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

Jun Luo and Jinwen Ma**

Department of Information Science
School of Mathematical Sciences and LMAM
Peking University, Beijing, 100871, China
jwma@math.pku.edu.cn

Abstract. This paper proposes a multi-population χ^2 test method for informative gene selection of a tumor from microarray data based on the statistical multi-population χ^2 test with the sample data being grouped evenly. To test the effectiveness of the multi-population χ^2 test method, we use the support vector machine (SVM) to construct a tumor diagnosis system (i.e., a binary classifier) based on the identified informative genes on the colon and leukemia data. 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.

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,

* This work was supported by the Natural Science Foundation of China for Project 60471054

** The corresponding author

M. Gallagher, J. Hogan, and F. Maire (Eds.): IDEAL 2005, LNCS 3578, pp. 406–413, 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

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

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 training set (22 normal and 22 tumorous) and the test set (18 tumorous) provided by the web site. The parameters in the SVM are set as u_PolySVC(⋆,⋆,3,0.001) for the 3-order polynomial kernel function, and u_ RbfSVC(⋆,⋆,0.01,100) for the RBF kernel function.

To utilize the multi-population χ^2 test method, we now group the samples evenly. On each gene, we first put the 22 normal and 40 tumorous expression values together. Suppose that we expect each group to contain 9 sample values. Then, we select six real numbers according to which the whole real region R can be divided into 7 intervals. By adjusting these six numbers properly, we can divide the 62 sample values into seven groups that contain 9, 9, 9, 9, 9, 9, and 8, respectively. In this way, the 62 sample values at each gene are grouped evenly, with each group containing almost 9 sample values.

From Table 1, we can find that the SVM with the multi-population χ^2 test method can lead to a very good classification accuracy on the colon data set when the significance level is 0.001 or 0.01 and the number of sample values per group is properly selected. From the perspective of kernel functions, the 3-order polynomial function performs best with the two 100% classification accuracies (at $\alpha = 0.001$), which outperform the SVM on the original colon data set. Actually, the classification accuracies of the SVM with the three kernel functions on the original colon data set are 94.4%, 94.4%, 94.4%, respectively.

In order to illustrate the potential relation between the number of sample values per group and the classification effectiveness, we provide the average classification accuracies over the three kernel functions and the number of sample values per group under the significance level 0.001 in Table 2.

[1] retrieved from
`http://microarray.princeton.edu/oncology/affydata/index.html`

Table 1. The result on the colon data set

Kernel Functions	#	$\alpha = 0.1$	$\alpha = 0.05$	$\alpha = 0.01$	$\alpha = 0.001$
Linear	8	88.9% / 408	88.9% / 265	94.4% / 95	94.4% / 21
	9	88.9% / 415	88.9% / 254	94.4% / 93	94.4% / 19
	10	88.9% / 409	88.9% / 276	94.4% / 87	94.4% / 29
RBF	8	94.4% / 408	94.4% / 265	94.4% / 95	94.4% / 21
	9	94.4% / 415	94.4% / 254	94.4% / 93	94.4% / 19
	10	94.4% / 409	94.4% / 276	94.4% / 87	94.4% / 29
3-order Polynomial	8	88.9% / 408	94.4% / 265	88.9% / 95	100% / 21
	9	94.4% / 415	94.4% / 254	94.4% / 93	100% / 19
	10	88.9% / 409	94.4% / 276	94.4% / 87	94.4% / 29

In this and the following tables, the symbol # represents the number of sample values per group. The two numbers on the sides of "/" represent the classification accuracy on the test set or the diagnosis accuracy, and the number of informative genes, respectively. α is the significance level for the multi-population χ^2 test.

Table 2. The relation between the classification accuracy and the number of sample vaules per group

#	6	7	8	9	10	12
Accuracy	87.0%	94.4%	96.3%	96.3%	94.4%	94.4%

From Table 2, we can find that the multi-population χ^2 test method reaches the optimum result when there are 8 or 9 sample values per group. Moreover, from the stability of classification accuracy, it performs better at 9 sample values per group, which is further shown by other experiments on a plenty of data sets constructed by randomly selecting 44 samples from the colon data set as a new training set and leaving the other 18 as a new test set.

B. The Experimental Results on the Leukemia Data Set

The leukemia cancer data set[2] consists of expression profiles of 7129 genes from 47 acute lymphoblastic leukemia (ALL) and 25 acute myeloid leukemia (AML) samples. Specifically, the training set contains 38 samples (27 ALL and 11 AML), while the test set contains 34 samples (20 ALL, 14 AML). Actually, the training and test sets are provided at the web site. The parameters in the SVM are selected as above.

From Table 3, we can find that the classification accuracy of the multi-population χ^2 test method on the leukemia data set is good and stable, compared to that on the original leukemia data set with 94.1%, 85.7% and 97.1% to three kernel functions respectively. Also, the number of related genes is reduced to a low level. We further apply our method to a plenty of data sets constructed using the same method on colon data set, and find that on some data sets, our method can reach the optimal average accuracy 100% over three kernel functions. Our

[2] retrieved from `http://www-genome.wi.mit.edu/cgi-bin/cancer/datasets.cgi`

Table 3. The result on the leukemia data set

Kernel Functions	#	$\alpha = 0.1$	$\alpha = 0.05$	$\alpha = 0.01$	$\alpha = 0.001$
Linear	8	97.1% / 1733	97.1% / 1039	97.1% / 557	97.1% / 260
	9	97.1% / 1948	97.1% / 1356	97.1% / 619	97.1% / 274
	10	97.1% / 1807	97.1% / 1320	97.1% / 605	97.1% / 257
RBF	8	97.1% / 1733	97.1% / 1039	97.1% / 557	97.1% / 260
	9	97.1% / 1948	97.1% / 1356	97.1% / 619	97.1% / 274
	10	97.1% / 1807	97.1% / 1320	97.1% / 605	97.1% / 257
3-order Polynomial	8	94.1% / 1733	97.1% / 1039	97.1% / 557	97.1% / 260
	9	97.1% / 1948	97.1% / 1356	97.1% / 619	97.1% / 274
	10	97.1% / 1807	97.1% / 1320	97.1% / 605	97.1% / 257

experiment results also show that the multi-population χ^2 test method performs best when the number of sample values per group is 9.

C. Further Discussions and Remarks

According to the experimental results, we give some further discussions and remarks on the multi-population χ^2 test method as follows.

(1). In general, we can select 0.01 as the best choice of the significance level for the multi-population χ^2 test method. The slight difference between the optimal significance levels on the colon and leukemia data sets may be owing to the characteristics of the distributions of the two data sets. However, the multi-population χ^2 test method is rather good on both the colon and leukemia data set at 0.01 significance level.

(2). As to the number of sample values per group, it should be determined through the experiments. However, we can set it 9 as an initial value and make some adjustments based on the experiment results.

(3). The principle of grouping the samples evenly is demonstrated to be effective in our experiments. However, it does not mean that this is just the optimal one. Actually, the optimal division of the sample values for the multi-population χ^2 test method should be theoretically studied in the future.

(4). By the experiments, we can find that the performance of the the multi-population χ^2 test method is sensitive to the number of sample values per group as well as the significance level α. However, it is interesting that when we chart the diagram with the number of sample values per group as x-coordinate and the number of identified informative genes as y-coordinate, we can discover that the optimum number "9" of samples per group is located near the so-called "plateau" of the polygonal line, which is shown in Fig. 1. If it can be proved theoretically, we will have a good method to get the best number of the sample values per group in the general case.

(5). Our experiments indicate that the classification accuracy of our method is related to which samples are used for training and for testing. It is possible to select proper training set to construct our diagnosis system.

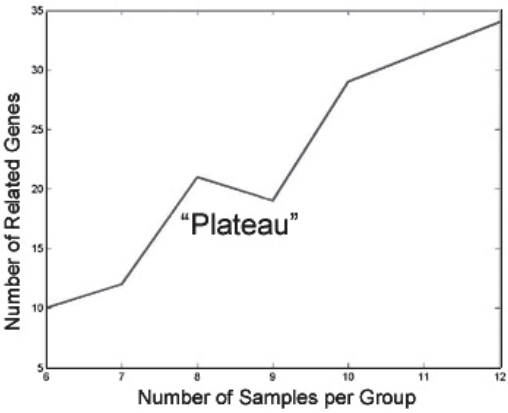

Fig. 1. The plateau for the best number of sample values per group

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. Comparison result between the multi-population χ^2 test method and the rank sum test method

Data set	α	0.1	0.05	0.01	0.001	0.0001
Colon cancer	χ^2	92.6%	92.6%	94.4%	96.3%	100%
	rank sums	92.6%	92.6%	92.6%	94.4%	96.3%
Leukaemia	χ^2	97.1%	97.1%	97.1%	97.1%	96.1%
	rank sums	96.1%	97.1%	97.1%	97.1%	97.1%

For the multi-population χ^2 test method, we use the experimental result with 9 sample values per group.

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. Golub, D. K. Slonim, P.Tamayo, et al., "Molecular classification of cancer: class discovery and class prediction by gene expression monitoring," *Science*, 286: 531-537, 1999.
2. C. Ding, "Analysis of gene expression profiles: class discovery and leaf ordering," *Proceedings of the 6th Annual International Conference on Computational Molecular Biology (RECOMB'02)*, Washington, D. C., USA, April 18-21, 2002, pp: 601-680.
3. A. Ben-Dor, N. Friedman, and Z. Yakhini, "Scoring Genes for Relevance," *Agilent Technical Report*, no. AGL-2000-13, 2000.
4. E. P. Xing, M¿ I. Jordan, and R. M. Karp, "Feature selection for high-dimensional genomic microarray data," *Proceedings of the 18th International Conference of Machine Learning (ICML'01)*, Massachusetts, USA, June 28-July 1, 2001, pp: 601-608.
5. BrownM. P. S. Brown, W. N. Grundy, D. Lin, N. Cristianini, et al., "Knowledge-based analysis of microarray gene expression data by using support vector machines," *Proc. Nat'l Acad Sci*, 97(1): 262-267, 2000.
6. D. S. Dudoit, J. Fridyand, and T. P. Speed, "Comparison of discrimination methods for the classification of tumor using gene expression data," *Univ. of California, Dept. of Statistics, Tech Report*, no.576, 2000.
7. T. Furey, N. Cristianini, N. Duffy, et al., "Support vector machine classification and validation of cancer tissue samples using microarray expression data," *Bioinformatics*, 16(10): 909-914, 2000.
8. Guyon I. Guyon, J. Weston, S. Barnhill, V. Vapnik, "Gene selection for cancer classification using support vector machine," *Machine Learning*, 46(1/3): 389-422, 2002.
9. V. Vapnik, *Statistical Learning Theory*, New Your: Wiley, 1998.
10. L. Deng, J. Ma, and J. Pei, "Rank sum method for related gene selection and its application to tumor diagnosis," *Chinese Science Bulletin,* 49(15): 1652-1657, 2004.
11. M. Hollander and D. A. Wolfe, *Nonparametric statistical method*, New York: Wiley, 1999.