality Relationship

In this section, we introduce a simple method for estimating error-dimensionality relationship for FLDs. The reason why we choose the FLD as our classifier is because it is a simple yet very powerful classifier. And complicated classifiers will make the estimation of classification errors also very complicated or even impossible.

A. Fisher Linear Discriminant

Mathematically, an FLD can be described as follows. There are two groups of samples, $\{\mathbf{Y}_{1i}, i = 1, 2, ..., n_1\}$ and $\{\mathbf{Y}_{2i}, i = 1, 2, ..., n_1\}$, which belong to class C_1 and class C_2, respectively. Their mean vectors $\mathbf{M}_1$ and $\mathbf{M}_2$ are given by

$$\mathbf{M}_1 = \frac{1}{n_1}\sum_{i=1}^{n_1}\mathbf{Y}_{1i} \tag{1}$$

$$\mathbf{M}_2 = \frac{1}{n_2}\sum_{i=1}^{n_2}\mathbf{Y}_{2i}. \tag{2}$$

And their within-class covariances, S_1 and S_2, are given by

$$\mathbf{S}_1 = \frac{1}{n_1 - 1}\sum_{i=1}^{n_1}(\mathbf{Y}_{1i} - \mathbf{M}_1)(\mathbf{Y}_{1i} - \mathbf{M}_1)' \tag{3}$$

$$\mathbf{S}_2 = \frac{1}{n_2 - 1}\sum_{i=1}^{n_2}(\mathbf{Y}_{2i} - \mathbf{M}_2)(\mathbf{Y}_{2i} - \mathbf{M}_2)'. \tag{4}$$

Therefore, the overall within-class covariance, $\mathbf{S}$, can be obtained by:

$$\mathbf{S} = \frac{(n_1 - 1)\mathbf{S}_1 + (n_2 - 1)\mathbf{S}_2}{n_1 + n_2 - 2}. \tag{5}$$

The basic idea of FLDs is to project multivariate samples into one dimension in which the samples belonging to different classes should be separated as much as possible. According to [9], such a projection can be achieved by the transformation below:

$$y = \mathbf{W}'\mathbf{Y} = (\mathbf{M}_1 - \mathbf{M}_2)'\mathbf{S}^{-1}\mathbf{Y}. \tag{6}$$

where $\mathbf{W}'$ is the transpose of $\mathbf{W}$, and $\mathbf{S}^{-1}$ is the inverse of $\mathbf{S}$. Noting that y is a sum of a number of random variables, the discriminant threshold that is obtained based on the *central limit theorem* (see e.g. [9]) should be

$$t = \frac{1}{2}(\mathbf{M}_1 - \mathbf{M}_2)'\mathbf{S}^{-1}(\mathbf{M}_1 + \mathbf{M}_2). \tag{7}$$

Hence, the FLD classifies a sample, e.g. $\mathbf{Y}_0$, with the following rule:

If $(\mathbf{M}_1 - \mathbf{M}_2)'\mathbf{S}^{-1}\mathbf{Y}_0 \geq t$, $\mathbf{Y}_0$ is categorized to C_1;
otherwise, if $(\mathbf{M}_1 - \mathbf{M}_2)'\mathbf{S}^{-1}\mathbf{Y}_0 < t$, $\mathbf{Y}_0$ is categorized to C_2.

B. Relationship between Dimensionality and Classification Errors

Since all the samples can be regarded as observations of random variables, a classifier designed with these samples can also be regarded as a function of random variables, i.e., the classifier's parameters are random variables. Furthermore, its probability of misclassification (PM) is also a random variable. Therefore, we can use the *expectation* of PM ($\hat{PM}$) to *estimate* the errors the classifier will make.

In fact, a lot of efforts have been made to determine the relationship between dimensionality and $\hat{PM}$ [10], [11], [12]. For an FLD, its $\hat{PM}^{(F)}$ can be calculated with the following equation [13]:

$$\hat{PM}^{(F)} = \Phi\left\{-\frac{\hat{\delta}}{2}\left[(1+\frac{p}{n_1+n_2-p})(1+\frac{(p-1)(n_1+n_2)}{n_1 n_2 \hat{\delta}^2})\right]^{-\frac{1}{2}}\right\} \tag{8}$$

where $\Phi(a)$ is the standard Gaussian cumulative distribution function

$$\Phi(a) = \int_{-\infty}^{a} \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}} dt. \tag{9}$$

$\hat{\delta}^2$ is an unbiased estimate of the Mahalanobis distance between the two classes.

$$\hat{\delta}^2 = \left[(\mathbf{M}_1 - \mathbf{M}_2)'\mathbf{S}^{-1}(\mathbf{M}_1 - \mathbf{M}_2)\frac{n_1+n_2-p-3}{n_1+n_2-2}\right] - \frac{n_1+n_2}{n_1 n_2}p. \tag{10}$$

p is dimensionality, i.e., how many features are used by the FLD.

III. A Novel Feature (Gene) Selection Approach

Based on equation 8, one can estimate the variation of $\hat{PM}^{(F)}$ with respect to p for a FLD. The $\hat{PM}^{(F)}$-p curve reaches its minimum very fast if all the features are properly ordered according to their discriminating ability. Here, "properly ordered" means that a feature with higher discriminating power should be put into a place with a higher rank. On the contrary, if features are not ordered properly, the $\hat{PM}^{(F)}$-p curve reaches its minimum slowly. Or ever worse, its minimum will be greatly enlarged by some "noise" features, which undermines the accuracy of

the estimator. Considering this characteristic of $P\hat{M}^{(F)}$-p curve, we propose a novel feature ranking approach that attempts to make $P\hat{M}^{(F)}$ equal to its minimum with the corresponding p.

Our feature selection approach is based on the "hill-climbing" technique [14]. First of all, we use a matrix, A, to contain the features to be ranked, and use another matrix, B, to contain the ranked features (B is an empty matrix before the ranking process). Then, we use the following steps to rank features.

1) Make a copy of B, and save it as B_s.
2) Select one feature from **A**. Then make a copy of this feature and save it as V_a.
3) Put V_a right behind the last row of B_s, i.e., if B_s has l rows, V_a is added into B_s as the $(l+1)$-th row.
4) Calculate $P\hat{M}^{(F)}$ for the B_s generated in the last step using equation 8.
5) Repeat from step 1 by selecting another feature until all the features in A have been selected once.
6) From A, pick out the feature that minimizes $P\hat{M}^{(F)}$ in step 4 and put it right behind the last row of B.
7) Repeat from step 1 until a required number of features have been moved to B.

After all these steps, the features in B are ranked with a decreasing order of their discriminant power in from the first row to the last row. And an FLD tends to make less errors when the features are input into it with this order than with other orders.

IV. Experimental Results

In this section, we test our methods in the liver cancer microarray data set. The liver cancer data set (http://genome-www.stanford.edu/hcc/) [5] has two classes, i.e. the nontumor liver and hepatocellular carcinoma (HCC). The data set contains 156 samples and the expression profiles of 1648 important genes. Among them, 82 are HCCs and the other 74 are nontumor livers. In this data set, there are some missing values. We used a k-nearest neighbor method to fill those missing values [15].

We firstly ranked all the genes with the feature (gene) selection approach that we proposed in Section III. We selected 150 genes from the total 1648 genes. After that, we estimated classification errors with the method introduced in Section II. For the sake of verification, we used an FLD to classify the liver cancer data set and validated the results with the *leave-one-out* technique. The estimation and classification results are shown in Fig. 1a and Fig. 1b, respectively. Fig. 1c is a comparison of the two results. From Fig. 1c, we find that the estimated error, i.e. $P\hat{M}^{(F)}$, is quite close to the LOOV error, which proves the effectiveness of $P\hat{M}^{(F)}$ in estimating classification errors.

V. Conclusions and Future Work

Accurate estimating error-dimensionality relationship is an important problem for gene- expression-based cancer classification. In this paper, we introduced an effective method to deal with this problem. Together with the feature (gene) selection method we proposed, our method obtained a very good estimation result in the liver cancer data set, which substantiated its effectiveness.

Until now, our estimating method only can deal with problems with 2 classes. In the future, we will extend it to multi-class problems. Besides, we will apply it to more date sets.

REFERENCES

[1] D. K. Slonim, "From patterns to pathways: gene expression data analysis comes of age," *Nature Genetics Suppl.*, vol. 32, pp. 502–508, Dec 2002.

[2] G. Russo, C. Zegar, and A. Giordano, "Advantages and limitations of microarray technology in human cancer," *Oncogene*, vol. 22, pp. 6497–6507, Sep 2003.

[3] A. A. Alizadeh, M. B. Eisen, R. E. Davis, C. Ma, I. S. Lossos, A. Rosenwald, J. C. Boldrick, H. Sabet, T. Tran, X. Yu, and *et al.*, "Distinct types of diffuse large b-cell lymphoma identified by gene expression profiling," *Nature*, vol. 403, pp. 503–511, Feb 2000.

[4] T. R. Golub, D. K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. P. Mesirov, H. Coller, M. L. Loh, J. R. Downing, M. A. Caligiuri, and *et al.*, "Molecular classification of cancer: class discovery and class prediction by gene expression monitoring," *Science*, vol. 286, pp. 531–537, Oct 1999.

[5] X. Chen, S. T. Cheung, S. So, S. T. Fan, and C. Barry, "Gene expression patterns in human liver cancers," *Molecular Biology of Cell*, vol. 13, pp. 1929–1939, Jun 2002.

[6] R. Tibshirani, T. Hastie, B. Narashiman, and G. Chu, "Diagnosis of multiple cancer types by shrunken centroids of gene expression," *Proc. Natl. Acad. Sci. USA*, vol. 99, pp. 6567–6572, May 2002.

[7] J. M. Deutsch, "Evolutionary algorithms for finding optimal gene sets in microarray prediction," *Bioinformatics*, vol. 19, pp. 45–52, Jan 2003.

[8] Y. Lee and C. K. Lee, "Classification of multiple cancer types by mulitcategory support vector machines using gene expression data," *Bioinformatics*, vol. 19, pp. 1132–1139, Jun 2003.

[9] C. M. Bishop, *Neural Networks for Pattern Recognition*. Oxford: Clarendon Press, 1995.

[10] D. H. Foley, "Considerations of sample and feature size," *IEEE Transactions on Information Theory*, vol. 18, pp. 618–626, Sep 1972.

[11] S. J. Raudys, "Determination of optimal dimensionality in statistical pattern classification," *Pattern Recognition*, vol. 11, pp. 263–270, 1979.

[12] S. Raudys, "On dimensionality, sample size, and classification error of nonparametric linear classification algorithms," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 19, pp. 667–671, Jun 1997.

[13] S. J. Raudys and A. K. Jain, "Small sample size effects in statistical pattern recognition: recommendations for practitioners," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 13, pp. 252–267, Mar 1991.

[14] R. Kohavi and G. H. John, "Wrappers for feature selection," *Artificial Intelligence, special issue on relevance*, vol. 97, pp. 273–324, 1996.

[15] O. Troyanskaya, M. Cantor, G. Sherlock, P. Brown, T. Hastie, R. Tibshirani, D. Botstein, and R. B. Altman, "Missing value estimation methods for dna microarrays," *Bioinformatics*, vol. 17, pp. 520–525, 2001.

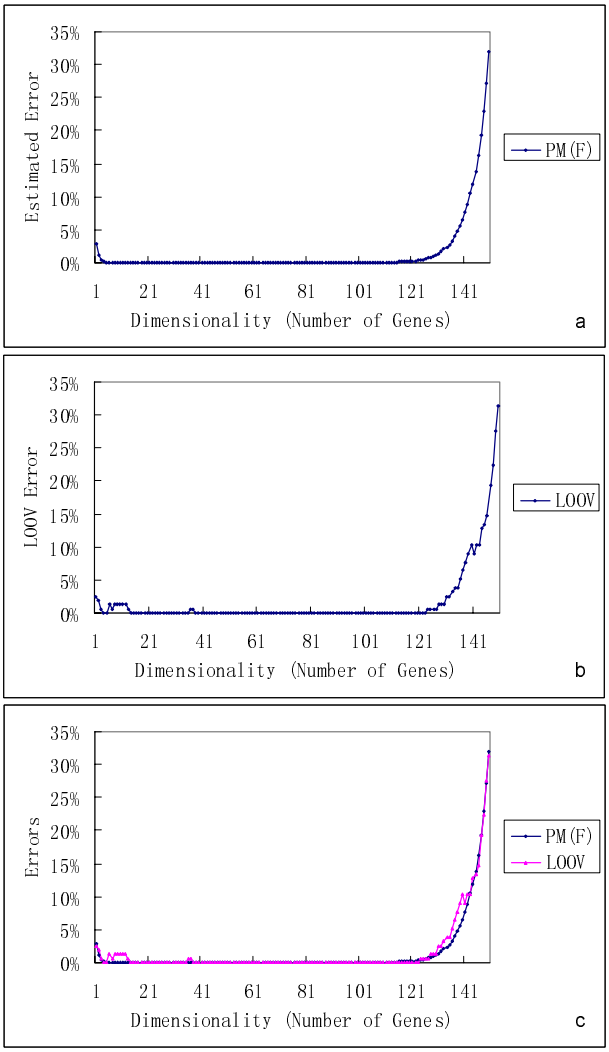

Fig. 1. The results for the liver cancer data set: (a) the estimated error with respective to dimensionality, (b) the LOOV error with respective to dimensionality, and (c) comparision of the estimated error and the LOOV error with respective to dimensionality.