Gene Selection Based on Mutual Information for the Classification of Multi-class Cancer

Sheng-Bo Guo[1,2], Michael R. Lyu[3], and Tat-Ming Lok[4]

[1] Department of Automation, University of Science and Technology of China, Hefei, Anhui, 230026, China
sbguo@iim.ac.cn
[2] Intelligent Computation Lab, Hefei Institute of Intelligent Machines, Chinese Academy of Sciences, P.O. Box 1130, Hefei, Anhui, 230031, China
[3] Computer Science & Engineering Dept., The Chinese University of Hong Kong, Shatin, Hong Kong
[4] Information Engineering Dept., The Chinese University of Hong Kong, Shatin, Hong Kong

Abstract. With the development of mirocarray technology, microarray data are widely used in the diagnoses of cancer subtypes. However, people are still facing the complicated problem of accurate diagnosis of cancer subtypes. Building classifiers based on the selected key genes from microarray data is a promising approach for the development of microarray technology; yet the selection of non-redundant but relevant genes is complicated. The selected genes should be small enough to allow diagnosis even in regular laboratories and ideally identify genes involved in cancer-specific regulatory pathways. Instead of the traditional gene selection methods used for the classification of two categories of cancers, in the present paper, a novel gene selection algorithm based on mutual information is proposed for the classification of multi-class cancer using microarray data, and the selected key genes are fed into the classifier to classify the cancer subtypes. In our algorithm, mutual information is employed to select key genes related with class distinction. The application on the breast cancer data suggests that the present algorithm can identify the key genes to the BRCA1 mutations/BRCA2 mutations/the sporadic mutations class distinction since the result of our proposed algorithm is promising, because our method can perform the classification of the three types of breast cancer effectively and efficiently. And two more microarray datasets, leukemia and ovarian cancer data, are also employed to validate the performance of our method. The performances of these applications demonstrate the high quality of our method. Based on the present work, our method can be widely used to discriminate different cancer subtypes, which will contribute to the development of technology for the recovery of the cancer.

1 Introduction

Microarray technology, a recent development in experimental molecular biology, provides biomedical researchers the ability to measure expression levels of thousands of genes simultaneously. Such gene expression profiles are used to understand the

D.-S. Huang, K. Li, and G.W. Irwin (Eds.): ICIC 2006, LNBI 4115, pp. 454 – 463, 2006.

molecular variations among disease related cellular processed, and also to help the increasing development of diagnostic tools and classification platforms in the cancer research.

With the development of the microarray technology, the necessary processing and analysis methods grow increasingly critical. It becomes gradually urgent and challenging to explore the appropriate approaches because of the large scale of microarray data comprised of the large number of genes compared to the small number of samples in a specific experiment. For the data obtained in a typical experiment, only some of genes are useful to differentiate samples among different classes, but many other genes are irrelevant to the classification. Those irrelevant genes not only introduce some unnecessary noise to gene expression data analysis, but also increase the dimensionality of the gene expression matrix, which results in the increase of the computational complexity in various consequent researches such as classification and clustering. As a consequence, it is significant to eliminate those irrelevant genes and identify the informative genes, which is a feature selection problem crucial in gene expression data analysis [1, 2].

In the present paper, we propose a novel gene selection method based on the mutual information for the multi-class cancer classification using microarray data. Our method firstly calculates the mutual information (MI) between the discreted gene expression profiles and the cancer class label vector for all the samples. Then, the genes are ranked according to the calculated MI. These selected genes with high ranks are fed into the nearest neighbor method.

The rest of the present paper is organized as follows. In section 2, we first introduce the method to discretize the gene expression data, and then we in detail formulate the principle of GSMI. Section 3 describes the test statistics. And Section 4 describes the experiment. In Section 5, GSMI is applied to analyze the breast cancer dataset. Section 6 contains the conclusions.

2 Methods

Among the many thousands of genes simultaneously measured in a specific microarray experiment, it is impossible that all of their expressions are related to a particular partition of the samples. In the analysis of a biological system, the following 'rules of thumb' regarding gene functions are often assumed. 1) A gene can be in either the 'on' or 'off' state; 2) not all genes simultaneously respond to a single physiological event; 3) gene functions are highly redundant [3]. According to these assumptions, we consider the genes as random variables with two values, in which 1 denotes the 'on' state and 0 denotes the 'off' state. As a consequence, the gene expression data can be discretized into two states 0 and 1, respectively. The discretization of the gene expression data will be formulated later in Section 3.

Assume that a microarray dataset can be represented as a $G \times S$ matrix A with generic element a_{gs} representing the expression level of the gene, g in sample, s. All the samples are divided into n categories, and with the class label denoted by C with its element a_{gs} standing for the class of *ith* sample. From the biological point of view, those genes, having higher mutual agreement with class label of the cancer microarray

data, contribute more significantly on the classification of the cancer subtypes. Consequently, these genes should be selected as the key genes and used to the sequent classification and clustering. According to the information theory, mutual information can be used to measure the mutual agreement between two object models. We then employ the mutual information to rank every gene according to mutual information between the gene and the class label of the cancer microarray data.

Based on the information-theoretic principle of mutual information, the mutual information of two random variables X and $h_i = w_i / \sum w_i$ with a joint probability mass function $p(x, y)$ and marginal probability mass functions $p(x)$ and $p(y)$ is defined as [4]:

$$I(X;Y) = \sum_{x_i, y_j} p(x_i, y_j) \log \frac{p(x_i, y_j)}{p(x_i) p(y_j)} \ . \tag{1}$$

Let us suppose that the domain of $G_i, i \in \{1,...,G\}$, is distcretized into two intervals. After discretization, the domains of all the genes can be represented by $dom(G_i) = \{v_{ik}\}, k = 1, 2$ where $v_{i1=0}$ and $v_{i2=0}$. Denoted by σ the SELECT operation from relational algebra and $|S|$ denote the cardinality of set S [8]. The probability of a gene in microarray data having $G_i = v_{ik}, i \in \{1,...,G\}, k \in \{1,2\}$ is then given by:

$$P(G_i = v_{ik}) = \frac{\left| \sigma_{G_i = ik}(A) \right|}{\left| \sigma_{G_i \neq \Phi}(A) \right|} \ . \tag{2}$$

And the joint probability of the gene in the gene expression data has $G_i = v_{ik}$ and the class label $C = c_i, i \in \{1,...,n\}$ is calculated by:

$$P(G_i = v_{ik} \wedge C = c_i) = \frac{\left| \sigma_{G_i = v_{ik} \wedge C = c_i}(A) \right|}{\left| \sigma_{G_i \neq NULL}(A) \right|} \ . \tag{3}$$

Definition 1. The interdependence measure I between the gene and the class label, G_i and $C, i \in \{1,...,G\}$, is defined as:

$$I(G_i : C) = \sum_{k=1}^{2} \sum_{l=1}^{n} P(G_i = v_{ik} \wedge C = c_i) \log \frac{P(G_i = v_{ik} \wedge C = c_i)}{P(G_i = v_{ik}) P(C = c_i)} \ . \tag{4}$$

$I(G_i : C)$ measures the average reduction in uncertainty about G_i that results from learning the value of C [9]. If $I(G_i : C) > I(G_j : C), i, j \in \{1,...,G\}, i \neq j$, the dependence of G_i and the class label C is greater than the dependence of G_j and C. Before ranking the genes according to the mutual information, the redundancy in the microarray should be decreased because of the fundamental principle of microarray technology. Due to the principle of micorarray technology, the gene expression matrix contains high redundancy since some genes are measured more than once.

Definition 2. The mutual information matrix of the microarray, named as M with its element m_{ij}, is given by:

$$m_{ij} = \sum_{k=1}^{2}\sum_{l=1}^{2} P(G_i = v_{ik} \wedge G_j = v_{jl}) \log \frac{P(G_i = v_{ik} \wedge G_j = v_{jl})}{P(G_i = v_{ik})P(G_j = v_{jl})} . \tag{5}$$

For simplicity, the mutual information matrix M is normalized to M^* with its element m^* given as follows,

$$m_{ij}^* = \frac{m_{ij}}{H(G_i, G_j)} \tag{6}$$

where the joint entropy of the gene G_i and G_j is denoted by $H(G_i, G_j)$, which is given by:

$$H(G_i, G_j) = -\sum_{k=1}^{2}\sum_{l=1}^{2} P(G_i = v_{ik} \wedge G_j = v_{jl}) \log P(G_i = v_{ik} \wedge G_j = v_{jl}) . \tag{7}$$

The redundancy in the microarray data is reduced by the following method. In the matrix M^*, the elements on the diagonal are all with the same value 1. These rows without containing the value less than 0.95 are labeled in the normalized mutual information matrix except these elements in the diagonal of the matrix. We then select these genes corresponding to these rows. Due to the error in the process of computing the mutual information, the cutoff value is set to 0.95 so that the redundancy can be reduced as much as possible. Otherwise, if the cutoff value is set to 1, the redundancy cannot be reduced to the expected extent.

After selecting the genes with little redundancy, if any, the selected genes (SGS) are ranked according to the interdependence measure I between the gene expression profiles and the class label. Then the SGS are used to train the RBF neural network to classify the cancer subtypes in the designed experiment formulated in the next section.

3 Test Statistics

A general statistical model for gene expression values will be firstly introduced followed by several test statistics in this section.

Assume that there are more than two kinds of distinct tumor tissue classes for the problem under consideration and there are p genes (variables) and n tumor mRNA samples (observations). After introducing the novel gene selection method, we now turn to some test statistics used for testing the equality of the class means for a fixed gene. The following five parametric test statistics will be considered [10].

3.1 ANOVA F Test Statistics

The definition of this test is given by:

$$F = \frac{(n-k)\sum_i n_i(\overline{Y}_{i\cdot} - \overline{Y}_{\cdot\cdot})^2}{(k-1)\sum_i (n_i - 1)s_i^2} \quad . \tag{8}$$

where $\overline{Y}_{\cdot} = \sum_{j=1}^{n_i} Y_{ij} / n_i, \overline{Y}_{\cdot\cdot} = \sum_{i=1}^{k} n_i \overline{Y}_{i\cdot} / n_i$, and $s_i^2 = \sum_{j=1}^{n_i} (Y_{ij} - \overline{Y}_{i\cdot})^2 /(n_i - 1)$.

3.2 Brown-Forsythe Test Statistic

Brown-Forsythe Test Statistic [11] is given by:

$$B = \frac{\sum_i n_i(\overline{Y}_{i\cdot} - \overline{Y}_{\cdot\cdot})^2}{\sum_i (1 - n_i / n)s_i^2} \quad . \tag{9}$$

3.3 Welch Test Statistics

Welch test statistics [12] is defined as

$$W = \frac{\sum_i w_i(\overline{Y}_{i\cdot} - \sum_i h_i \overline{Y}_{i\cdot})^2}{(k-1) + 2(k-2)(k+1)^{-1}\sum_i (n_i - 1)^{-1}(1 - h_i)^2} \tag{10}$$

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

4 Introduction to the Experiment

To evaluate the performance of GSMI, we applied it to the well-known gene expression data sets: the breast cancer data [5], in which RNA from samples of primary breast tumors from 7 carriers of the BRCA1 mutation, 8 carries of the BRCA2 mutation, and 7 patients with sporadic cases of breast caner have been hybridized to a cDNA microarray containing 6512 complementary DNA clones of 5361 genes [6]; Leukemia 72 with 6817 genes, 38 ALL-Bcell, 9 ALL-Tcell, and 25 AML, and Ovarian with 7129 genes, 27 epithelial ovarian cancer cases, 5 normal tissues, and 4 malignant epithelial ovarian cell lines [13].

Before calculated the mutual information, the microarray expression level should firstly be preprocessed according to an alternative idea of the Optimal Class-Dependence Discretization Algorithm (OCDD) [11]. OCDD is a new method to convert variables into discrete variables for inductive machine learning, which can thus be employed for pattern classification problems. The discretization process is formulated as an optimization problem, then the normalized mutual information that measures the interdependence between the class labels and the variable to discretized as the objective function, and then iterative dynamic programming is applied to find its optimum [14]. For each continuous gene expression profile in microarray expression matrix A , its domain is typically discretized into two intervals for gene selection, which are denoted by 0 and 1, respectively. We then use the normalized mutual information measure that reflects interdependence between the class label and the attribute to be discretized as the objective function to find a global optimal solution separating the domain of the gene expression data.

We employ the nearest neighbor method to classify the cancer cases with different cancer subtypes. The leave-one-out cross-validation (LOOCV) is used to evaluate the accuracy of classification.

5 Experiment Results and Discussion

By using our method, genes are ranked by the mutual information between the genes and the class label. Then, the nearest neighbor classifier is employed as the benchmark to classify the three cancer microarray expression datasets.

In the classification performance evaluation process, we employed LOOCV, which is a widely used method for evaluating the performance of the classification of gene expression data [7].

The results of our method on the three datasets are given in Figure 1, Figure 2 and Figure 3, respectively. From figure 1, the classification error rate minimized to 0% when 18 genes are selected according to our method, but the genes selected by all the test statistics used for classification are not as effective as ours since the classification accuracies maximize to 73% at 404 genes for the ANOVA test statistic, 78% at 119 genes for Brown-Forsythe test statistic, 73% at 116 genes for Welch test statistic. From figure 2, the classification error rate minimized to 1.39% when 19 genes are

selected according to our method, but the genes selected by all the test statistics used for classification are not as effective as ours since the classification accuracies maximize to 85% at 70 genes for the Brown-Forsythe test statistic, 71% at 156 genes for ANOVA test statistic, 80.5% at 354 genes for Welch test statistic. From figure 3, the classification error rate minimized to 0 % when 21 genes are selected according to our method, but the genes selected by all the test statistics used for classification are not as effective as ours since the classification accuracies maximize to 97.3% at 38 genes for the Welch test statistic, 97.3% at 370 genes for ANOVA test statistic, 94.45% at 105 genes for Brown-Forsythe test statistic.

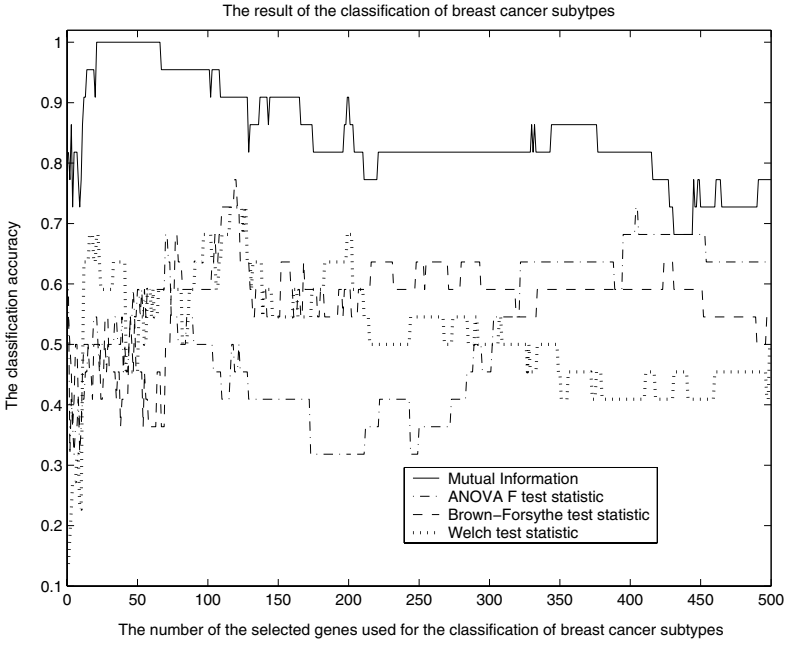

Fig. 1. Comparison on the breast cancer cases

Demonstrated by these results, the present method based on the mutual information obviously outperforms the three test statistics. The major reason for the superiority of our method is that mutual information between the genes and the class labels reflects the potential relation and correlation, and thus indicates the discriminability of genes. What's more, there is no assumption about the probability distribution of the microarray data. The three test statistics are based on the default probability distribution, but it is not clear whether microarray data are according to the default probability till now without adequate samples of the cancer cases.

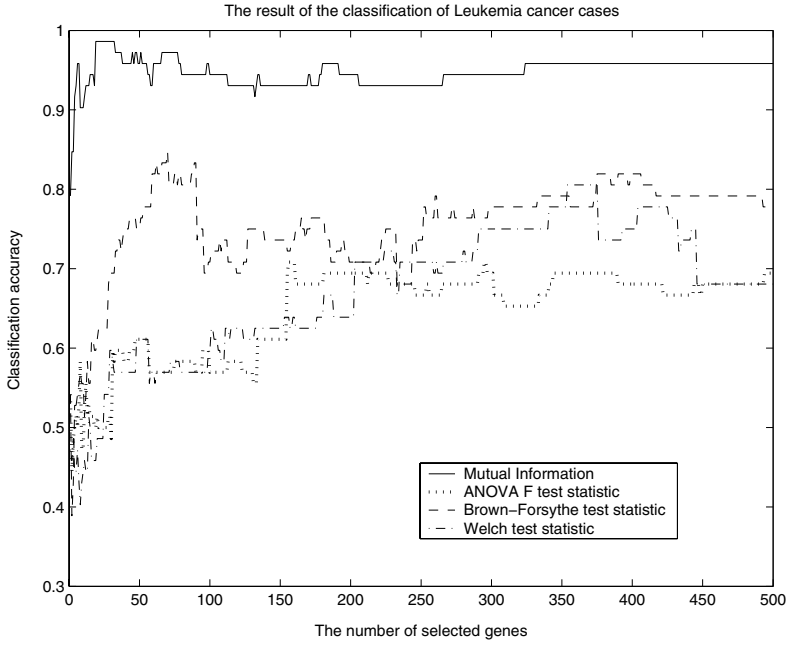

Fig. 2. Comparison on the leukemia cancer cases

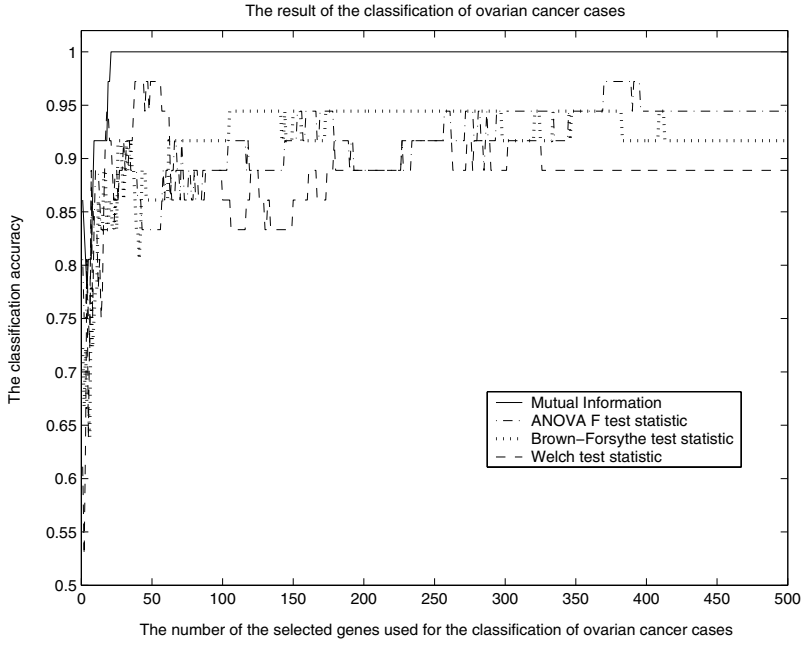

Fig. 3. Comparison on the ovarian cancer cases

6 Conclusions

In this paper, contrary to other work managing discriminating one cancer subtype from the other, we present the novel method for the classification of the multi-class cancer subtypes, which will contribute the development of the technology for the research and cure of cancer. In our work, an information theoretic approach is proposed for gene selection algorithm based on the mutual information, which proves promising in the classification of multi-class cancer microarray datasets. The mutual information between the gene expression data and the class label is calculated. And the genes are selected according to the calculated mutual information after removing the redundancy. The successful applications on the breast, as well as ovarian and leukemia cancer datasets prove that our algorithm is effective, robust and appropriate to the classification of the multi-class cancers since it can discovery the informative key genes.

Future work will concentrate on the further research on the method for extracting the key features from gene expression data. And we will try to fully evaluate the performance of the classification by employing SVM and neural network, such as RBF neural network. Furthermore, we will also concentrate on the research of biological significance of the found key genes and try to find the specific feature of the key genes and understand the function of these genes. By doing so, the microarray technology can be fully used.

Acknowledgement

The authors are grateful to Dechang Chen for sharing the micorarray data sets of breast, ovarian as well as leukemia with us.

References

1. Ben-Dor, A.: Tissue Classification with Gene Expression Profiles, Journal of Computational Biology, 7 (2000) 559-583
2. Weston, J., Mukherjee, S., Chapelle, O., Pontil, M., Poggio, T., Vapnik, V.: Feature Selection for SVMs. In Advances in Neural Information Processing Systems, MIT Press, 13 (2001)
3. Xing, E. P., Richard, M. K.: CLIFF: Clustering of High-Dimensional Microarray Data via Iterative Feature Filtering Using Normalized Cuts. Bioinformatics, 17 (1) (2001) 306-315
4. Cover, T., Thomas J.: Elements of Information Theory. John Wiley and Sons, Inc (1991)
5. Hedenfalk, I., Duggan, D., Chen, Y., Radmacher, M., Bittner, M., Simon, R., Meltzer, P., Gusterson, B., Esteller, M., Raffeld, M.: Gene-Expression Profiles in Hereditary Breast Cancer. New Eng. J. Med. 344 (2001) 539-548
6. Nathalie, P., Frank, D. S., Johan, A. K. S., Bart, L. R. D. M.: Systematic Benchmarking of Microarray Data Classification: Assessing the Role of Non-linearity and Dimensionality Reduction. Bioinformatics, 20 (17) (2004) 3185-3195
7. Simon, R.: Supervised Analysis when The Number of Candidate Features Greatly Exceeds the Number of the Cases. SIGKDD Explorations, 5 (2) (2003) 31-36

8. Au, W. H., Keith, C.C. C., Andrew, K.C. W., Wang, Y.: Attribute Clustering for Grouping, Selection and Classification of Gene Expression Data. IEEE/ACM Transactions on computational biology and bioinformatics, April-June, 2 (2) (2005) 83-101
9. MacKay, D. J. C.: Information Theory, Inference, and Learning Algorithm. Cambridge Univ. Press (2003)
10. Chen, D. C., Liu, Z. Q., Ma, X. B., Hua, D.: Selecting Genes by Test Statistics. Journal of Biomedicine and Biotechnology, 2 (2005) 132-138
11. Brown, M. B., Forsythe, A. B.: The Small Sample Behavior of Some Statistic which Test the Equality of Several Means. Technometrics (1974) 129-132
12. Welch, B. L.: On the Comparison of Several Mean Values: An Alternative Approach. Biometrika., 38 (1951) 330-336
13. Chen, D. C., Hua, D., Jaques, R., Cheng, X. Z.: Gene Selection for Multi-class Prediction of Microarray Data. Bioinformatics Conference, 2003, CSB'03, Proceedings of the 2003 IEEE, (2003) 492-495
14. Liu, L., Andrew, K.C. W., Wang, Y.: A Global Optimal Algorithm for Class-Dependent Discretization of Continuous Data, 8 (2) (2004) 151-170