Non-parametric Statistical Tests for Informative Gene Selection*

Jinwen Ma, Fuhai Li, and Jianfeng Liu

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

Abstract. This paper presents two non-parametric statistical test methods, called Kolmogorov-Smirnov (KS) and U statistic test methods, respectively, for informative gene selection of a tumor from microarray data, with help of the theory of false discovery rate. To test the effectiveness of these non-parametric statistical test methods, 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 tumor diagnosis system with both the KS and U statistic test methods can reach a good prediction accuracy on both the colon and leukemia data sets.

1 Introduction

With the rapid development of DNA microarray technology, we can now get rapid, large-scale screening for patterns of gene expression. These microarray data corresponding to certain biological feature are generally represented by a gene expression genes of a tumor from microarray data. Essentially, the informative genes can not only provide valuable information for discovering the crucial reasons of the tumor as well as the treatment methods, but also support to construct an efficient tumor diagnosis system from their expression levels directly without any influence of the other irrelevant genes. In fact, these have been already many methods for informative gene selection. However, most of the existing methods are based on ranking the important genes according to a certain criterion which requires that the data follows a normal distribution (e.g., [1]-[4]). But the normality assumption is often violated in real data sets [5]. In order to avoid the normality assumption, a rank sum test method (as a typical non-parametric statistical method) has been suggested to select informative genes with a considerably improved performance of tumor diagnosis on the colon and leukemia data [5].

In this paper, we further study this problem via two other non-parametric statistical tests, called Kolmogorov-Smirnov (KS) test and U statistic test, respectively, for informative gene selection with help of the theory of false discovery rate [6]-[8]. To test the

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

J. Wang, X. Liao, and Z. Yi (Eds.): ISNN 2005, LNCS 3498, pp. 697–702, 2005.

effectiveness of these non-parametric statistical test methods, 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. Our experiments show that the constructed tumor diagnosis system with both the KS and U test methods through SVM can reach a good prediction accuracy on both the colon and leukemia data sets.

2 The KS and U Statistic Tests

In this section, we introduce the KS and U statistic tests as the bases for informative gene selection. We consider a data set of two classes, i.e., $\mathcal{S} \rightarrow \{x_{11}, \cdots, x_{1n}\ x_{21}, \cdots, x_{2m}\}$ in which $x_{11}, \cdots, x_{1n}$ come from one population being subject to the probability distribution $F_1(x)$, while $x_{21}, \cdots, x_{2m}$ come from another population being subject to the probability distribution $F_2(x)$. We need to make a non-parametric statistical test of hypothesis $H_0 \quad F_1(x) \rightarrow F_2(x)$ without any information on these two probability distributions.

2.1 The KS Test

The KS test is a typical non-parametric statistical test based on the ranks of the observations. Actually, we first rank all the observations in $\mathcal{S}$ in an ascending order. Then, each observation has a ranking number, being called its rank in statistics. We now define a discrete probability distribution $F \rightarrow \{f_1, f_2, \cdots, f_{n+m}\}$ according to the ranks $\{k_1, k_2, \cdots, k_n\}$ of the observations $\{x_{11}, \cdots, x_{1n}\}$ of the first class as follows (assuming that $k_1 \leq k_2 \leq \cdots \leq k_n$).

$$f_j \rightarrow \begin{cases} 0, \text{ if } j < k_1; \\ \frac{i}{n}, \text{ if } j \rightarrow k_i, i \rightarrow 1, 2, \cdots, n; \\ \frac{i}{n}, \text{ if } k_i < j < k_{i+1}, i \rightarrow 1, 2, \cdots, n-1; \\ 1, \text{ if } j > k_n, \end{cases} \tag{1}$$

for $j \rightarrow 1, 2, \cdots, n+m$. In the same way, we can define $G \rightarrow \{g_1, g_2, \cdots, g_{n+m}\}$ according to the ranks of the observations $\{x_{21}, \cdots, x_{2n}\}$ of the second class. Finally, we construct the KS statistic D_{nm} as follows.

$$D_{nm} \rightarrow max\{|f_i - g_i| \quad i \rightarrow 1, 2, \cdots, n+m.\}. \tag{2}$$

For a significance level $\alpha > 0$, we can get the threshold value $V(\alpha)$ of D_{nm} from a general KS test table. If $D_{nm} > V(\alpha)$, we reject H_0; otherwise, we accept H_0.

2.2 The U Statistics Test

Supposing that $n \leq m$ and R_1 is the sum of ranks of the observations of the first class, i.e., $R_1 \rightarrow \sum_{i=1}^{n} k_i$, we have the following (Mann-Whitney) U statistic:

$$U \rightarrow nm + \frac{n(n+1)}{2} - R_1. \tag{3}$$

On the other hand, if $n > m$, we let R_1 be the sum of ranks of the observations of the second class and the U statistic is defined by

$$U = nm + \frac{m(m+1)}{2} - R_1. \tag{4}$$

When H_0 holds, the U statistic tends to be subject to a normal distribution with the mean $\mu_U = nm/2$ and the standard variance $\sigma_U = \sqrt{nm(n+m+1)/12}$ as the number of observations in each class becomes large. That is,

$$Z = \frac{U - \mu_U}{\sigma_U} = \frac{U - nm/2}{\sqrt{nm(n+m+1)/12}} \sim N(0,1). \tag{5}$$

Thus, we can make the statistical test on H_0 from Z. For a significance level $\alpha > 0$, we can get the threshold value $V(\alpha)$ from the standard normal distribution function. In the same way, if $Z > V(\alpha)$, we reject H_0; otherwise, we accept H_0.

3 Informative Gene Selection and Tumor Diagnosis System via SVM

We now consider the informative gene selection based on these two non-parametric statistical tests. For informative gene selection, we can make a statistical test on each gene with its expressions on the two classes of samples(tumor and normal tissues or two kinds of tumor tissues). Clearly, if a gene is informative to the tumor, the probability distributions on the two classes should be quite different; otherwise they should be the same. On the other hand, we generally know nothing about the structures of these distributions. In these situations, it is reasonable to apply the KS or U statistic test to the informative gene selection via a microarray data set. That is, on each gene, when H_0 is rejected by the test of hypothesis, we consider this gene is informative; otherwise, we consider it is not informative.

Generally, there are thousands of genes in a microarray data and thus the informative gene selection is a large multiple-hypothesis testing problem. In this case, we must control the false discovery rate (FDR), i.e., the ratio of the number of falsely discovered (or selected) informative genes over that of all discovered informative genes [6]-[7]. Only when the FDR is controlled in a certain degree, we are sure that the informative gene selection is reliable. In order to do so, Storey and Tibshirani [8] proposed a q-value method which can be used to select the informative genes with the FDR being controlled directly. Actually, we first make the KS or U statistic test for each gene from the microarray data independently. Then, we can calculate the p-values of these statistical tests according to the statistics, say $p_1 \leq p_2 \leq \cdots, \leq p_n$ in an ascending order. By the q-value estimation algorithm given in [8] (which is now available on the web site: http://faculty.washington.edu/ jstorey/qvalue.), we can get their corresponding q-values, say $q_1 \leq q_2 \leq \cdots, \leq q_n$. Finally, if we want to control the FDR by $\alpha > 0$, we need only to select the informative genes by the selection criterion that $q_i \leq \alpha$. We will use this q-value method via the KS and U statistic tests to select the informative genes from a microarray data set.

To test the effectiveness of our non-parametric statistical test method for informative gene selection, we build a tumor diagnosis system (i.e., a binary classifier) using the support vector machine (SVM) [9]. Actually, SVM has been proved to be the most effective machine learning algorithm for processing large scale gene expression profiles. It has been derived from the optimal classification problem in the sample space with a finite number of samples. There are many softwares of SVM available on the web and we will use the software of SVM in Matlab. For comparison, we also try the following kinds of support vector machines: (1). Radial basis function SVM (RBF kernel); (2). 3-poly SVM (cubic polynomial kernel); and (3). Linear SVM (no kernel).

4 Experiment Results

We test the effectiveness of our non-parametric statistical test methods for informative gene selection through SVM for tumor diagnosis using two real data sets as follows.

The colon cancer data set. It contains the expression profiles of $2,000$ genes in 22 normal tissues and 40 colon tumor tissues (retrieved from http://micro-array.princeton.edu /oncology/database.html). In our experiment, we use the train set (22 normal and 22 tumorous tissues) and test set (1 tumorous tissues) provided at the web site.

The leukemia cancer data set. It consists of $,12$ genes in 4 acute lymphoblastic leukemia (ALL) and 2 acute myeloid leukemia (AML) samples (retrieved from http://www-genome.wi.mit.edu/cgi-bin/cancer/datasets.cgi). In our experiments, we use the train set (2 ALL, 11 AML) and the test set (20 ALL, 14 AML) provided at the web site.

We use MATLAB toolbox OSU_SVM 3.0 (which can be obtained from the web site: `http://www.ece.osu.edu/~maj/osu_svm/`.) to implement the three kinds of SVMs. In the radial basis function and 3-poly SVMs, there are two parameters γ and C. In our experiments, we generally select $\gamma = 0.02$ and $C = 0.0$ on the colon dataset, and $\gamma = 0.002$ and $C = 10$ on the leukemia dataset. Sometimes, they are slightly adjusted to get the best performance of the SVMs. For the KS and U statistic test methods, we try three FDR α: 0.0 , 0.0 , and 0.0 , respectively. The informative gene selection returns different numbers of informative genes on both the colon and leukemia data sets with slightly different prediction accuracies on the test sets. The results of the two non-parametric statistical test methods on the colon and leukemia data sets are given in Table 1-4, respectively.

From Tables 1& 2, we find that on the colon data set, the SVM tumor diagnosis system can reach an optimum prediction accuracy 1 by both the KS and U statistic test

Table 1. The result of the KS test method on the colon data set. Here and in the following tables, each number in the second to the fourth rows represents the prediction accuracy of the SVM on the test set.

α (# informative genes)	0.03 (130)	0.05 (187)	0.07 (216)
RBF SVM	0.9444	1.0000	1.0000
3-poly SVM	0.9444	0.9444	0.9444
Linear SVM	0.9444	0.9444	0.9444

Table 2. The result of the U test method on the colon data set.

α (# informative genes)	0.03 (130)	0.05 (187)	0.07 (216)
RBF SVM	0.9444	1.0000	1.0000
3-poly SVM	0.9444	0.9444	0.9444
Linear SVM	0.9444	0.9444	0.9444

Table 3. The result of the KS test method on the leukemia data set.

α (# informative genes)	0.03 (775)	0.05 (946)	0.07 (1108)
RBF SVM	0.9706	0.9706	0.9706
3-poly SVM	0.9706	0.9706	0.9706
Linear SVM	0.9706	0.9706	0.9706

methods. From Tables 3& 4, we further find that on the leukemia data set, the prediction accuracy of the SVM tumor diagnosis system by using the informative genes of the KS test method is always $0.\quad 0$, where only one prediction error happens in our test experiments. As for the U statistic test method, the optimum prediction accuracy of the SVM tumor diagnosis system is even 1. Therefore, the SVM tumor diagnosis system with both the KS and U statistic test methods for informative gene selection can reach a good prediction accuracy on both the colon and leukemia data sets.

We also make the experiments on these two microarray data sets by the rank sum test method proposed in [5] with help of false discovery rate theory. It is found by the experiments that the rank sum method is as good as the U statistic test method and better than the original one to select the informative genes directly by the test of hypothesis. It is also found by the experiments that these non-parametric statistical test methods considerably outperforms the typical ranking methods [1]-[4] through the same SVM software.

5 Conclusions

We have investigated the informative gene selection problem from a microarray data set via non-parametric statistical test with help of the theory of false discovery rate theory. We apply the Kolmogorov-Smirnov (KS) and U statistic tests and the q-value algorithm to the informative gene selection of a tumor and use the support vector machine (SVM) to construct a tumor diagnosis system with the identified informative genes on the colon and leukemia data. Our experiments show that the constructed tumor diagnosis system with both the KS and U statistic test methods can lead to a good prediction accuracy on both the colon and leukemia data sets.

Table 4. The result of the U statistic test method on the leukemia data set.

α (# informative genes)	0.03 (1042)	0.05 (1255)	0.07 (1416)
RBF SVM	0.9706	0.9706	0.9706
3-poly SVM	0.9706	0.9706	1.0000
Linear SVM	0.9706	0.9706	0.9706

References

1. Alon, U., Barkai, N., Notterman, D.A., et al.: Broad Patterns of Gene Expression Revealed by Clustering Analysis of Tumor and Normal Colon Tissues Probed by Oligonucleotide Arrays. it Proc Natl Acad Sci USA, **96** (1999) 6745-6750
2. Golub, T.R., Slonim, D.K. Tamayo, P., et al.: Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring. it Science, **286** (1999) 531-537
3. Furey, T.S., Cristianini, N., Duffy, N., Bednarski, D.W., Schummer, M., and Haussler, D.: Support Vector Machine Classification and Validation of Cancer Tissue Samples Using Microarray Expression Data. it Bioinformatics, **16** (2000) 906-914
4. Ben-Dor, A., Friedman, N., and Yakhini, Z.: Scoring Genes for Relevance. Agilent Technical Report, no. AGL-2000-13(2000)
5. Deng, L., Ma, J., and Pei, J.: Rank Sum Method for Related Gene Selection and Its Application to Tumor Diagnosis . Chinese Science Bulletin, **49** (2004) 1652-1657
6. Benjamini, Y., and Hochberg, Y.: Controlling the False Discovery Rate: a Practical and Powerful Approach to Multiple Testing . J. R. Statist. Soc. B, **57** (1995) 289-300
7. Storey, J.D.: A Direct Approach to False Discovery Rates. J. R. Statist. Soc. B, **64** (2002) 479-498
8. Storey, J.D., and Tibshirani, R.: Statistical Significance for Genomewide Studies. Proc Natl Acad Sci USA, **100** (2003) 9440-9445
9. Vapnik, V.: The Nature of Statistical Learning Theory. Springer,New York (1995)