An Information Criterion for Informative Gene Selection*

Fei Ge and Jinwen Ma**

Department of Information Science
School of Mathematical Sciences and LMAM
Peking University, Beijing 100871, China

Abstract. It is important in bioinformatics research and applications to select or discover informative genes of a tumor from microarray data. However, most of the existing methods are based on models which assume that the gene expressions are normal distributed, which is often violated in practice. In this paper, we propose an information criterion for informative gene selection by ranking the genes with the Kullback-Leiber discrimination information of two probability distributions of the expression levels on the tumor and normal (or another type of tumor) samples. We use support vector machine (SVM) to construct the tumor diagnosis system using certain top informative genes. The experiments on two well-known data sets (colon data and leukemia data) show that the information criterion can make the tumor diagnosis system reach 94.4% and 100% correctness rate of diagnosis on these two datasets, respectively.

1 Introduction

With the development of DNA microarray technology, we can now quickly obtain large-scale gene expression profiles, i.e., the microarray data, which provide important and detailed evidences to health state of human tissues for disease analysis and diagnosis. Moreover, as gene studies are shifting from DNA sequencing to function analysis, the microarry data will play a more important role since they can help us to discover and understand the biological characteristics from a group of genes.

The microarray data can be represented by a matrix $\mathbf{A} = (a_{ij})_{N\times k}$, where the i-th row corresponds to gene i, the j-th column corresponds to sample j, and a_{ij} denotes the mRNA expression level of gene i in sample j. Generally, it is a large matrix with thousands of rows according to such a number of genes in a microarray chip. As for tumor diagnosis, each sample is labelled to be of a certain tumor or not and the tumor diagnosis system can be trained with the supervised learning on these data. However, the computing complexity due to the high dimension of the data has made it hard to train the learning system. Moreover, not all these genes are relevant to the tumor and the irrelevant genes will contribute nothing to the learning system but noise. In order to achieve a high diagnosis accuracy, we should first select the informative genes that are discriminative among the tumor and normal phenotypes. Meanwhile, the informative genes provide clues to medical or biological studies.

* This work was supported by the Natural Science Foundation of China for Project 60471054.
** The corresponding author, Email:jwma@math.pku.edu.cn.

J. Wang, X. Liao, and Z. Yi (Eds.): ISNN 2005, LNCS 3498, pp. 703–708, 2005.

The problem of informative gene selection has been studied extensively in the last five years. Golub et al.[1] proposed a kind of discrimination measurement on the genes via a simple statistic: $(\mu_1 - \mu_2)/(\sigma_1 + \sigma_2)$, similar to the t-statistic from expression levels in two different classes. The t-statistic method and its variations are the most popular in gene selection[2][3][4]. The statistic value is considered as a score for each gene. Then all the genes are ranked according to their scores, and the group of top genes are candidates for informative genes. The most serious problem for t-statistic method is that it assumes the expression levels of each gene follow a normal distribution. However, this is not always true in practice[5].

Many other scores have been proposed, such as NToM score[6] and BSS/WSS score[7]. They even don't consider the probability distributions of two-class gene expressions. According to the dimension reduction or filtering theory, some other methods are also proposed for informative gene selection, e.g., [8] [9] [10]. However, these methods are not only lack of theoretic foundation on informative gene selection, but also difficult to deal with, since the dimension of the data, i.e., the number of genes, is so large.

In this paper, we propose an information criterion for informative gene selection by measuring the discriminate power of a gene with the Kullback-Leiber discrimination information between the probability distributions of the expression level on the tumor and normal (or another type of tumor) samples, which doesn't need the normality assumption on the expression levels. We then construct a tumor diagnosis system by the support vector machine trained on the data set of certain top informative genes. In our experiments, the information criterion can make the tumor diagnosis system reach 94.4% correctness rate of diagnosis on colon dataset and 100% correctness rate of diagnosis on leukemia dataset, respectively.

In the sequel, we propose our information criterion for informative gene selection in Section 2. In Section 3, the experiments are conducted to demonstrate the information criterion, being compared with the t-statistic method. A brief conclusion is made in Section 4.

2 The Information Criterion

From the point of view of probability theory, the expression level of an informative gene should subject to different probability distributions on the tumor and normal tissues, respectively. Moreover, as this gene becomes more discriminative to the tumor, the two probability distributions should be more different. Otherwise, the expression level of an irrelevant gene should subject to the same probability distribution on both the tumor and normal tissues. That is, the two probability distributions become the same in this case. Based on this fact, we can use Kullback-Leiber discrimination information (also known as Kullback-Leiber divergence) of these two probability distributions to evaluate the discriminative power of the gene to the tumor.

If $p(x)$ and $q(x)$ are the probability distributions of the expression level on tumor and normal tissues, respectively, the Kullback-Leiber discrimination information is computed by $K(p||q) = \int p(x) \log(p(x)/q(x))dx$. Since there are only a finite number of data available, we can only use the empirical distributions instead of $p(x)$ and $q(x)$. For clarity, we rewrite the gene expression data by the following matrix:

$$\left(\begin{array}{ccc|ccc} a_{11} & \cdots & a_{1m} & b_{11} & \cdots & b_{1n} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ a_{N1} & \cdots & a_{Nm} & b_{N1} & \cdots & b_{Nn} \end{array}\right), \tag{1}$$

where a_{ij} is the expression level of gene i for the j-th tumor sample, while each b_{ij} is the expression level of gene i for the j-th normal sample.

For a set of i.i.d. sample data $x_1, x_2, \cdots, x_N$, the empirical distribution can be estimated by

$$\hat{p}_N(x) = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{h}K\left(\frac{x - x_i}{h}\right), \tag{2}$$

where $K(x)$ is called the *kernel function*, h is the *bandwidth* of Parzen window (related with N). In fact, Parzen[11] proved that under certain regular conditions, $\hat{p}_N(x)$ is an asymptotically unbiased estimator of the actual probability density function. In our experiments we use Gaussian kernel function $K(x) = \frac{1}{\sqrt{2\pi}}\mathrm{e}^{-x^2}$ and choose h by $\hat{h}_{NS} = (\frac{4}{3N})^{1/5}S$, where S denotes the standard deviation of the sample data.

According to Eq.(2), from the data $a_{i1}, a_{i2}, \cdots, a_{im}$, we can get $\hat{p}_i(x)$, the empirical distribution of the expression level of gene i on the tumor sample data, and from the data $b_{i1}, b_{i2}, \cdots, b_{in}$ we can get $\hat{q}_i(x)$, the empirical distribution for the expression of gene i on the normal sample data. For symmetry, we define

$$K_i = K(\hat{p}_i||\hat{q}_i) + K(\hat{q}_i||\hat{p}_i) = \int_{-\infty}^{+\infty}(\hat{p}_i(x) - \hat{q}_i(x))\log\frac{\hat{p}_i(x)}{\hat{q}_i(x)}dx \tag{3}$$

as the discriminative power of gene i to the tumor. Then, all genes can be ranked in the descending order of this criterion. We can select a certain number of genes ranked first as the informative genes and discard the rest ones.

After we select a group of informative genes, we can construct the tumor diagnosis system by a binary (or bipolar) supervised classifier trained with the expression levels of these informative genes. Since SVM owns a better generalization ability on a small sample set[12], we use it as our tumor diagnosis system, with radial basis function as the kernel function.

For comparison, we also give the t-statistic method for informative gene selection. Actually, the t-statistic method ranks genes in the descending order of the absolute value of t-statistic calculated from data samples in two classes as follows:

$$t_i = (\bar{a}_i - \bar{b}_i)/(\sqrt{\frac{s_{a,i}^2}{m} + \frac{s_{b,i}^2}{n}}), \tag{4}$$

where $\bar{a}_i$ and $s_{a,i}^2$ are the mean and variance of gene i's expression level on the tumor samples, respectively, while $\bar{b}_i$ and $s_{b,i}^2$ are the mean and variance of gene i's expression level on the normal samples, respectively. Traditionally, t-statistic is used to test whether two normal distributions have the same mean.

3 Experiment Results

We test the effectiveness of the information criterion for informative gene selection through the SVM for tumor diagnosis using the two real data sets as follows:

The colon cancer dataset[1] contains the expression profiles of 2000 genes in 22 normal tissues and 40 tumor tissues[2]. In our experiments, we randomly select 44 samples as the training set, and use the other 18 samples as test data.

The leukemia dataset[2] consists of expression profiles of 7129 genes from 47 acute lymphoblastic leukemia (ALL) and 25 acute myeloid leukemia (AML) samples[1]. Specifically, the training dataset contains 38 samples (27 ALL and 11 AML), while the test dataset contains 34 samples (20 ALL, 14 AML).

We calculate all K_i and then rank the genes with these values. As shown above, genes with larger K_i will have stronger discriminate powers. In the experiments, we select the top k genes with the highest ranks, and train the SVM with the expression levels of these k genes. We test the performance with k gradually increasing from some initial value k^0.

We use MATLAB toolbox OSU SVM 2.0 (which can be obtained from `http://eewww.eng.ohio-state.edu/~maj/osu_svm/`) to implement the SVM with the RBF kernel functions. In this situation, there are only two parameters γ and C to be determined. Actually, the selection of γ and C affects the performance of the SVM. In the experiments, we take a grid search procedure from 16×16 pairs of γ and C, and choose the optimal values by cross validation. Then, the SVM is trained for the tumor diagnosis. The correctness rates of tumor prediction on the two datasets with k from 1 to 300 are sketched in Figs 1 & 2, respectively.

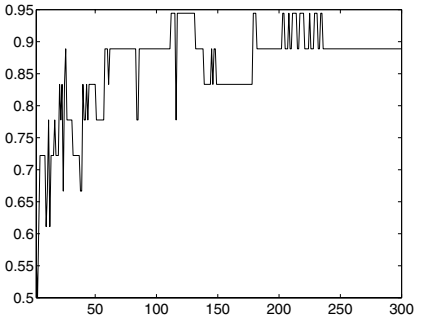

Fig. 1. The correctness rate of tumor prediction on colon cancer data

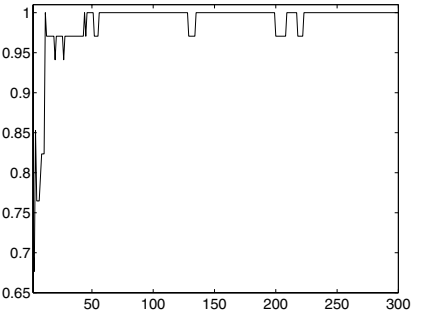

Fig. 2. The correctness rate of tumor prediection on leukemia data

From the curves in Figs 1 & 2, we can observe that as the number k of selected informative genes increases from the beginning, the prediction accuracy tends to increase, too. When the highest correctness rate of tumor prediction is reached, there exists certain interval in which each k can maintain a good prediction accuracy, although the correctness rate may fluctuate slightly. But if k further increases, the correctness rate begins to decrease. This means that the information criterion is significant on the selection of informative genes of a tumor. It can be also found that when the number of informative genes is properly selected, the information criterion can make the tumor

[1] retrieved from `http://microarray.princeton.edu/oncology/database.html`
[2] retrieved from `http://www-genome.wi.mit.edu/cgi-bin/cancer/datasets.cgi`

diagnosis system reach 94.4% correctness rate of diagnosis on the colon dataset and 100% correctness rate of diagnosis on the leukemia dataset, respectively.

On the other hand, the experiment results show that the number of informative genes should be carefully selected and is essential to the performance of SVM for tumor prediction. With the more or less genes selected, the SVM would result in a drop of prediction accuracy. However, there exist a number of weakly related genes that are not very sensitive to the performance of the SVM for tumor prediction. It can be found by the experiments that the optimal number of informative genes of colon cancer is about 125, while that of leukemia is about 75.

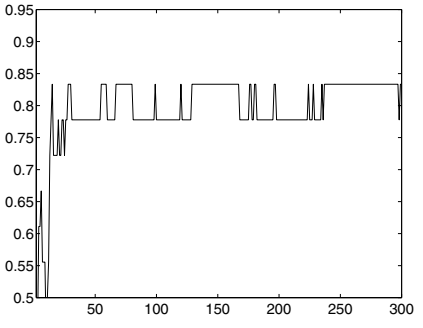

Fig. 3. The correctness rate of tumor prediction on colon cancer data

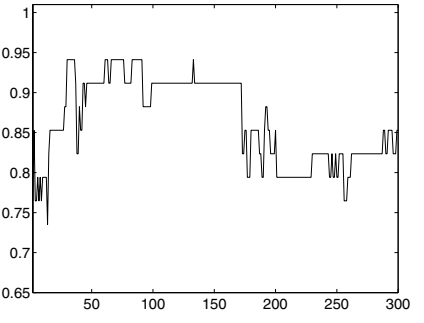

Fig. 4. The correctness rate of tumor prediection on leukemia data

For comparison, we replace the gene selection criterion with the t-statistic and the results on the two datasets are shown in Figs 3 & 4. Clearly, our information criterion is superior to the t-statistic method on the tumor diagnosis.

4 Conclusions

We have proposed an information criterion for informative gene selection of a tumor according to the Kullback-Leiber discriminative information between the two probability distributions of the expression levels on the tumor and normal tissues. By experiments on real data sets through the SVM for tumor diagnosis, we show that the information criterion is significant and even better than the t-statistic method. Moreover, the experiments also show that the information criterion can make the tumor diagnosis system reach 94.4% correctness rate of diagnosis on colon dataset and 100% correctness rate of diagnosis on leukemia dataset, respectively.

References

1. Golub, T.R., Slonim, D.K., Tamayo, P., et al.: Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring. Science, **286** (1999) 531-537

2. 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. Proceedings of the National Academy of Sciences, **96** (1999) 6745–6750
3. Furey, T.S., Cristianini, N., Duffy, N., et al.: Support Vector Machine Classification and Validation of Cancer Tissue Samples Using Microarray Expression Data. Bioinformatics, **16** (2000) 906-914
4. Nguyen, D., and Rocke, D.: Tumor Classification by Partial Least Squares Using Microarray Gene Expression Data. Bioinformatics, **18** (2002) 39-C50.
5. Deng, L., Ma, J., and Pei, J.: Rank Sum Method for Related Gene Selection and Its Application to Tumor Diagnosis. Chinese Science Bulletin, **15** (2004)1652-1657
6. Ben-Dor, A., Friedman, N., and Yakhini, Z.: Scoring Genes for Relevance. Agilent Technical Report, no. AGL-2000-13 (2000)
7. Dudoit, S., Fridlyand, J., Speed, T.: Comparison of Discrimination Methods for The Classification of Tumor Using Gene Expression Data. Journal of American Statistical Association, **97** (2002)77–87
8. Liu, J., Iba, H., and Ishizuka, M.: Selecting Informative Genes with Parallel Genetic Algorithms in Tissue Classification. Genome Informatics, **12** (2001) 14-23
9. Xiong, M., Li, W., Zhao, J., et al.: Feature (Gene) Selection in Gene Expression-based Tumor Classification. Mol Genet Metab, **73** (2001) 239-47
10. Hellem Bø , T., and Jonassen, I.: New Feature Subset Selection Procedures for Classification of Expression Profiles. Genome Biology, 3(4): research0017.1-C0017.11, (2002)
11. Parzen, E.: On the Estimation of a Probability Density Function and Mode. Annals of Mathematical Statistics, **33** (1962)1064–1076
12. Vapnik, V.: The Nature of Statistical Learning Theory. Springer,New York (1995)