Proceedings of the IEEE Computer Society Bioinformatics Conference, August 11 – 12, Stanford, CA

Gene Selection for Multi-Class Prediction of Microarray Data

Dechang Chen
Uniformed Services University of the Health Sciences
dchen@usuhs.mil

Dong Hua
The George Washington University
gwuhua@gwu.edu

Jaques Reifman
US Army Medical Research and Materiel Command
reifman@tatrc.org

Xiuzhen Cheng
The George Washington University
cheng@gwu.edu

Abstract

Gene expression data from microarrays have been successfully applied to class prediction, where the purpose is to classify and predict the diagnostic category of a sample by its gene expression profile. A typical microarray dataset consists of expression levels for a large number of genes on a relatively small number of samples. As a consequence, one basic and important question associated with class prediction is: how do we identify a small subset of informative genes contributing the most to the classification task? Many methods have been proposed but most focus on two-class problems, such as discrimination between normal and disease samples. This paper addresses selecting informative genes for multi-class prediction problems by jointly considering all the classes simultaneously. Our approach is based on the power of the genes in discriminating among the different classes (e.g., tumor types) and the existing correlation between genes. We formulate the expression levels of a given gene by a one-way analysis of variance model with heterogeneity of variances, and determine the discriminatory power of the gene by a test statistic designed to test the equality of the class means. In other words, the discriminatory power of a gene is associated with a Behrens-Fisher problem. Informative genes are chosen such that each selected gene has a high discriminatory power and the correlation between any pair of selected genes is low. Test statistics considered in this paper include the ANOVA F test statistic, the Brown-Forsythe test statistic, the Cochran test statistic, and the Welch test statistic. Their performances are evaluated over several classification methods applied to two publicly available microarray datasets. The results show that Brown-Forsythe test statistic achieves the best performance.

1. Introduction

DNA microarrays provide a very effective approach to interrogate hundreds or thousands of genes simultaneously. Such high throughput capability offers great opportunities in terms of data collection but also poses great challenges in terms of analyzing the data and transforming the data into useful information. Essentially, microarrays provide information about the expression level of the genes represented on the array. Such gene expression profiling has been successfully applied to class prediction, where the purpose is to classify and predict the diagnostic category of a sample by its gene expression profile [5, 6, 14, 16]. Various machine learning methods are currently used for class prediction. However, the task of prediction by microarrays is challenging. One main reason is that the number of genes is large but the number of samples is relatively small. As a consequence, one basic and important question required to be answered is: how do we identify a small subset of informative genes (or features) contributing most to the classification task?

Performing feature selection is essential in microarray prediction problems. High-dimensional problems usually involve higher computational complexity and larger prediction errors. Besides, a large number of genes usually comprise many irrelevant genes with respect to classification and sometimes can skew the result [7]. Furthermore, because the number of genes is much larger than the number of samples, one can find features that discriminate by chance.

Many methods have been proposed to select informative genes in microarray data analysis. They include TNoM score [1], naive Bayes global relevance method [3] , nonparametric scoring [13], information gain [19], and t-score [20]. Most recently, Lee et al. [8] obtained informative genes through a Bayesian variable selection approach, and Li and Grosse [10] conducted gene selection based on the

extreme value distribution. We note that the majority of the discussions on gene selection have focused on two-class problems such as the discrimination between normal and cancerous samples.

In this paper, we address selecting informative genes for multi-class prediction problems by jointly considering all the classes simultaneously. Our approach is based on the power of the genes in discriminating among the different classes (e.g., tumor types) and the existing correlation between genes. We formulate the expression levels of a given gene by a one-way analysis of variance model with heterogeneity of variances, and determine the power of the gene in discriminating between classes by a test statistic designed to test the equality of the class means. Naturally, one would like to choose as informative genes those genes having high discriminatory power. However, selecting the high power genes may involve correlated genes, based on the correlation. There is not much additional information added to the prediction task if correlated features are kept. Therefore we may need to discard irrelevant genes. Our final list of informative genes should be chosen such that each selected gene has a high discriminatory power and the correlation between any pair of selected genes is low. Our gene selection procedure depends on the choices of the test statistics. Different test statistics may lead to different sets of informative genes. For this reason, we study four test statistics in this paper. These test statistics are extensions of the t-statistic used in the two-class prediction problems.

The paper is organized as follows. In Section 2, we describe statistical models for gene expression levels, test statistics, and our approach to select genes using discriminatory power and correlation. In Section 3, we investigate the effect of test statistics on the classification results by using our gene selection approach and several machine learning techniques applied to two publicly available microarray datasets. Conclusion and future research activities are discussed in Section 4.

2. Models and Methods

In this section, we will first introduce general statistical models for gene expression values and describe test statistics for testing the equality of the class means. We then present our approach to select informative genes using discriminatory power and correlation.

2.1. Statistical Models

Assume the prediction problem involves $k(\geq 2)$ distinct tumor tissue classes, p genes, and n tumor mRNA samples (observations). Furthermore, suppose that X_{rs} represents the measurement of the expression level of the rth gene from the sth sample for $r = 1, \ldots, p$ and $s = 1, \ldots, n$. In terms of an expression matrix $\mathbf{G}$, we may write

$$\mathbf{G} = \begin{pmatrix} X_{11} & X_{12} & \cdots & X_{1n} \\ X_{21} & X_{22} & \cdots & X_{2n} \\ \vdots & \vdots & \vdots & \vdots \\ X_{p1} & X_{p2} & \cdots & X_{pn} \end{pmatrix}.$$

The columns and rows of the expression matrix $\mathbf{G}$ correspond to samples and genes, respectively. Given a fixed gene, let Y_{ij} be the expression level from the jth sample of the ith class. Note that these Y_{ij} come from the corresponding row of $\mathbf{G}$. For example, for gene 1, Y_{ij} are a rearrangement of the first row of $\mathbf{G}$. Schematically, the expression levels Y_{ij} look like the following:

Classes				
1	2	3	...	k
Y_{11}	Y_{21}	Y_{31}	...	Y_{k1}
Y_{12}	Y_{22}	Y_{32}	...	Y_{k2}
$\vdots$	$\vdots$	$\vdots$	...	$\vdots$
	Y_{2n_2}			
		Y_{3n_3}		
Y_{1n_1}				Y_{kn_k}

For a given gene, we consider the following general model for Y_{ij}:

$$Y_{ij} = \mu_i + \epsilon_{ij}, \quad \text{for} \quad i = 1, 2, \ldots, k;\ j = 1, 2, \ldots, n_i$$

with $n_1 + n_2 + \ldots + n_k = n$. In this model, μ_i is a parameter representing the mean expression level of the gene in class i, ϵ_{ij} are the error terms such that ϵ_{ij} are independent normal random variables and $E(\epsilon_{ij}) = 0$, $V(\epsilon_{ij}) = \sigma_i^2 < \infty$, for $i = 1, 2, \ldots, k$; $j = 1, 2, \ldots, n_i$.

Note that if the variances are equal, i.e., $\sigma_1^2 = \sigma_2^2 = \ldots = \sigma_k^2$, then the above model is simply the commonly used one-way ANOVA model. For the microarray data, heterogeneity in the variances may be more realistic, since different σ_i may describe different variation of the gene expression across classes.

One of the main tasks associated with the above model is to detect whether or not there is some difference among the means μ_1, μ_2, ..., μ_k. For the case of homogeneity of variances, the well known ANOVA F test is the optimal test to accomplish the task [9, 12]. However, with heterogeneity of the variances, the task is challenging and is closely related to the well known Behrens-Fisher problem [15]. When the sample sizes in all classes are equal, i.e., $n_1 = n_2 = \ldots = n_k$, the presence of heterogeneous variances of the errors only slightly affects the F test. That is not the case when the sample sizes are unequal [11]. The actual type I error is inflated if smaller sizes n_i are associated with larger variances σ_i^2, and the significance levels are smaller than anticipated if larger sizes n_i are associated with larger variances σ_i^2. This indicates that for

our model, the F test may not be appropriate for testing $H_0 : \mu_1 = \mu_2 = \ldots = \mu_k$ versus H_1 : not all the μ_i are equal. Therefore, some alternatives to the F test are worth investigation.

2.2. Test Statistics

We consider the following four test statistics.

a) ANOVA F test statistic [12]. It is defined as

$$F = \frac{(n-k)\sum n_i(\bar{Y}_{i\cdot} - \bar{Y}_{\cdot\cdot})^2}{(k-1)\sum(n_i-1)s_i^2},$$

where $\bar{Y}_{i\cdot} = \sum_{j=1}^{n_i} Y_{ij}/n_i$, $\bar{Y}_{\cdot\cdot} = \sum_{i=1}^{k} n_i\bar{Y}_{i\cdot}/n$, and $s_i^2 = \sum_{j=1}^{n_i}(Y_{ij} - \bar{Y}_{i\cdot})^2/(n_i - 1)$. For simplicity, we use $\sum$ to indicate the sum is taken over the index i.

b) Brown-Forsythe test statistic [2]. It is given by

$$B = \frac{\sum n_i(\bar{Y}_{i\cdot} - \bar{Y}_{\cdot\cdot})^2}{\sum(1 - n_i/n)s_i^2}.$$

c) Cochran test statistic [4]. It is defined as

$$C = \sum w_i(\bar{Y}_{i\cdot} - \sum h_i\bar{Y}_{i\cdot})^2$$

with $w_i = n_i/s_i^2$ and $h_i = w_i/\sum w_i$.

d) Welch test statistic [17] W defined to be

$$\frac{\sum w_i(\bar{Y}_{i\cdot} - \sum h_i\bar{Y}_{i\cdot})^2}{(k-1) + 2(k-2)(k+1)^{-1}\sum(n_i-1)^{-1}(1-h_i)^2}.$$

2.3. Gene Selection

Our method of gene selection is based on correlation of genes and power of genes in discriminating between tumor types. Given the gene expression matrix $\mathbf{G}$, the correlation between gene r and gene r' is

$$\frac{\sum_s(X_{rs} - \bar{X}_r)(X_{r's} - \bar{X}_{r'})}{\sqrt{\sum_s(X_{rs} - \bar{X}_r)^2\sum_s(X_{r's} - \bar{X}_{r'})^2}},$$

where $\bar{X}_r = \sum_s X_{rs}/n$ represents the average level of the gene r, based on the n samples.

Given a test statistic $\mathcal{T}$ described above, we define the *discriminatory power* of a gene as the value of $\mathcal{T}$ evaluated over the n expression levels of the gene. This definition is based on the fact that with larger $\mathcal{T}$ the null hypothesis $H_0 : \mu_1 = \mu_2 = \ldots = \mu_k$ will be rejected more likely. Therefore, the higher the discriminatory power is, the more powerful the gene is in discriminating between tumor types.

Assume that for the purpose of classification, we want to select a set of m informative genes out of a total number of p genes, where m is a predetermined number. The selection requires that: a) each of the m genes has a high score of discriminatory power and b) the correlation between any pair of genes is low. The selection process can be made in may different ways. This paper focuses on the following greedy algorithm. Choose a test statistic $\mathcal{T}$. The first selected gene is the gene having the highest discriminatory power determined by $\mathcal{T}$. Consider the set of all genes whose correlation to the chosen gene is not larger than the specified threshold θ. The gene with the highest discriminatory power from this set is selected as the second informative gene. In general, the lth informative gene is the gene with the highest discriminatory power from the set of all genes whose correlation to each of the chosen $l-1$ genes is not larger than θ. The process is repeated and the selection is terminated after m informative genes are obtained. This greedy algorithm applied to two-class classification is used in [7]. If m is not given, the selection process could be continued until the correlation requirement fails. We use ω to denote the total number of informative genes selected in this scenario.

3. Results and Discussion

The gene selection procedure described above depends on the test statistics. Different test statistics may eventually lead to different prediction errors. In this section, we investigate such effects of the test statistics.

We used the following two datasets: LEUKEMIA72 with 6817 genes, 38 ALL-Bcell, 9 ALL-Tcell, and 25 AML [6, 22] and OVARIAN with 7129 genes, 27 epithelial ovarian cancer cases, 5 normal tissues, and 4 malignant epithelial ovarian cell lines [18, 21]. The classification methods we used include the Nearest Neighbor, LogitBoost (DecisionStump as the classifier), C4.5, and Naive Bayes from the software Weka developed by Ian H. Witten and Eibe Frank. In addition, we chose $\theta = 0.1, 0.2, \ldots, 1.0$ with increment 0.1. Before any analysis, the expression values from each dataset were standardized so that the genes have mean 0 and variance 1 across the samples. We performed gene selection for each possible combination of test statistic and θ. The number of genes selected was set to be the maximum of 100 and ω. For each possible combination of test statistic, classification method, and θ , we evaluated the prediction error using leave-one-out cross-validation.

Table 1 shows the summarized performance over the four test statistics. In the table, the first number in each cell represents the average of 40 prediction errors based on 10 values of θ and four classification algorithms. The second number in each cell is the median of the 40 prediction errors. It is seen that Brown-Forsythe test statistic B has the best performance, since B always achieves the lowest prediction error (average or median) for each dataset. These results indicate that the proposed models (Section 2.1) without assuming equal variances are appropriate in practice.

Table 1. Performance of the test statistics

	F	B	C	W
LEUKEMIA	6.08% 5.56%	6.01% 4.86%	7.40% 5.56%	7.53% 5.56%
OVARIAN	5.21% 2.78%	5.00% 2.78%	10.83% 11.11%	10.97% 11.11%

4. Conclusions

In this paper we present a general procedure of gene selection, based on correlation and test statistics, for multiclass prediction of microarray data. Demonstration of the procedure on two datasets shows that Brown-Forsythe test statistic B has the best performance among several test statistics, including the traditional ANOVA F test statistic. For a more detailed examination of these statistics, more experiments on real datasets are needed. We expect to report further results along this direction in the near future.

References

[1] A. Ben-Dor, et al. Tissue classification with gene expression profiles. *J. Comput. Biol.* 7, 559-583, 2000.

[2] M. B. Brown, and A. B. Forsythe. The small sample behavior of some statistics which test the equality of means. *Technometrics*, 16, 129-132, 1974.

[3] M. L. Chow, E. J. Moler, and I. S. Mian. Identifying marker genes in transcription profiling data using a mixture of feature relevance experts. *Physiol Genomics*, 5, 99-111, 2001.

[4] W. G. Cochran. Problems arising in the analysis of a series of similar experiments. *J. Roy. Stat. Soc. Supp.*, 4, 102-118, 1937.

[5] S. Dudoit, J. Fridlyand, and T. P. Speed. Comparison of discrimination methods for the classification of tumors using gene expression data. *Journal of the American Statistical Association*, 97, 77-87, 2002.

[6] T. R. Golub, et al. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, 286, 531-537, 1999.

[7] J. Jaeger, R. Sengupta, and W. L. Ruzzo. Improved gene selection for classification of microarrays. *Pacific Symposium on Biocomputing*, 8, 53-64, 2003.

[8] K. E. Lee, et al. Gene selection: a Bayesian variable selection approach. *Bioinformatics*, 19, 90-97, 2003.

[9] E. L. Lehman. *Testing Statistical Hypotheses*, 2nd edition. New York: Wiley, 1986.

[10] W. Li, and I. Grosse. Gene selection criterion for discriminant microarray data analysis based on extreme value distributions. *Proceedings of the Seventh International Conference on Research in Computational Molecular Biology*, 217-223, 2003.

[11] D. C. Montgomery. *Design and Analysis of Experiments*, 5th edition. New York: Wiley, 2001.

[12] J. Neter, M. H. Kutner, C. J. Nachtsheim, and W. Wasserman. *Applied Linear Statistical Models*, 4th edition. McGraw-Hill, 1996.

[13] P. J. Park, M. Pagano, and M. Bonetti. A nonparametric scoring algorithm for identifying informative genes from microarray data. *Pacific Symposium on Biocomputing*, 6, 52-63, 2001.

[14] S. Ramaswamy, et al. Multiclass cancer diagnosis using tumor gene expression signatures. *Proc. Natl. Acad. Sci. USA*, 98, 15149-15154, 2001.

[15] A. Stuart, J. K. Ord, and S. Arnold. *Advanced Theory of Statistics, Volume 2A: Classical Inference and the Linear Model*, 6th edition. London: Oxford University Press, 1999.

[16] R. Tibshirani, T. Hastie, B. Narasimhan, and G. Chu. Diagnosis of multiple cancer types by shrunken centroids of gene expression. *Proc. Natl. Acad. Sci. USA*, 99, 6567-6572, 2002.

[17] B. L. Welch. On the comparison of several mean values: An alternative approach. *Biometrika*, 38, 330-336, 1951.

[18] J. B. Welsh, et al. Analysis of gene expression profiles in normal and neoplastic ovarian tissue samples identifies candidate molecular markers of epithelial ovarian cancer. *Proc Natl Acad Sci USA*, 98:1176-1181, January 2001.

[19] E. P. Xing, M. I. Jordan, and R. Karp. Feature selection for high-dimensional genomic microarray data. *ICML*, 601-608, 2001.

[20] M. Xiong, et al. Computational Methods for Gene Expression-Based Tumor Classification. *Biotechniques*, 29, 1264-1270, 2000.

[21] www.gnf.org/cancer/ovary

[22] www.wi.mit.edu/MPR/data_set_ALL_AML.html