Journal of Biological Systems, Vol. 12, No. 3 (2004) 371–386

World Scientific
www.worldscientific.com

NONLINEAR PROBIT GENE CLASSIFICATION USING MUTUAL INFORMATION AND WAVELET-BASED FEATURE SELECTION

XIAOBO ZHOU
Department of Electrical Engineering, Texas A&M University
College Station, TX 77843, USA

XIAODONG WANG
Department of Electrical Engineering, Columbia University
New York, NY 10027, USA

EDWARD R. DOUGHERTY*
Department of Electrical Engineering, Texas A&M University
and
Department of Pathology, University of Texas M.D. Anderson Cancer Center
Houston, TX 77030, USA
edward@ee.tamu.edu

Received 7 June 2003
Revised 29 December 2003

We consider the problem of cancer classification from gene expression data. We propose using a mutual information-based gene or feature selection method where features are wavelet-based. The bootstrap technique is employed to obtain an accurate estimate of the mutual information. We then develop a nonlinear probit Bayesian classifier consisting of a linear term plus a nonlinear term, the parameters of which are estimated using the Gibbs sampler. These new methods are applied to analyze breast-cancer data and leukemia data. The results indicate that the proposed gene and feature selection method is very accurate in breast-cancer and leukemia classifications.

Keywords: Wavelet transformation; gene microarray; gene selection; mutual information; nonlinear probit classifier; Gibbs sampler.

1. Introduction

cDNA microarrays make it possible to measure simultaneously the expression levels for thousands of genes. In addition to their enormous scientific potential to help understand gene regulation and interactions [12, 14, 24, 29], DNA microarrays have very important applications in pharmaceutical and clinical research. By comparing gene expressions in normal and diseased cells, we may use microarrays to identify disease related genes and targets for therapeutic drugs. Therefore, the huge amount

*Corresponding author.

of data provided by cDNA microarray measurement is explored in order to answer fundamental questions about gene functions and their inter-dependence, and hopefully to provide answers to questions like what type of disease affects a cell or which genes have strong influence on this disease.

Since there exist thousands of genes and only tens of samples in microarray data, feature reduction or gene selection is the necessary step for cancer classification. There are mainly two approaches: one is feature reduction. For example, transformation methods such as principal component analysis (PCA) have been applied in cancer classification [11, 20]. Another is gene selection [9]. In this paper, we consider both.

Singular value decomposition has been applied to represent genome-wide expression data [2]. The wavelet transform is a very powerful tool and finds applications in many areas including bioinformatics [18]. In this paper, we propose to use the wavelet transform to represent microarray data. Each feature (wavelet coefficient) of the wavelet transformation is related to several genes of the original gene microarray data. We then adopt a mutual information-based feature selection method to select the strongest features from the wavelet coefficients. The recognition accuracy of this new proposed method is very high, as seen in Sec. 5.

Gene selection is often a very important step in cancer classification and the discovery of gene pathways. A number of variable (or gene) selection methods have been proposed, e.g., the support vector machine method [10], the genetic algorithm [17], the perceptron method [13], Bayesian variable selection [16, 22, 28, 30], and the voting technique [9]. In this paper, we develop a mutual information-based gene and feature selection method. To cope with the small sample size, the bootstrap technique [32] is employed to obtain more accurate estimation of the mutual information. Although gene selection by using entropy and Kulback–Leibler divergence is discussed in [27], that work is based on estimating the distribution of many genes, which is not feasible for classification because the sample size is very small.

A linear probit regression classifier was proposed in [1], which is very effective in cancer classification [16]. The strongest genes and features selected by the mutual information criterion actually show a strong nonlinear property because this selection method is a nonlinear procedure. Hence the recognition accuracy is not high when the linear probit regression classifier is applied based on these strongest features. Here we propose a nonlinear probit regression classifier consisting of a linear term plus a nonlinear term, and the parameters are estimated using the Gibbs sampler. The nonlinear probit classifier then outperforms the linear probit classifier.

An outline of the proposed classifier with the gene and feature selection algorithm is as follows:

- Feature extraction from wavelet representation of microarray: Assume there are M microarrays and N genes with the microarray data being $\xi_{ij}, i = 1, \ldots, M$, $j = 1, \ldots, N$. The wavelet features obtained from a wavelet transformation of each microarray are denoted by $x_{ij}, j = 1, \ldots, N$ for $i = 1, \ldots, M$. See Appendix 1.

- Mutual information estimation: The features (or genes) are denoted by x_{ij}, $i = 1, \ldots, M, j = 1, \ldots, N$. We estimate the class-feature (or class-gene) mutual information using a bootstrap technique. See Appendix 2.
- Gene and feature selection: After obtaining the mutual information $\{\hat{I}_j\}_{j=1}^{N}$, we select the P strongest features (or genes) based on mutual information maximization. The profiles of the strongest features (or genes) are denoted by $y_{ij}, i = 1, \ldots, M, j = 1, \ldots, P$. See Appendix 2.
- Nonlinear probit classifier: Given the class label $w_i(0$ or $1)$, and the corresponding features (or genes) y_{ij} (e.g., $i = 1, \ldots, M$ and $j = 1, \ldots, P$), we construct a nonlinear probit regression model and estimate the parameters in this model using the Gibbs sampler. Then, based on the estimated model, we compute $P(w_i = 1|\boldsymbol{y}_i)$ with $\boldsymbol{y}_i = \{y_{i1}, \ldots, y_{iP}\}$ and decide whether w_i is 0 (class 0) or 1 (class 1). See Appendix 3.

We next provide some experimental results. The mathematical details are presented in the sections of the appendix: Appendix 1 gives the wavelet representation for gene microarrays, Appendix 2 describes the gene and feature selection approach using mutual information, along with the bootstrap technique for estimating the mutual information, and Appendix 3 describes the nonlinear probit classifier.

2. Experimental Results

In order to speed up the computation of mutual information, as the first step of our proposed algorithms, we pre-select genes based on the following criterion: the smaller is the sum of squares within groups and the bigger is the sum of squares between groups, the better is the classification accuracy. Therefore we can define a score using the above two statistics to pre-select genes, i.e., the ratio of the between-group to within-group sum of squares. It is actually the F-test. Denote $R(j)$ as the score for gene j $(1 \leq j \leq Q)$, where Q is the total number of original genes. We select a threshold Γ and keep those genes j such that $R(j) \geq \Gamma$. The pre-selection procedure yields N genes such that $R(j) \geq \Gamma$.

2.1. *Experimental results for breast cancer data*

We first focus on hereditary breast cancer data, which can be downloaded from the web page for the original paper [11]. In [11], cDNA microarrays are used in conjunction with classification algorithms to show the feasibility of using differences in global gene expression profiles to separate BRCA1 and BRCA2 mutation-positive breast cancers. Twenty-two breast tumor samples from 21 patients were examined: 7 BRCA1, 8 BRCA2, and 7 sporadic. There are 3226 genes for each tumor sample. We use our methods to classify BRCA1 versus the others (BRCA2 and sporadic). The cross-validation (leave-one-out) method is employed to compute all classification errors.

2.1.1. *Mutual information-based gene selection*

Table 1 describes the strongest genes selected from the hereditary breast cancer data using the mutual information-based gene selection method. (For reading convenience, instead of Clone ID, we use the gene index number in the data base [11].) Note that the gene keratin 8 is one of the strongest genes. This is consistent with the discussion in other references [13, 16]. It is a member of the cytokeratin family of genes. Cytokeratins are frequently used to identify breast cancer metastases by immunohistochemistry. Many of the strongest genes (denoted by the stars in Table 1) are same as the strongest genes selected by a Bayesian method [16].

We test recognition accuracy using two models: models 1 and 2 consist of the first five and first ten genes, respectively, in Table 1. The nonlinear Bayesian probit classifier (NLProbit) has no errors in both models, whereas the linear Bayesian probit classifier (LProbit) [16] has two errors in each model. In practice we would prefer the 5-gene model on account of the small sample size and greater ease in recognizing biological relations.

Looking at the probabilities for model 1 using the linear probit and the nonlinear probit classification methods in Table 2, we see that the nonlinear probit classifier corrects the two errors (sample 16 and sample 17) caused by the linear probit classifier. Moreover, for samples 6 and 12, the results of the LProbit classifier appear weak because the probabilities are very close to 0.5, whereas the NLProbit classifier shows very good separability for all samples. As foreshadowed in the Introduction, we conjecture that the nonlinear approach works better because the strongest genes found by mutual information exhibit nonlinear properties.

Table 1. Strongest genes selected by mutual information.

No.	Mutual information	Index No.	Gene description
1	1.6165	556*	tumor protein p53-binding protein, 2
2	1.6018	2670	interleukin enhancer binding factor 2, 45kD
3	1.4723	1008*	keratin 8
4	1.3969	2893	mutS (*E. coli*) homolog 2 (colon cancer, nonpolyposis type 1)
5	1.3890	1065*	signal transducing adaptor molecule 1
6	1.3889	1999*	ESTs
7	1.3858	1345	Major histocompatibility complex, class II
8	1.3837	1859*	ESTs
9	1.3719	1443*	chromobox homolog 3 (Drosophila HP1 gamma)
10	1.3527	2734*	minichromosome maintenance deficient (S. cerevisiae) 7
11	1.3518	3009*	ESTs, Highly similar to titin-like protein [H. sapiens]
12	1.3231	1466	polymerase (RNA) II (DNA directed) polypeptide F
13	1.3037	609	CHK1 (checkpoint, S. pombe) homolog
14	1.3034	806	CTP synthase
15	1.2987	272*	splicing factor, arginine/serine-rich 10
16	1.2915	2951	cyclin-dependent kinase inhibitor 2C (p18, inhibits CDK4)
17	1.2820	963	splicing factor, arginine/serine-rich 4
18	1.2730	2272	ESTs
19	1.2670	2423	very low density lipoprotein receptor
20	1.2511	1179*	KIAA0090 protein

Table 2. The probabilities for each sample in model 1.

Sample index No.	w	LProbit $\mathrm{P}(w = 1)$	NLProbit $\mathrm{P}(w = 1)$
1	0	0.1862	0.3781
2	0	0.3009	0.1438
3	0	0.1584	0.0559
4	0	0.0375	0.3105
5	0	0.0130	0.1011
6	0	0.4513	0.0879
7	1	0.8219	0.8983
8	1	0.9758	0.9701
9	1	0.9270	0.9813
10	1	1.0000	0.9990
11	1	1.0000	1.0000
12	1	0.5089	0.9839
13	1	0.9930	0.9996
14	1	0.8150	0.9060
15	1	0.9992	0.9826
16	1	0.4715	0.7774
17	1	0.3046	0.7906
18	0	0.0523	0.0027
19	1	1.0000	0.9746
20	1	1.0000	0.9998
21	1	0.9326	0.9817
22	1	0.9898	0.9879
No. of misclassification		2	0

2.1.2. *Mutual information-based feature selection for wavelet representation*

We select the Daubechies' bases 4 which has four non-zero coefficients $\{h_n\}_{n=0}^{3}$ [6] of the compact support wavelet orthogonal basis. We set $\Gamma = 0.0452$ as the threshold for the pre-selection procedure, which yields 173 genes, and then do wavelet decomposition using the gene-expression profiles of these genes. We treat all coefficients of $\{\boldsymbol{x}_i\}_{i=1}^{M}$ in (A.1.7) as features. They include the approximation coefficients and wavelet coefficients. Using different wavelet bases in different scales reveals similar phenomena. Since many wavelet coefficients are very small, the result shows that only some large coefficients are useful for representing the original gene-expression profiles. That is in agreement with the description of the last paragraph in Sec. 2. Two models are used to test the recognition accuracy: model 1 uses the five strongest features and model 2 uses the ten strongest features. For both models, the linear probit classifier has two errors, whereas no error is found with the nonlinear probit classifier.

2.2. *Experimental results for leukemia data set*

We have also applied the proposed methods to the leukemia data of [9], which is publicly available at http://www-genome.wi.mit.edu/cgi-bin/cancer/publications. The microarray data contains 7129 human genes, sampled from 72 cases of cancer, of which 38 are of type B-cell ALL, 9 are of type T-cell ALL and 25 of type AML.

The data are preprocessed as recommended in [25]: gene values are truncated from below at 100 and from above at 16,000; genes having the ratio of the maximum over the minimum less than five or the difference between the maximum and the minimum less than 500 are excluded; and finally the base-10 logarithm is applied to the 3571 remaining genes.

Following the experimental setup in [9], the data is split into a training set consisting of 38 samples of which 27 are ALL and 11 are AML, and a test set of 34 samples, 20 ALL and 14 AML. In [9], a classifier is trained using a weighted voting scheme on the training samples, and correctly classifies 29 of the 34 samples. Table 3 lists the ten strongest genes from the mutual information method. The index number is the Clone ID. Some strongest genes, such as index numbers 2288, 1882, 4847, belong to the set of the genes used in [9]. Using the five strongest genes, the LProbit linear classifier has two misclassifications and there is one error for the NLProbit classifier. Moreover, for the five strongest features consisting of the wavelet basis representation for this data set, one error is found for the NLProbit classifier, but the LProbit linear classifier has four errors. Note that the authors also got very good results using this data set in [16].

Finally we also compare the top ten genes listed in Table 3 with the top ten genes selected using a t-test score [26], which are listed in Table 4. Although the two

Table 3. Strongest genes selected by mutual information.

No.	MI	Index No.	Gene description
1	1.7971	5376	Cyclooxygenase-2 (hCox-2) gene
2	1.7681	3640	Zinc finger transcription factor hEZF (EZF) mRNA
3	1.6920	6219	GB DEF = Neutrophil elastase gene, exon 5
4	1.6890	2288*	DF D component of complement (adipsin)
5	1.6632	5599	Cyclooxygenase-2 (hCox-2) gene
6	1.6632	5598	PTGS2 Prostaglandin-endoperoxide synthase 2
7	1.6632	1926	PTX3 Pentaxin-related gene, rapidly induced by IL-1 beta
8	1.6582	1604	RB1 Retinoblastoma 1 (including osteosarcoma)
9	1.6578	1882*	CST3 Cystatin C (amyloid angiopathy and cerebral hemorrhage)
10	1.6151	4847*	Zyxin

Table 4. Strongest genes selected by t-test score.

No.	Index No.	Gene description
1	1882	CST3 Cystatin C (amyloid angiopathy and cerebral hemorrhage)
2	760	CYSTATIN A
3	1834	CD33 CD33 antigen (differentiation antigen)
4	6218	ELA2 Elastatse 2, neutrophil
5	3320	Leukotriene C4 synthase (LTC4S) gene
6	2288	DF D component of complement (adipsin)
7	4847	Zyxin
8	5772	C-myb gene extracted from Human (c-myb) gene
9	6200	Interleukin 8 (IL8) gene
10	4499	CHRNA7 Cholinergic receptor, nicotinic, alpha polypeptide 7

gene selection criteria are different, it is interesting to notice that the gene index numbers 2288, 1882 and 4847 appear in both Tables 3 and 4. In [16], the three genes are also listed as important genes using Bayesian gene selection (note that there the initial important genes are selected using a t-test).

3. Conclusions

In this paper, we have treated the problem of cancer classification based on microarray expression data. First, like the PCA transformation and the singular value decomposition for genome-wide expression data, we have proposed a wavelet representation of the gene microarray data. We have also proposed a mutual information-based gene and feature selection method. The bootstrap technique is employed to obtain an accurate estimate of the mutual information. Furthermore, we have developed a nonlinear probit Bayesian classifier consisting of a linear term plus a nonlinear term, the parameters of which are estimated using the Gibbs sampler. These methods are applied to analyze both breast cancer data and leukemia data. The results show that the proposed gene selection method is very efficient in breast cancer and leukemia classifications. The classification error is very low in our experiments on the two data sets using our proposed methods. Moreover, the proposed nonlinear probit classifier outperforms the linear probit classifier using the mutual-information selected features.

Appendix 1. Wavelet Representation of Microarray Data

Suppose a discrete signal $\boldsymbol{c}^0$ has a finite length K_0, i.e., $\boldsymbol{c}^0 = [c_1^0, \ldots, c_{K_0}^0]^T$. In Mallat's algorithm [19] of multi-resolution analysis, a one-stage FIR filter, say low-pass filter, and down-sampling by a factor of two can be given in a matrix form [6],

$$\underbrace{\begin{bmatrix} c_0^1 \\ c_2^1 \\ c_4^1 \\ \vdots \\ \vdots \\ c_{K_1-3}^1 \\ c_{K_1-1}^1 \end{bmatrix}}_{\triangleq \boldsymbol{c}^1{}_{\downarrow 2}} = \underbrace{\begin{bmatrix} h_{2N-2} & h_{2N-1} & & & & & h_0 & \cdots & h_{2N-3} \\ & & & & \ddots & & & & \\ h_0 & h_1 & h_2 & \dots & h_{2N-1} & & & & \\ & & h_0 & h_1 & \dots & h_{2N-1} & & & \\ & & & & & & & & \\ & & & & \ddots & & & & \\ h_2 & h_3 & \cdots & h_{2N-1} & & & & h_0 & h_1 \end{bmatrix}_{\frac{K_1}{2} \times K_0}}_{\triangleq \boldsymbol{H}^0} \underbrace{\begin{bmatrix} c_0^0 \\ c_1^0 \\ c_2^0 \\ \vdots \\ \vdots \\ c_{K_0-2}^0 \\ c_{K_0-1}^0 \end{bmatrix}}_{\triangleq \boldsymbol{c}^0}, \tag{A.1.1}$$

where 2κ is the length of the low-pass filter $\mathcal{H}$ and the length of $\boldsymbol{c}^1$ is $K_1 \overset{\triangle}{=} K_0 + 2\kappa - 1$. The impulse response of the low-pass filter is given by $h_n = \frac{1}{\sqrt{2}} \int \phi(\frac{x}{2}) \times \phi(x - n)dx$, where ϕ is the generator of the multi-resolution wavelet basis. Specifically, in $\boldsymbol{H}^0$, we adopt a commonly used periodic method to mitigate edge effects [6]. Similarly, the detailed signal of $\boldsymbol{c}^0$ after down-sampling is given by

$$\boldsymbol{d}^1 \downarrow 2 = \boldsymbol{G}^0 \boldsymbol{c}^0, \tag{A.1.2}$$

where $\boldsymbol{G}^0$ is similarly defined as $\boldsymbol{H}^0$ in (A.1.1) with $\{h_n\}$ replaced by $\{g_n\}$, which is in general given by $g_n = (-1)^n h_{1-n}$. Here we can write the one-step wavelet decomposition of $\boldsymbol{c}^0$ as

$$\boldsymbol{x}^1 \overset{\triangle}{=} \begin{bmatrix} \boldsymbol{c}^1 \downarrow 2 \\ \boldsymbol{d}^1 \downarrow 2 \end{bmatrix}_{K_1 \times 1} = \begin{bmatrix} \boldsymbol{H}^0 \\ \boldsymbol{G}^0 \end{bmatrix}_{K_1 \times K_0} \boldsymbol{c}^0, \tag{A.1.3}$$

where $\boldsymbol{x}^1$ contains the first-level wavelet coefficients. Repeating the above decomposition L times and keeping all the detailed signals and the last level signal approximation, we then obtain the Lth level wavelet coefficients $\boldsymbol{x}^L$ given by

$$\begin{aligned} \boldsymbol{x}^L &\overset{\triangle}{=} \begin{bmatrix} \boldsymbol{c}^L \downarrow 2 \\ \boldsymbol{d}^L \downarrow 2 \\ \vdots \\ \boldsymbol{d}^1 \downarrow 2 \end{bmatrix}_{K \times 1} \\ &= \underbrace{\begin{bmatrix} \begin{bmatrix} \boldsymbol{H}^{L-1} \\ \boldsymbol{G}^{L-1} \end{bmatrix}_{K_L \times \frac{K_{L-1}}{2}} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{I}_{\sum_{i=0}^{L-2} \frac{K_i}{2}} \end{bmatrix} \cdots \begin{bmatrix} \begin{bmatrix} \boldsymbol{H}^{1} \\ \boldsymbol{G}^{1} \end{bmatrix}_{K_2 \times \frac{K_1}{2}} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{I}_{\frac{K_1 + K_0}{2}} \end{bmatrix} \begin{bmatrix} \boldsymbol{H}^0 \\ \\ \boldsymbol{G}^0 \end{bmatrix}}_{\overset{\triangle}{=} \mathcal{W}} \boldsymbol{c}^0. \end{aligned} \tag{A.1.4}$$

Since $K_i = K_{i-1} + 2\kappa - 1, i = 1, \ldots, L$, the length K of the Lth level wavelet coefficients is

$$K = \frac{L+2}{2} K_0 + \frac{L^2 + 3L}{4} (2\kappa - 1). \tag{A.1.5}$$

It then follows from (A.1.4) that $\boldsymbol{c}^0 = \boldsymbol{\Phi} \boldsymbol{x}^L$, where the perfect reconstruction matrix is given by

$$\begin{aligned} \boldsymbol{\Phi} &\overset{\triangle}{=} (\mathcal{W}^T \mathcal{W})^{-1} \boldsymbol{W}^T \\ &= \begin{bmatrix} \boldsymbol{H}^{0T} & \boldsymbol{G}^{0T} \end{bmatrix} \begin{bmatrix} \boldsymbol{H}^{1T} & \boldsymbol{G}^{1T} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{I}_{\frac{K_1}{2}} \end{bmatrix} \cdots \begin{bmatrix} \boldsymbol{H}^{(L-1)T} & \boldsymbol{G}^{(L-1)T} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{I}_{\sum_{i=0}^{L-1} \frac{K_i}{2}} \end{bmatrix}. \end{aligned} \tag{A.1.6}$$

For the ith microarray, $\boldsymbol{\xi}_i = [\xi_{i,1}, \ldots, \xi_{i,K_0}]^T$, applying the wavelet decomposition (A.1.4) to $\boldsymbol{\xi}$ yields

$$\boldsymbol{x}_i = \boldsymbol{\mathcal{W}}\boldsymbol{\xi}_i, \quad i = 1, \ldots, M. \tag{A.1.7}$$

As noted in (A.1.5), the number of the wavelet coefficients K is a function of the length of the original signal K_0, the length of the low-pass (or high-pass) filter κ, and the number of levels of the decomposition L. Note that the wavelet coefficient space is structured, roughly, according to the location and scale of the functional information contained each coefficient. Only a few large coefficients explain most of the functional form in the signal, while the remaining majority are small and therefore can be discarded. Various wavelet shrinkage methods exist for choosing the number of wavelet coefficients [7]. Since different microarrays usually have different noise levels, it is difficult to determine a threshold for each microarray. In this paper, our objective is to extract important features from the wavelet coefficients for cancer classification, hence after obtaining the wavelet coefficients $\{\boldsymbol{x}_i\}$ using the wavelet transformation, we will apply a mutual information-based feature selection method developed in the next section to select some important features.

Appendix 2. Mutual Information-Based Gene Selection

Mutual information

The motivation for considering mutual information is its capacity to measure a general dependence among random variables. Shannon's information theory provides a suitable formalism for quantifying such a concept. Denote X as the random variable describing the jth gene-expression and C as the random variable describing the classes. Assume N_c classes to be studied. If the probabilities for the different classes are $P(c)$; $c = 1, \ldots, N_c$, the initial uncertainty in the output class is measured by the entropy

$$H(C) = -\sum_{c=1}^{N_c} P(c) \log P(c). \tag{A.2.1}$$

The *entropy* of a gene expression pattern is a measure of the uncertainty information content in that pattern. Given a continuous random vector X and its probability distribution $p(x)$, the *entropy* is defined as

$$H(X) \triangleq -\int p(x) \log p(x) dx. \tag{A.2.2}$$

Higher entropy for gene variables means that their expression levels are more randomly distributed. Given the joint distribution $p(c, x)$, the *joint entropy* of X and C is a measure of the uncertainty information between X and C, and is defined by

$$H(C; X) \triangleq -\sum_{c=1}^{N_c} \int p(c, x) \log p(c, x) dx. \tag{A.2.3}$$

When certain variables are known and others are not, the uncertainty is measured by the *conditional entropy*

$$\begin{aligned}H(C|X) &\triangleq -\sum_{c=1}^{N_c}\int p(c,x)\log p(c|x)dx.\\ &= -\sum_{c=1}^{N_c}\int p(c,x)\log\frac{p_c(x)\,P(c)}{p(x)}dx,\end{aligned} \tag{A.2.4}$$

where $p_c(x)$ is the conditional distribution given class c. In general, the conditional entropy will be less than or equal to the initial entropy. It is equal if and only if one has independence between features and output class. The mutual information $I(C;X)$ is a measure of the information given for C by observing X. It can also be interpreted as the reduction of uncertainty of one random variable due to the knowledge of the other [8]. Thus, $I(C;X)$ provides a criterion for measuring the effectiveness of a gene variable for class separability [8]. The *mutual information* $I(C;X)$ between variables C and X is defined as

$$I(C;X) \triangleq H(C) - H(C|X). \tag{A.2.5}$$

This function is symmetric with respect to C and X, i.e., $I(C;X) = I(X;C)$. If X has continuous components, then the definition of $I(C;X)$ is given by

$$I(C;X) \triangleq H(X) - H(X|C) = H(C) - H(C|X) \tag{A.2.6}$$

$$= H(C) + H(X) - H(C;X) = \sum_{c=1}^{N_c}\int p(c,x)\log\frac{p(c,x)}{P(c)\,p(x)}dx. \tag{A.2.7}$$

It is known that mutual information is always non-negative, i.e., $I(C;X) \geq 0$ [5].

Estimation of mutual information

To compute the mutual information between the random variable describing the class C and the random variable describing a feature X, it is necessary to know the distributions $P(c)$ and $p_c(x)$. The probability function $P(c)$ is usually estimated using the ratio of samples in the different classes of the training set, i.e., $P(c) \approx \frac{N_c}{M}$, where M is the sample size.

To estimate the conditional distributions $p_c(x)$, a frequently used technique is the Gaussian density estimate or Parzen density estimate. Here we adopt log normal density estimate given by

$$p_c(x) \approx \frac{1}{(2\pi)^{1/2}\sigma_c x}\exp\left\{-\frac{1}{2\sigma_c^2}(\log x - \mu_c)^2\right\},\quad c = 1,\ldots,N_c, \tag{A.2.8}$$

where μ_c and σ_c^2 are the estimates of the mean and the variance of the log data in the cth class. To estimate the mutual information, the integral can be approximated by a discrete sum using sample observations. It follows that the mutual information $I(C;X)$ is estimated using

$$H(C) \approx -\sum_{c=1}^{N_c} P(c) \log P(c), \tag{A.2.9}$$

and

$$H(C|X) \approx -\sum_{c=1}^{N_c} P(c) \sum_{x \in X} p(c,x) \log \frac{P(c)\, p_c(x)}{p(x)}, \tag{A.2.10}$$

where $p_c(x)$ is the estimated conditional distribution and $p(x)$ is the estimated marginal distribution obtained according to

$$p(x) \approx \sum_{c=1}^{N_c} P(c)\, p_c(x).$$

Note that the mutual information defined in (A.2.7) is not normalized; and $I(C;X)$ can be quite small even if X and C are highly correlated since $H(X)$ and $H(C)$ may be small. Therefore, we normalize $I(X;C)$ by the joint entropy of each of the contributing sequences [23]:

$$\bar{I}(C;X) = \frac{I(C;X)}{H(C;X)} = \frac{H(C) + H(X)}{H(C;X)} - 1. \tag{A.2.11}$$

Note that the above discussion is suitable for multivariate X.

Mutual information estimation based on bootstrap

In practice, the sample size M is typically small. In order to get a more accurate estimate of the mutual information, we resort to the bootstrap technique [32].

Let $\boldsymbol{z}$ denote the vector of N gene variables. Denote $\boldsymbol{Z} = [\boldsymbol{z}(1), \boldsymbol{z}(2), \ldots, \boldsymbol{z}(M)]$ as M realizations (i.e., samples) of $\boldsymbol{z}$. At each iteration of the bootstrap procedure, M random draws are performed on $\boldsymbol{Z}$ to form a "resample" $\boldsymbol{Z}^* = [\boldsymbol{z}^*(1), \boldsymbol{z}^*(2), \ldots, \boldsymbol{z}^*(M)]$, and the mutual information is computed based on the resample. The bootstrap method for estimating the mutual information is summarized as follows:

- For $n = 1, 2, \ldots, Q$
  - — Resample: Draw a random sample $\boldsymbol{Z}_n^*$ of M values from $\boldsymbol{Z}$;
  - — Calculate the estimated mutual information $\bar{I}_n$ based on the resample $\boldsymbol{Z}_n^*$;
- Sort the bootstrap estimates $\bar{I}_n, n = 1, \ldots, Q$, according to increasing order to obtain $\bar{I}_{k_1}, \bar{I}_{k_2}, \ldots, \bar{I}_{k_Q}$;
- The desired $(1 - \alpha)100\%$ bootstrap confidence interval is $(\bar{I}_{k_p}, \bar{I}_{k_q})$, where $p = \lfloor Q\alpha/2 \rfloor$ and $q = Q - p + 1$;
- The final estimated mutual information $\hat{I}$ is the mean of the mutual information values in the interval $(\bar{I}_{k_p}, \bar{I}_{k_q})$.

We set $\alpha = 0.95$ and $Q = 1000$ in our simulations.

Gene selection

The existing feature selection methods using mutual information are the MIFS (mutual information feature selector) algorithm [3] and its improved algorithm MIFS-U (mutual information feature selector under uniform information distribution) algorithm [15]. The greedy acquisition algorithm in [3] is as follows: Given an initial set $V = \{X_1, X_2, \ldots, X_N\}$ with N random variables and the class variable C, find subset $S \subset V$ with k variables that maximizes the mutual information $I(C; S)$. A local suboptimal search algorithm [15] is developed to find the subset S. The two algorithms do not work well for problems with very small sets of data samples.

Here the proposed gene selection procedure using mutual information is quite simple. The genes are selected according to the mutual information maximization criterion, namely, the strongest gene is given by

$$\hat{X}_1 = \arg\max_{X \in V} \hat{I}(C; X), \tag{A.2.12}$$

and the second strongest gene is given by

$$\hat{X}_2 = \arg\max_{X \in V \setminus X_1} \hat{I}(C; X). \tag{A.2.13}$$

We repeat the above procedure until we get the given number of strongest genes. The same procedure can be applied to the feature selection for the wavelet coefficients obtained from wavelet representations of microarray. After obtaining the P strongest genes or features, we denote the profiles as y_{ij} for $i = 1, \ldots, M$ and $j = 1, \ldots, P$. The P is selected by experience. In cancer classification, we cannot select a large value for P due to the small sample size. We will specify this value for each experiment in Sec. 5.

Appendix 3. Nonlinear Probit Bayesian Classifier

Let $\boldsymbol{w} = [w_1, \ldots, w_M]^T$ denote the class labels, where $w_i = 0$ indicates the sample i being cancer 1, and $w_i = 1$ indicates the sample i being cancer 2 or no cancer, for $i = 1, 2, \ldots, M$. Assume $y_1, \ldots, y_P$ are the P genes or features. Let y_{ij} be the measurement of the expression level of the jth gene or feature for the ith sample where $j = 1, 2, \ldots, P$. Define the gene expression matrix as

$$\boldsymbol{Y} = \begin{bmatrix} y_{11} & y_{12} & \cdots & y_{1P} \\ y_{21} & y_{22} & \cdots & y_{2P} \\ \vdots & \vdots & \ddots & \vdots \\ y_{M1} & y_{M2} & \cdots & y_{MP} \end{bmatrix}. \tag{A.3.1}$$

Due to the small sample size, here we adopt a probit regression model composed of a linear term plus a nonlinear term. Then w_i and the gene expression levels are related through

$$P(w=1|y_1,\ldots,y_P)=\Phi\left(\sum_{i=1}^{P}\alpha_i y_i+\sum_{k=1}^{2}\beta_k\phi_k(y_1,\ldots,y_P)\right), \tag{A.3.2}$$

$$\text{with}\quad \phi_1(y_1,\ldots,y_P)\triangleq\exp\{-\lambda_1\|\boldsymbol{y}-\boldsymbol{\mu}_1\|^2\}, \tag{A.3.3}$$

$$\text{and}\quad \phi_2(y_1,\ldots,y_P)\triangleq\exp\{-\lambda_2\|\boldsymbol{y}-\boldsymbol{\mu}_2\|^2\}, \tag{A.3.4}$$

where $\boldsymbol{\alpha}=(\alpha_1,\alpha_2,\ldots,\alpha_P)^T$ and $\boldsymbol{\beta}=[\beta_1,\beta_2]^T$ are regression parameters, Φ is the standard normal cumulative distribution function, $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$ are the centers of the two clusters obtained by using the fuzzy c means clustering algorithm, and the parameters λ_1 and λ_2 are empirically set as 2.0 and 4.0, respectively. Define

$$z=\sum_{i=1}^{P}\alpha_i y_i+\sum_{k=1}^{2}\beta_k\phi_k(y_1,\ldots,y_P)+e, \tag{A.3.5}$$

where $e\sim\mathcal{N}(0,1)$. Denote $\boldsymbol{y}_i=[y_{i1},\cdots,y_{iP}], i=1,\ldots,M$. Since z can take $z_1,\ldots,z_M$ corresponding to $\boldsymbol{y}_i$, the above equation can be rewritten in a matrix form as

$$\boldsymbol{z}=\boldsymbol{Y}_1\boldsymbol{\alpha}+\boldsymbol{Y}_2\boldsymbol{\beta}+\boldsymbol{e}, \tag{A.3.6}$$

where $\boldsymbol{z}\triangleq[z_1,\ldots,z_M]^T$, $\boldsymbol{Y}_1\triangleq[\boldsymbol{y}_1^T,\ldots,\boldsymbol{y}_M^T]^T$, $\boldsymbol{\alpha}\triangleq[\alpha_1,\ldots,\alpha_P]^T$, $\boldsymbol{\beta}\triangleq[\beta_1,\beta_2]^T$, $\boldsymbol{e}\triangleq[e_1,\ldots,e_M]\sim\mathcal{N}(0,\mathcal{I}_M)$, and

$$\boldsymbol{Y}_2\triangleq\begin{bmatrix}\phi_1(\boldsymbol{y}_1) & \phi_2(\boldsymbol{y}_1)\\ \vdots & \vdots\\ \phi_1(\boldsymbol{y}_M) & \phi_2(\boldsymbol{y}_M)\end{bmatrix}.$$

Here we assume $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are independent of each other. The prior distributions of $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are assumed as $\boldsymbol{\alpha}\sim\mathcal{N}(0,c_1\boldsymbol{\mathcal{I}}_P)$ and $\boldsymbol{\beta}\sim\mathcal{N}(0,c_2\boldsymbol{\mathcal{I}}_2)$, where c_1 and c_2 are two constants set as 100 in this study. We first transform $\boldsymbol{Y}_i$ such that $\boldsymbol{Y}_i^T\boldsymbol{Y}_i$ becomes an orthogonal matrix, e.g., suppose $\boldsymbol{Y}_i^T\boldsymbol{Y}_i=\boldsymbol{Q}_i\boldsymbol{D}_i\boldsymbol{Q}_i^T$ such that $\boldsymbol{Q}_i\boldsymbol{Q}_i^T=\boldsymbol{\mathcal{I}}$ and $\boldsymbol{D}_i$ is diagonal, and then set $\boldsymbol{Y}_i\leftarrow\boldsymbol{Y}_i\boldsymbol{Q}_i\boldsymbol{D}_i^{-\frac{1}{2}}$ for $i=1,2$. Denote $V=(1+c_1^{-1})^{-1}\boldsymbol{\mathcal{I}}_P$ and $U=(1+c_2^{-1})^{-1}\boldsymbol{\mathcal{I}}_2$. After straightforward computation,

$$\begin{aligned}p(\boldsymbol{\alpha}|\boldsymbol{z},\boldsymbol{\beta})&\propto p(\boldsymbol{z}|\boldsymbol{\alpha},\boldsymbol{\beta})\,p(\boldsymbol{\alpha})\\ &\propto\|V\|^{-\frac{1}{2}}\exp\left[-\frac{1}{2}(\boldsymbol{\alpha}-\Xi_1)^TV^{-1}(\boldsymbol{\alpha}-\Xi_1)^T\right],\end{aligned} \tag{A.3.7}$$

$$\text{with}\quad \Xi_1\triangleq V^{-1}\boldsymbol{Y}_1^T(\boldsymbol{z}-\boldsymbol{Y}_2\boldsymbol{\beta}),$$

namely $(\boldsymbol{\alpha}|\boldsymbol{z},\boldsymbol{\beta})\sim\mathcal{N}(\Xi_1,V)$. Similarly, we have

$$\begin{aligned}p(\boldsymbol{\beta}|\boldsymbol{z},\boldsymbol{\alpha})&\propto p(\boldsymbol{z}|\boldsymbol{\alpha},\boldsymbol{\beta})p(\boldsymbol{\beta})\\ &\propto\|U\|^{-\frac{1}{2}}\exp\left[-\frac{1}{2}(\boldsymbol{\beta}-\Xi_2)^TU^{-1}(\boldsymbol{\beta}-\Xi_2)\right],\end{aligned} \tag{A.3.8}$$

$$\text{with}\quad \Xi_2\triangleq U^{-1}\boldsymbol{Y}_2^T(\boldsymbol{z}-\boldsymbol{Y}_1\boldsymbol{\alpha}),$$

namely $(\boldsymbol{\beta}|\boldsymbol{z}, \boldsymbol{\alpha}) \sim \mathcal{N}(\Xi_2, U)$. The Gibbs sampling algorithm for estimation of $\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{z}$ is as follows:

- Set initial values of $\boldsymbol{z}^{(0)}, \boldsymbol{\alpha}^{(0)}, \boldsymbol{\beta}^{(0)}$.
- Draw $\boldsymbol{\alpha}^{(t)}$ according to $\boldsymbol{\alpha} \sim \mathcal{N}(\Xi_1, V)$.
- Draw $\boldsymbol{\beta}^{(t)}$ according to $\boldsymbol{\beta} \sim \mathcal{N}(\Xi_2, U)$.
- Sample $z_i^{(t)}$, $i = 1, \ldots, M$ from a truncated normal distribution as follows [21]:

$$p(z_i|\boldsymbol{\alpha}^{(t)}, \boldsymbol{\beta}^{(t)}, w_i = 1) \propto \mathcal{N}(\boldsymbol{y}_i\boldsymbol{\alpha}^{(t)} + \boldsymbol{\vartheta}_i\boldsymbol{\beta}^{(t)}, 1)1_{\{z_i>0\}}, \tag{A.3.9}$$

$$p(z_i|\boldsymbol{\alpha}^{(t)}, \boldsymbol{\beta}^{(t)}, w_i = 0) \propto \mathcal{N}(\boldsymbol{y}_i\boldsymbol{\alpha}^{(t)} + \boldsymbol{\vartheta}_i\boldsymbol{\beta}^{(t)}, 1)1_{\{z_i<0\}}, \tag{A.3.10}$$

where $\boldsymbol{\vartheta}_i = [\phi_1(\boldsymbol{y}_i), \phi_2(\boldsymbol{y}_i)]$.

In this study, 1,500 iterations are implemented with the first 500 as the burn-in period. We obtain the Monte Carlo samples $\{\boldsymbol{\alpha}^{(t)}, \boldsymbol{\beta}^{(t)}, \boldsymbol{z}^{(t)}, t = 1, \ldots, T\}$. The estimate of the class label is given by

$$P(w = 1|\boldsymbol{Y}_1, \boldsymbol{Y}_2) = \frac{1}{T}\sum_{t=501}^{T} \Phi(\boldsymbol{Y}_1\boldsymbol{\alpha}^{(t)} + \boldsymbol{Y}_2\boldsymbol{\beta}^{(t)}). \tag{A.3.11}$$

The classification is given by

$$\hat{w} = \begin{cases} 1, & \text{if } P(w|\boldsymbol{Y}_1, \boldsymbol{Y}_2) \geq 0.5; \\ 0, & \text{if } P(w|\boldsymbol{Y}_1, \boldsymbol{Y}_2) < 0.5. \end{cases} \tag{A.3.12}$$

Acknowledgments

This research was supported by the National Human Genome Research Institute. X. Wang was supported in part by the US National Science Foundation under grant DMS-0225692.

References

[1] Albert J, Chib S, Bayesian analysis of binary and polychotomous response data, *J. Am. Stat. Assoc.* **88**:669–679, 1993.

[2] Alter O, Brown PO, Botstein D, Singular value decomposition for genome-wide expression data processing and modeling, *Proc. Natl. Acad. Sci. USA* **97**:10101–10106, 2000.

[3] Battiti R, Using mutual information for selecting features in supervised neural net learning, *IEEE T. Neural Networ.* **5**:537–550, 1994.

[4] Butz T, Thiran J-P, Feature-space mutual information for multi-modal signal processing, with application to medical image registration, *EUSIPCO 2002*, Toulouse, Switzerland.

[5] Cover TM, Thomas JA, *Elements of Information Theory*, Wiley, New York, 1991.

[6] Daubechies I, *Ten Lectures on Wavelets*, Philadelphia, 1992.

[7] Donoho D, Johnstone I, Multiple shrinkage and subset selection in wavelets, *Biometrica* **81**:425–455, 1994.

[8] Grall-Maes E, Beauseroy P, Mutual information-based feature extraction on the time-frequency plane, *IEEE T. Signal Proces.* **50**:779–790, 2002.

[9] Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, Mesirov JP, Coller H, Loh ML, Downing JR, Caligiuri MA, Bloomfield CD, Lander ES, Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring, *Science* **286**:531–537, 1999.

[10] Guyon I, Weston J, Barnhill S, Vapnik V, Gene selection for cancer classification using support vector machines, *Machine Learning* **46**:389–422, 2002.

[11] Hedenfalk I, Duggan D, Chen Y, Radmacher M, Bittner M, Simon R, Meltzer P, Gusterson B, Esteller M, Raffeld M, Yakhini Z, Ben-Dor A, Dougherty ER, Kononen J, Bubendorf L, Fehrle W, Pittaluga S, Gruvberger S, Loman N, Johannsson O, Olsson H, Wilfond B, Sauter G, Kallioniemi O-P, Borg A, Trent J, Gene expression profiles in hereditary breast cancer, *New Engl. J. Med.* **344**:539–548, 2001.

[12] Kim S, Dougherty ER, Chen Y, Sivakumar K, Meltzer P, Trent JM, Bittner M, Multivariate measurement of gene expression relations, *Genomics* **67**:201–209, 2000.

[13] Kim S, Dougherty ER, Barrea J, Chen Y, Bittner M, Trent JM, Strong feature sets from small samples, *J. Comput. Biol.* **9**:127–146, 2002.

[14] Kim S, Li H, Chen Y, Cao N, Dougherty ER, Bittner ML, Suh EB, Can Markov chain models mimic biological regulation? *J. Biol. Syst.* **10**:337–357, 2002.

[15] Kwak N, Choi CH, Input feature selection for classification problems, *IEEE T. Neural Networ.* **13**:143–159, 2002.

[16] Lee KE, Sha N, Dougherty ER, Vannucci M, Mallick BK, Gene selection: A Bayesian variable selection approach, *Bioinformatics* **19**:90–97, 2003.

[17] Li L, Weinberg CR, Darden TA, Pedersen LG, Gene selection for sample classification based on gene expression data: Study of sensitivity to choice of parameters of the GA/KNN method, *Bioinformatics* **17**:1131–1142, 2001.

[18] Lio P, Wavelets in bioinformatics and computational biology: State of art and perspectives, *Bioinformatics* **19**:2–9, 2003.

[19] Mallat SG, A theory for multiresolution signal decomposition: The wavelet representation, *IEEE T. Pattern Anal. Machine Intell.* **11**:674–693, 1989.

[20] Nguyen DV, Rocke DM, Tumor classification by partial least squares using microarray gene expression data, *Bioinformatics* **18**:39–50, 2002.

[21] Robert C, Simulation of truncated normal variables, *Stat. Comput.* **5**:121–125, 1995.

[22] Smith M, Kohn R, Nonparametric regression using Bayesian variable selection, *J. Econometrics* **75**:317–344, 1997.

[23] Studholme C, Hill DLG, Hawkes DJ, An overlap invariant entropy measure of 3D medical image alignment, *Pattern Recogn.* **32**:71–86, 1999.

[24] Tabus I, Astola J, On the use of MDL principle in gene expression prediction, *J. Appl. Signal Proces.* **81**:297–303, 2001.

[25] Tabus I, Rissenan J, Astola J, Classification and feature gene selection using the normalized maximum likelihood for discrete regression, *Signal Proces.* **83**:713–727, 2003.

[26] Walpole RE, Myers RH, *Probability and Statistics for Engineers and Scientists*, Macmillan, New York, 1989.

[27] Xing E, Jordan M, Karp R, Feature selection for high dimensional genomic microarray data, *Proc. 8th Int. Conf. Machine Learning*, Williams College, Massachussets, 2001.

[28] Yau P, Kohn R, Wood S, Bayesian variable selection and model averaging in high dimensional multinomial nonparametric regression, to appear in *J. Comput. Graph. Stat.*

[29] Zhou X, Wang X, Dougherty ER, Construction of genomic networks using mutual-information clustering and reversible-jump Markov-Chain-Monte-Carlo predictor design, *Signal Proces.* **83**:745–761, 2003.

[30] Zhou X, Wang X, Dougherty ER, Binarization of microarray data based on a mixture model, *J. Mol. Cancer Ther.* **2**:679–684, 2003.

[31] Zhou X, Wang X, Dougherty ER, Gene clustering based on cluster-wide mutual information, to appear in *J. Comput. Biol.*

[32] Zoubir AM, Boashash B, The bootstrap and its application in signal processing, *IEEE Signal Proces. Mag.* **15**:56–76, 1998.