A Multi-population χ^2 Test Approach to Informative Gene Selection[*]

Jun Luo and Jinwen Ma[**]

D[illegible] I[illegible] S[illegible],
S[illegible] M[illegible] S[illegible] LMAM
P[illegible] U[illegible], B[illegible], 100871, C[illegible]
jwma@math.pku.edu.cn

Abstract. [illegible] -o [illegible] χ^2 [illegible] -o [illegible] χ^2 [illegible] -o [illegible] χ^2 [illegible] (S M) [illegible] -o [illegible] χ^2 [illegible] 100% [illegible] 97.1% [illegible].

1 Introduction

With the rapid development of DNA microarray technology, we can now get the expression levels of thousands of genes via one single experiment. Certainly, these gene expression profiles or simply called microarray data provide important and detailed evidences to health state of human tissues for tumor analysis and diagnosis. Mathematically, the microarray data corresponding to a tumor can be represented by a matrix $A = (a_{ij})n \times m$, where the $i-$th row represents the $i-$th gene, the $j-$th column represents the $j-$th sample, and the element a_{ij} represents the expression level of the $i-$th gene in the $j-$th sample. Many microarray data sets are now available on the web.

In tumor diagnosis, each sample can be identified as "tumorous" or "normal", and it is expected to construct a binary classifier as a diagnosis system to classify them as correctly as possible. Clearly, this is just a problem of supervised binary classification. However, as there are always thousands of genes in a microarray chip, the microarray data are generally complete, but may be redundant since some irrelevant genes can be involved. The existence of irrelevant genes not only increases the computational complexity, but also impairs the efficiency of the diagnosis system with the noise. In order to achieve a higher diagnosis accuracy,

[*] [illegible] 60471054

[**] [illegible]

M. Gallagher, J. Hogan, and F. Maire (Eds.): IDEAL 2005, LNCS 3578, pp. 406–413, 2005.
© Springer-Verlag Berlin Heidelberg 2005

we should first select the informative or related genes that are discriminative between the tumor and normal or two kinds of tumor phenotypes. Meanwhile, the informative genes provide clues to medical or biological studies.

The problem of informative gene selection or discovery has been studied extensively in the past several years. In 1999, Golub et al. [1] proposed a kind of discrimination measurement or criterion on the genes via a simple statistic similar to t statistic. In their experiments, 50 most informative genes were selected and used to construct the tumor classifier with a good result on the leukemia data set. Later on, some other ranking criteria were proposed sequentially, such as F statistic method [2], mutual information scoring method [3], Markov blanket method [4], etc.. Moreover, the experiments carried out by Brown et al. [5], Dudoit[6], Furey et al. [7] and Guyon et al. [8] have shown that the support vector machine (SVM) [9] is one optimal choice for constructing the classifier or tumor diagnosis system on a microarray data set.

However, there exist two serious problems in former methods. On the one hand, these methods require a user-specified threshold on the number of informative genes. That is, they select the top k genes as the informative ones. However, it is often difficult for a user to specify such a parameter. Certainly, we can use the SVM to test the best number for k, but the testing process incurs a large computational cost. On the other hand, some methods use the t-statistic or its variations as the selection criteria. The t-statistic requires that the data follows the normal (or Gaussian) distribution. However, the assumption of normal distribution often does not hold in gene expression data [10]. In order to solve these problems, Deng et al. [10] proposed a rank sum test method that utilizes a significance level to select informative genes through the rank sum test (as a typical non-parametric statistical method) with the quality guarantee in statistics. It was shown by the experiments that the rank sum test method considerably improves the performance of tumor diagnosis on the colon and leukemia data.

In this paper, we further propose a non-parametric statistical test method, called the multi-population χ^2 test method, to select informative genes from a microarray data. It is based on the statistical multi-population χ^2 test with the sample data being grouped evenly. It is shown by the experiments that the constructed diagnosis system with the multi-population χ^2 test method can 100% correctness rate of diagnosis on colon dataset and 97.1% correctness rate of diagnosis on leukemia dataset, respectively.

2 Multi-population χ^2 Test Method and Tumor Diagnosis System via SVM

We begin to introduce the multi-population χ^2 test [11]. Suppose that there are k populations, denoted by $X_1, \cdots, X_k$, with their cumulative distribution functions denoted by $F_1(x), \cdots, F_k(x)$, respectively. From each population, we have collected a number of samples and the whole samples from these k populations, denoted by the sample set A, are divided into r exclusive groups or subsets $A_1, \cdots, A_r$ such that $A = \bigcup_{i=1}^{r} A_i,\ A_i \subset A,\ A_i \cap A_j = \emptyset (i \neq j)$. Our aim is to test the hypothesis $H0 : F_1(x) = \cdots = F_k(x)$, i.e., the identity of the distributions of these k populations.

In order to do so, we define the number n_{ij} as the number of samples from the $i-$th population in the $j-$th group, with $n_{i\,\cdot} = \sum_{j=1}^{r} n_{ij}$, $n_{\cdot\,j} = \sum_{i=1}^{k} n_{ij}$ and $n = \sum_{j=1}^{k} n_{\cdot\,j} = \sum_{i=1}^{k} n_{i\,\cdot}$. We then calculate the statistic χ_n^2 by

$$\chi_n^2 = \sum_{i=1}^{k} \sum_{j=1}^{r} \frac{(n_{ij} - n_{i.}\widehat{p}_j)^2}{n_{i.}\widehat{p}_j} \tag{1}$$

where $\widehat{p}_j = \frac{n_{\cdot j}}{n} (j = 1, 2, \cdots, r)$. In fact, it has been proved in statistics[11] that the distribution of χ_n^2 approximates $\chi^2((k-1)(n-1))$ as $n \to \infty$. So, we can use this statistic to test the hypothesis of identical distributions of the k populations via a given significance level α. That is, according to α, we get the rejection field $(\chi_\alpha^2((r-1)(k-1)), +\infty)$ or the threshold value $\chi_\alpha^2((r-1)(k-1))$. If $\chi_n^2 > \chi_\alpha^2((r-1)(k-1))$, we reject the hypothesis $H0$; otherwise, we accept it. Clearly, the multi-population χ^2 test is non-parametric.

We now consider how to utilize the multi-population χ^2 test for informative gene selection. From the perspective of statistics, the distribution of expression level of one informative gene for a tumor should be quite different between the normal and tumorous samples. That is, this difference can be checked or proved by a statistical hypothesis test method. In this way, we can apply the multi-population χ^2 test to informative gene selection on the microarray data collected from both tumorous and normal tissues. In this case, the number of populations is just 2. Correspondingly, the hypothesis becomes $H0 : F_1(x) = F_2(x)$, where $F_1(x)$ and $F_2(x)$ represent the cumulative distribution functions of expression level on the normal and tumorous samples, respectively. However, there exists one problem: how are these samples (or sample values at one gene) divided into groups (or subsets as described above)? It is clear that the number of groups should be neither too large nor too small. In fact, a small number of groups makes the division too rough, with certain differences being obscured, whereas a large number of groups makes the division too precise, with an exaggerated interference from noise. Therefore, the number of groups should be proper to the total number of samples, which will be further discussed in the following experiments. On the other hand, the number of samples in each group should also be neither too large nor too small. One particular idea is that, we can divide the samples evenly so that each group approximately has the same number of samples, which will be detailed in the following experiments. After the samples are divided into a number of groups, we can use the multi-population χ^2 test to select the informative genes only if the hypotheses on these genes are rejected.

To test the effectiveness of the multi-population χ^2 test method for informative gene selection, we build a tumor diagnosis system (i.e., a binary classifier) using the support vector machine (SVM). It has been derived from the optimal classification problem in the sample space with a finite number of samples under the statistical learning theory. Actually, there are many softwares of SVM available on the web and we will use the version OSU SVM 3.0 in the toolbox of MATLAB (It can be downloaded from http://eewww.eng.ohio-state.edu/ ˜ maj /osu svm). Three types of kernel functions are used for comparison in our experiments: (1). Linear kernel function (no kernel); (2). RBF kernel function $K(x, xi) = exp\{-\frac{|x-xi|}{\sigma^2}\}$; and (3). 3-order Polynomial kernel function $K(x, xi) = [(x \cdot xi) + 1]^3$.

3 Experimental Results and Comparisons

3.1 The Experimental Results on the Colon and Leukemia Datasets

In our experiments, we use the multi-population χ^2 test method to select the informative genes for both the colon and leukemia data sets, and then apply the SVM to constructing a tumor diagnosis system with the identified informative genes on the colon and leukemia data sets. Before the experiments, we normalize each microarray data set column by column with zero mean and unit variance, which can eliminate some possible noises in the data set.

A. The Experimental Results on the Colon Data Set

The colon cancer data set[1] contains the expression profiles of 2000 genes from 22 normal tissues and 40 tumorous tissues. In most of our experiments, we use the

Table 1. T[illegible]

K[illegible] F[illegible]	#	$\alpha = 0.1$	$\alpha = 0.05$	$\alpha = 0.01$	$\alpha = 0.001$
	8	88.9% / 408	88.9% / 265	94.4% / 95	94.4% / 21

Table 3. T[illegible] l o [illegible] d [illegible]

K [illegible] F [illegible] o [illegible]	#

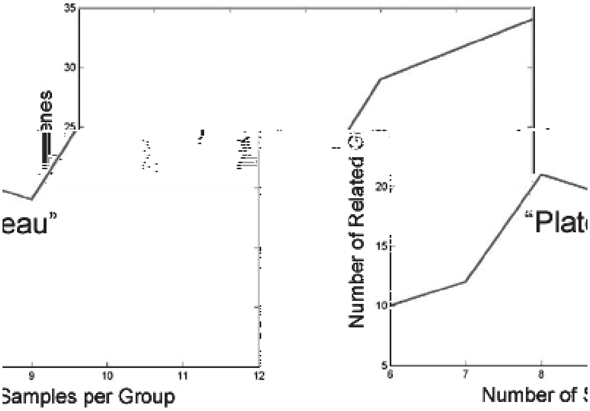

Fig. 1. T[illegible]

3.2 Comparisons with the Rank Sum Test Method

We now compare the multi-population χ^2 test method with the rank sum test method [10] on these two data sets. Obviously, these two test methods are both non-parametric, getting rid of the normality assumption on the microarray data. We implemented the rank sum test to select the informative genes on each data set and obtained the classification result through the SVMs with the same three kernel functions. The comparison results are listed in Table 4. Since we use the test set provided in the website, our results of the rank sum test method are different from those in [10].

Table 4. [illegible] χ^2 [illegible]

D[illegible]	α	0.1	0.05	0.01	0.001	0.0001
[illegible]	χ^2	92.6%	92.6%	94.4%	96.3%	100%
	[illegible]	92.6%	92.6%	92.6%	94.4%	96.3%
L[illegible]	χ^2	97.1%	97.1%	97.1%	97.1%	96.1%
	[illegible]	96.1%	97.1%	97.1%	97.1%	97.1%

[illegible] χ^2 [illegible]

From Table 4, we can find that the multi-population χ^2 test method outperforms the rank sum test method on both the diagnostic accuracy and the stability on results. However, since the multi-population χ^2 test method needs to group the samples evenly on each gene, its computational cost is higher than that of the rank sum test method. Nevertheless, this does not impair its efficiency in practice.

4 Conclusions

We have investigated the informative gene selection problem on a microarray data set via the multi-population χ^2 test. When the sample data are grouped evenly, the multi-population χ^2 test can be applied to selecting the informative genes of a tumor. The evenly grouping method on the sample data is suggested and demonstrated. By the experiments on real data sets utilizing the SVM for tumor classification or diagnosis, we show that this multi-population χ^2 test method is efficient and even better than the rank sum test method. However, there are still circumstances where the diagnostic accuracy under the selected informative genes is not satisfactory. This may be due to an unreasonable grouping on the sample data. However, in general, the multi-population χ^2 test method can reach excellent results, even without any diagnostic error when the parameters are set properly.

References

1. T. R. G[illegible], D. K. S[illegible], P. T[illegible], [illegible] al., M[illegible]: [illegible]," *Science*, 286: 531-537, 1999.
2. C. D[illegible], A[illegible]: [illegible]," *Proceedings of the 6th Annual International Conference on Computational Molecular Biology (RECOMB'02)*, [illegible] D. C., USA, A[illegible] 18-21, 2002, [illegible]: 601-680.
3. A. B[illegible], N. F[illegible]; [illegible] So[illegible] G[illegible] R[illegible]," *Agilent Technical Report*