Journal of Bioinformatics and Computational Biology
Vol. 3, No. 2 (2005) 185–205
© Imperial College Press

Imperial College Press
www.icpress.co.uk

MINIMUM REDUNDANCY FEATURE SELECTION FROM MICROARRAY GENE EXPRESSION DATA

CHRIS DING[†] and HANCHUAN PENG[‡]

[†]Computational Research Division and [‡]Life Sciences/Genomics Division
Lawrence Berkeley National Laboratory, University of California
Berkeley, CA, 94720, USA
[†]chqding@lbl.gov
[‡]hpeng@lbl.gov

Received 1 October 2003
Revised 16 June 2004
Accepted 16 June 2004

How to selecting a small subset out of the thousands of genes in microarray data is important for accurate classification of phenotypes. Widely used methods typically rank genes according to their differential expressions among phenotypes and pick the top-ranked genes. We observe that feature sets so obtained have certain redundancy and study methods to minimize it. We propose a minimum redundancy — maximum relevance (MRMR) feature selection framework. Genes selected via MRMR provide a more balanced coverage of the space and capture broader characteristics of phenotypes. They lead to significantly improved class predictions in extensive experiments on 6 gene expression data sets: NCI, Lymphoma, Lung, Child Leukemia, Leukemia, and Colon. Improvements are observed consistently among 4 classification methods: Naïve Bayes, Linear discriminant analysis, Logistic regression, and Support vector machines.

Supplimentary: The top 60 MRMR genes for each of the datasets are listed in http://crd.lbl.gov/∼cding/MRMR/. More information related to MRMR methods can be found at http://www.hpeng.net/.

Keywords: Cancer classification; gene selection; gene expression analysis; SVM; Naïve Bayes.

1. Introduction

Discriminant analysis is now widely used in bioinformatics tasks, such as distinguishing cancer tissues from normal tissues[2] or one cancer subtype from another,[1] predicting protein fold or super-family from its sequence,[8,16] etc. A critical issue in discriminant analysis is feature selection: instead of using all available variables (features or attributes) in the data, one selectively chooses a subset of features to be used in the discriminant system. There are a number of advantages of feature selection, to mention a few:

- dimension reduction to reduce the computational cost;

- reduction of noise to improve the classification accuracy; and
- more interpretable features or characteristics that can help identify and monitor the target diseases or function types.

These advantages are typified in DNA microarray gene expression profiles. Of the tens of thousands of genes in experiments, only a smaller number of them show strong correlation with the targeted phenotypes. For example, for a two-class cancer subtype classification problem, 50 informative genes are usually sufficient.[13] There are studies suggesting that only a few genes are sufficient.[23,39] Thus, computation is reduced while prediction accuracy is increased via effective feature selection. When a small number of genes are selected, their biological relationship with the target diseases is more easily identified. These "marker" genes thus provide additional scientific understanding of the problem. Selecting an effective and more representative feature set is the subject of this paper.

There are two general approaches to feature selection: filters and wrappers.[18,20] Filter type methods are essentially data pre-processing or data filtering methods. Features are selected based on the intrinsic characteristics, which determine their relevance or discriminant power with regard to the target classes. Simple methods based on mutual information,[4] statistical tests (t-test, F-test) have been shown to be effective.[7,10,13,25] More sophisticated methods are also developed.[3,19] Filter methods can be computed easily and very efficiently. The characteristics in the feature selection are uncorrelated to that of the learning methods, therefore they have better generalization property.

In wrapper type methods, feature selection is "wrapped" around a learning method: the usefulness of a feature is directly judged by the estimated accuracy of the learning method. One can often obtain a set with a small number of non-redundant features,[5,18,23,39] which gives high prediction accuracy, because the characteristics of the features match well with the characteristics of the learning method. Wrapper methods typically require extensive computation to search the best features.

2. Minimum Redundancy Gene Selection

One common practice of filter type methods is to simply select the top-ranked genes, say the top 50.[13] More sophisticated regression models or tests along this line were also developed.[29,34,38] So far, the number of features, m, retained in the feature set is set by human intuition with trial-and-error, although there are studies on setting m based on certain assumptions on data distributions.[23] A deficiency of this simple ranking approach is that the features could be correlated among themselves.[9,17] For example, if gene g_i is ranked high for the classification task, other genes highly correlated with g_i are also likely to be selected by the filter method. It is frequently observed[23,39] that simply combining a "very effective" gene with another "very effective" gene often does not form a better feature set. One reason is that these two genes could be highly correlated. This raises the issue of "redundancy" of feature set.

The fundamental problem with redundancy is that the feature set is not a comprehensive representation of the characteristics of the target phenotypes. There are two aspects of this problem. (1) Efficiency. If a feature set of 50 genes contains quite a number of mutually highly correlated genes, the true "independent" or "representative" genes are therefore much fewer, say 20. We can delete the 30 highly correlated genes without effectively reducing the performance of the prediction; this implies that 30 genes in the set are essentially "wasted". (2) Broadness. Because the features are selected according to their discriminative powers, they are not maximally representative of the original space covered by the entire dataset. The feature set may represent one or several dominant characteristics of the target phenotypes, but these could still be narrow regions of the relevant space. Thus, the generalization ability of the feature set could be limited.

Based on these observations, we propose to expand the representative power of the feature set by requiring that features are maximally dissimilar to each other, for example, their mutual Euclidean distances are maximized, or their pair-wise correlations are minimized. These minimum redundancy criteria are supplemented by the usual maximum relevance criteria such as maximal mutual information with the target phenotypes. We therefore call this approach the minimum redundancy — maximum relevance (MRMR) approach. The benefits of this approach can be realized in two ways. (1) With the same number of features, we expect the MRMR feature set to be more representative of the target phenotypes, therefore leading to better generalization property. (2) Equivalently, we can use a smaller MRMR feature set to effectively cover the same space as a larger conventional feature set does.

The main contribution of this paper is to point out the importance of minimum redundancy in gene selection and provide a comprehensive study. One novel point is to directly and explicitly reduce redundancy in feature selection via filter approach. Our extensive experiments indicate that features selected in this way lead to higher accuracy than features selected via maximum relevance only.

3. Criterion Functions of Minimum Redundancy

3.1. *MRMR for categorical (discrete) variables*

If a gene has expressions randomly or uniformly distributed in different classes, its mutual information with these classes is zero. If a gene is strongly differentially expressed for different classes, it should have large mutual information. Thus, we use mutual information as a measure of relevance of genes.

For discrete/categorical variables, the mutual information I of two variables x and y is defined based on their joint probabilistic distribution $p(x,y)$ and the respective marginal probabilities $p(x)$ and $p(y)$:

$$I(x,y) = \sum_{i,j} p(x_i, y_j) \log \frac{p(x_i, y_j)}{p(x_i)p(y_j)}. \tag{1}$$

For categorical variables, we use mutual information to measure the level of "similarity" between genes. The idea of minimum redundancy is to select the genes such that they are mutually maximally dissimilar. Minimal redundancy will make the feature set a better representation of the entire dataset. Let S denote the subset of features we are seeking. The minimum redundancy condition is

$$\min W_I, \quad W_I = \frac{1}{|S|^2} \sum_{i,j \in S} I(i,j), \tag{2}$$

where we use $I(i,j)$ to represent $I(g_i, g_j)$ for notational simplicity, and $|S|(= m)$ is the number of features in S.

To measure the level of discriminant powers of genes when they are differentially expressed for different target classes, we again use mutual information $I(h, g_i)$ between targeted classes $h = \{h_1, h_2, \ldots, h_K\}$ (we call h the classification variable) and the gene expression g_i. $I(h, g_i)$ quantifies the relevance of g_i for the classification task. Thus the maximum relevance condition is to maximize the total relevance of all genes in S:

$$\max V_I, \quad V_I = \frac{1}{|S|} \sum_{i \in S} I(h,i), \tag{3}$$

where we refer to $I(h, g_i)$ as $I(h,i)$.

The MRMR feature set is obtained by optimizing the conditions in Eqs. (2) and (3) simultaneously. Optimization of both conditions requires combining them into a single criterion function. In this paper we treat the two conditions equally important, and consider two simplest combination criteria:

$$\max(V_I - W_I), \tag{4}$$

$$\max(V_I / W_I). \tag{5}$$

Our goal here is to see whether the MRMR approach is effective in its simplest forms. More refined variants can be easily studied later on.

Exact solution to the MRMR requirements requires $\mathrm{O}(N^{|S|})$ searches (N is the number of genes in the whole gene set, Ω). In practice, a near optimal solution is sufficient. In this paper, we use a simple heuristic algorithm to resolve this MRMR optimization problem.

In our algorithm, the first feature is selected according to Eq. (3), i.e. the feature with the highest $I(h,i)$. The rest features are selected in an incremental way: earlier selected features remain in the feature set. Suppose m features are already selected for the set S, and we want to select additional features from the set $\Omega_S = \Omega - S$ (i.e.

all genes except those already selected). We optimize the following two conditions:

$$\max_{i\in\Omega_S} I(h,i), \tag{6}$$

$$\min_{i\in\Omega_S} \frac{1}{|S|}\sum_{j\in S} I(i,j). \tag{7}$$

The condition in Eq. (6) is equivalent to the maximum relevance condition in Eq. (3), while Eq. (7) is an approximation of the minimum redundancy condition of Eq. (2). The two ways to combine relevance and redundancy, Eqs. (4) and (5), lead to the selection criteria of a new feature:

(1) MID: Mutual Information Difference criterion,
(2) MIQ: Mutual Information Quotient criterion,

as listed in Table 1. These optimizations can be computed efficiently in $O(|S|\cdot N)$ complexity.

3.2. *MRMR for continuous variables*

For continuous data variables (or attributes), we can choose the F-statistic between the genes and the classification variable h as the score of maximum relevance. The F-test value of gene variable g_i in K classes denoted by h has the following form[7,10]:

$$F(g_i,h) = \left[\sum_k n_k(\bar{g}_k - \bar{g})/(K-1)\right] \Big/ \sigma^2, \tag{8}$$

where $\bar{g}$ is the mean value of g_i in all tissue samples, $\bar{g}_k$ is the mean value of g_i within the kth class, and $\sigma^2 = [\sum_k (n_k - 1)\sigma_k^2]/(n-K)$ is the pooled variance (where n_k and σ_k are the size and the variance of the kth class). F-test will reduce to the t-test for 2-class classification, with the relation $F = t^2$. Hence, for the feature set S, the maximum relevance can be written as:

$$\max V_F, \quad V_F = \frac{1}{|S|}\sum_{i\in S} F(i,h). \tag{9}$$

Table 1. Different schemes to search for the next feature in MRMR optimization conditions.

Type	Acronym	Full Name	Formula
Discrete	MID	Mutual information difference	$\max_{i\in\Omega_S}\left[I(i,h) - \frac{1}{\vert S\vert}\sum_{j\in S} I(i,j)\right]$
	MIQ	Mutual information quotient	$\max_{i\in\Omega_S}\left\{I(i,h)\Big/\left[\frac{1}{\vert S\vert}\sum_{j\in S} I(i,j)\right]\right\}$
Continuous	FCD	F-test correlation difference	$\max_{i\in\Omega_S}\left[F(i,h) - \frac{1}{\vert S\vert}\sum_{j\in S} \vert c(i,j)\vert\right]$
	FCQ	F-test correlation quotient	$\max_{i\in\Omega_S}\left\{F(i,h)\Big/\left[\frac{1}{\vert S\vert}\sum_{j\in S} \vert c(i,j)\vert\right]\right\}$

The minimum redundancy condition may be specified in several ways. If we use Pearson correlation coefficient $c(g_i,\, g_j) = c(i,\, j)$, the condition is

$$\min W_c, \quad W_c = \frac{1}{|S|^2} \sum_{i,j} |c(i,j)|, \tag{10}$$

where we have assumed that both high positive and high negative correlation mean redundancy, and thus take the absolute value of correlations. (We may also use Euclidean distance as a measure of redundancy. As shown in our preliminary results,[9] Euclidean distance is not as effective as correlation.)

Now the simplest MRMR optimization criterion functions involving above conditions are:

(1) FCD: combine F-test with correlation using difference, and
(2) FCQ: combine F-test with correlation using quotient,

as shown in Table 1.

We use the same linear incremental search algorithm as in the discrete variable case in Sec. 3.1. Assume m features have already been selected; the next feature is selected via a simple linear search based on the criteria listed in Table 1 for the above four criterion functions.

4. Class Prediction Methods

4.1. *Naïve Bayes (NB) classifier*

The Naïve Bayes (NB)[24] is one of the oldest classifiers. It is obtained by using the Bayes rule and assuming features (variables) are independent of each other given its class. For a tissue sample s with m gene expression levels $\{g_1, g_2, \ldots, g_m\}$ for the m features, the posterior probability that s belongs to class h_k is

$$p(h_k|s) \propto \prod_{i \in S} p(g_i \,|\, h_k), \tag{11}$$

where $p(g_i \,|\, h_k)$ are conditional tables (or conditional density) estimated from training examples. Despite the independence assumption, NB has been shown to have good classification performance for many real data sets, especially for documents,[24] on par with many more sophisticated classifiers.

4.2. *Support Vector Machine (SVM)*

SVM is a relatively new and promising classification method.[35] It is a margin classifier that draws an optimal hyperplane in the feature vector space; this defines a boundary that maximizes the margin between data samples in two classes, therefore leading to good generalization properties. A key factor in SVM is to use kernels to construct nonlinear decision boundary. We use linear kernels.

Standard SVM is for 2 classes. For multi-class problems, one may construct a multi-class classifier using binary classifiers such as one-against-others or all-against-all.[8] Another approach is to directly construct a multi-class SVM.[37] In this paper, we used the Matlab version of LIBSVM,[15] which uses the one-against-others approach.

4.3. *Linear Discriminant Analysis (LDA)*

Fisher's LDA is a very old classification method. It assumes samples in each class follow a Gaussian distribution. The center and covariance matrix are estimated for each class. We assume that the off-diagonal elements in the covariance are all zero, i.e. different features are uncorrelated. A new sample is classified to the class with the highest probability. Different from other classifiers in this section, LDA assumes data distribution to be Gaussian.

4.4. *Logistic Regression (LR)*

LR[6] forms a predictor variable that is a linear combination of the feature variables. The values of this predictor variable are then transformed into probabilities by a logistic function. This method is widely used for 2-class prediction in biostatistics. It can be extended to multi-class problems as well.

5. Experiments

5.1. *Data sets*

To evaluate the usefulness of the MRMR approach, we carried out experiments on six data sets of gene expression profiles. Two expression datasets popularly used in research literature are the leukemia data of Golub *et al.*[13] and the Colon cancer data of Alon *et al.*[2] As listed in Table 2, both leukemia and colon data sets have two classes. The colon dataset contains both normal and cancerous tissue samples. In the leukemia dataset, the target classes are leukemia subtypes AML and ALL. Note that in the leukemia dataset, the original data come with training and test samples that were drawn from different conditions. Here, we combined them together for the purpose of leave-one-out cross validation.

Table 2. Two-class datasets used in our experiments.

Dataset	Leukemia		Colon Cancer	
Source	Golub *et al.* (1999)		Alon *et al.* (1999)	
# Gene	7070		2000	
# Sample	72		62	
Class	Class name	# Sample	Class name	# Sample
C1	ALL	47	Tumor	40
C2	AML	25	Normal	22

Although two-class classification problems are an important type of tasks, they are relatively easy, since a random choice of class labels would give 50% accuracy. Classification problems with multiple classes are generally more difficult and give a more realistic assessment of the proposed methods. In this paper, we used three multi-class microarray data sets: NCI,[32,33] lung cancer,[12] lymphoma[1] and child leukemia.[22,40] The details of these data sets are summarized in Table 3. For the child leukemia data, for each class, the number of training samples is listed followed by the respective number of testing samples. We note that the number of tissue samples per class is generally small (e.g. < 10 for NCI data) and unevenly distributed (e.g. from 46 to 2 in lymphoma data). This, together with the larger number of classes (e.g. 9 for lymphoma data), makes the classification task more complex than two-class problems. These six data sets provide a comprehensive test suit.

For the two-class problems, we used the two-sided t-test selection method, i.e. we imposed the condition that in the features selected, the number of features with positive t-value is equal to that with negative t-value. Compared to the standard F-test selection, since $F = t^2$, two-sided t-test gives more balanced features whereas F-test does not guarantee the two sides have the equal number of features. The MRMR feature selection schemes of the F-test (as shown in Table 1) can be modified to use two-sided t-test. We denote them as TCD (vs FCD) and TCQ (vs FCQ) schemes.

5.2. *Assessment measure*

For the first 5 datasets, we assessed classification performance using the "Leave-One-Out Cross Validation" (LOOCV). CV accuracy provides more realistic assessment of classifiers which generalize well to unseen data. For presentation clarity, we give the number of LOOCV errors in Tables 4–8.

For the child leukemia data, we selected features using only the training data, and show the testing errors on the testing set in Table 9. This gives examples where the testing samples have never been met in the feature selection process. We considered both cross-validation where features are selected using all samples together and this training/testing procedure, and believed it is a more comprehensive study of the MRMR performance.

In experiments, we compared the MRMR feature sets against the baseline feature sets obtained using standard mutual information, F-statistic or t-statistic ranking to pick the top m features.

5.3. *Discretization for noise reduction*

The original gene expression data are continuous values. We directly classified them using SVM, LDA, and LR. We pre-processed the data so each gene has zero mean value and unit variance.

We also discretized the data into categorical data for two reasons. The first reason is noise reduction because the original readings contain substantial noise.

Table 3. Multi-class datasets used in our experiments (# S is the number of samples).

Dataset	NCI		Lung Cancer		Lymphoma		Child Leukemia	
Source	Ross *et al.* (2000) Scherf *et al.* (2000)		Garber *et al.* (2001)		Alizadeh *et al.* (2000)		Yoeh *et al.* (2002) Li *et al.* (2003)	
# Gene	9703		918		4026		4026	
# S	60		73		96		96	
# Class	9		7		9		9	
Class	Class name	# S	Class name	# S	Class name	# S	Class name	# S
C1	NSCLC	9	AC-group-1	21	Diffuse large B cell lymphoma	46	BCR-ABL	9/6
C2	Renal	9	Squamous	16	Chronic Lympho. leukemia	11	E2A-PBX1	18/9
C3	Breast	8	AC-group-3	13	Activated blood B	10	Hyperdiploid> 50	42/22
C4	Melanoma	8	AC-group-2	7	Follicular lymphoma	9	MLL	14/6
C5	Colon	7	Normal	6	Resting/activated T	6	T-ALL	28/15
C6	Leukemia	6	Small-cell	5	Transformed cell lines	6	TEL-AML1	52/27
C7	Ovarian	6	Large-cell	5	Resting blood B	4	Others	52/27
C8	CNS	5			Germinal center B	2		
C9	Prostate	2			Lymph node/tonsil	2		

Table 4. Lymphoma data (96 samples for 9 classes) LOOCV errors.

			M														
Classifier	Data Type	Method	3	6	9	12	15	18	21	24	27	30	36	42	48	54	60
NB	Discrete	Baseline	38	39	25	29	23	22	22	19	20	17	19	18	18	17	17
		MID	31	15	10	9	9	8	6	7	7	7	4	7	5	5	8
		MIQ	38	26	17	14	14	12	8	8	6	7	5	6	4	3	3
LDA	Discrete	Baseline	40	42	28	26	20	21	21	20	18	19	14	15	13	14	15
		MID	32	15	14	10	7	5	4	5	4	6	5	3	3	4	3
		MIQ	40	29	12	8	8	7	5	6	4	1	1	2	1	2	2
	Continuous	Baseline	66	26	26	17	17	18	18	18	15	11	14	12	11	11	13
		FCD	33	17	16	10	13	11	11	9	8	8	8	8	7	10	9
		FCQ	32	18	11	7	7	8	8	7	8	9	9	9	8	6	6
SVM	Discrete	Baseline	32	29	25	23	20	22	18	13	14	15	11	10	10	8	9
		MID	24	10	7	4	2	3	3	3	3	3	3	3	3	3	3
		MIQ	26	21	13	9	8	7	6	5	5	2	1	1	2	1	2
	Continuous	Baseline	30	24	14	13	12	13	10	11	13	6	8	9	5	6	7
		FCD	24	19	11	13	11	9	10	8	7	8	7	6	5	4	5
		FCQ	31	17	9	7	6	6	8	8	6	7	7	8	7	4	4

Table 5. NCI data (60 samples for 9 classes) LOOCV errors.

			M														
Classifier	Data Type	Method	3	6	9	12	15	18	21	24	27	30	36	42	48	54	60
NB	Discrete	Baseline	29	26	20	17	14	15	12	11	11	13	13	14	14	15	13
		MID	28	15	13	13	6	7	8	7	7	5	8	9	9	8	10
		MIQ	27	21	16	13	13	8	5	5	4	3	1	1	1	1	2
LDA	Discrete	Baseline	35	25	23	20	21	18	19	19	16	19	17	19	17	16	17
		MID	31	20	21	19	16	16	16	16	15	17	16	15	16	16	15
		MIQ	34	31	26	21	21	17	15	14	14	14	10	9	9	8	8
	Continuous	Baseline	41	35	23	21	22	21	20	17	16	17	17	21	19	19	18
		FCD	36	27	21	20	19	18	17	15	18	17	17	17	16	15	14
		FCQ	35	25	23	22	17	18	17	18	13	14	14	12	13	15	15
SVM	Discrete	Baseline	34	29	27	25	21	19	19	19	20	18	17	18	18	18	16
		MID	33	20	19	20	18	17	17	16	17	15	14	14	14	15	16
		MIQ	33	32	20	23	22	22	14	13	13	13	9	8	7	7	8
	Continuous	Baseline	50	33	27	27	24	22	22	20	23	20	17	18	15	16	15
		FCD	41	28	27	22	24	22	20	20	20	19	19	20	17	16	16
		FCQ	44	30	26	26	25	24	23	23	19	19	17	18	17	15	18

Second, prediction methods such as NB prefer categorical data so that conditional probability can be described using a small table. We discretized the observations of each gene expression variable using the respective σ (standard deviation) and μ (mean) for this gene's samples: any data larger than $\mu + \sigma/2$ were transformed to state 1; any data between $\mu - \sigma/2$ and $\mu + \sigma/2$ were transformed to state 0; any data

Table 6. Lung data (73 samples for 7 classes) LOOCV errors.

			M														
Classifier	Data Type	Method	3	6	9	12	15	18	21	24	27	30	36	42	48	54	60
NB	Discrete	Baseline	29	29	24	19	14	15	10	9	12	11	12	12	10	8	9
		MID	31	14	12	11	6	7	7	7	8	6	6	6	6	5	5
		MIQ	40	29	17	9	5	8	6	2	4	3	3	2	4	4	3
LDA	Discrete	Baseline	32	31	22	16	13	10	10	10	10	10	9	9	10	10	10
		MID	32	14	10	9	8	8	7	6	6	6	4	7	6	8	8
		MIQ	36	26	14	7	7	7	8	8	7	7	6	5	6	6	7
	Continuous	Baseline	36	26	14	15	10	9	8	9	12	10	8	10	9	10	10
		FCD	18	13	10	8	8	6	6	7	5	6	7	6	7	6	7
		FCQ	27	12	9	8	7	8	8	7	6	6	6	6	6	6	6
SVM	Discrete	Baseline	38	26	18	21	13	6	10	10	12	11	8	9	10	10	9
		MID	19	11	7	4	7	8	5	5	6	5	5	6	6	7	7
		MIQ	41	28	12	9	8	8	8	7	7	6	6	6	6	6	6
	Continuous	Baseline	30	23	14	15	11	9	9	10	9	8	9	10	10	9	8
		FCD	24	11	13	9	8	7	6	8	7	7	8	5	5	6	7
		FCQ	31	13	12	10	10	6	7	8	8	7	5	6	6	6	7

Table 7. Leukemia data (72 samples for 2 classes) LOOCV errors.

			M														
Classifier	Data Type	Method	1	2	3	4	5	6	7	8	10	12	15	20	30	40	50
NB	Discrete	Baseline	4	2	1	1	1	0	0	0	0	0	1	2	1	1	3
		MID	4	3	1	1	1	0	0	0	0	0	0	1	1	2	1
		MIQ	4	2	1	0	0	0	0	0	0	0	0	0	0	0	0
LDA	Discrete	Baseline	4	2	2	1	1	1	1	1	2	1	2	2	2	2	3
		MID	4	3	2	1	1	1	1	1	1	1	1	2	2	2	1
		MIQ	4	2	2	2	2	2	2	2	2	1	1	0	0	0	0
	Continuous	Baseline	12	4	2	2	3	3	2	3	3	3	2	3	2	2	2
		TCD	12	4	2	2	2	2	2	2	1	2	2	1	2	1	1
		TCQ	12	4	2	2	1	1	1	2	2	2	1	2	1	1	1
SVM	Discrete	Baseline	4	7	4	3	1	2	2	1	2	1	1	2	2	4	3
		MID	4	3	4	3	3	2	2	2	2	1	1	1	2	2	4
		MIQ	4	6	8	2	1	0	0	0	0	0	0	0	0	0	0
	Continuous	Baseline	9	3	2	2	2	3	3	4	2	3	3	3	3	4	1
		TCD	9	3	2	3	3	3	2	4	2	1	3	5	1	1	1
		TCQ	9	3	3	2	2	1	3	0	0	0	1	1	1	1	1
LR	Discrete	Baseline	11	7	2	3	3	1	1	1	3	4	5	3	4	5	11
		MID	11	3	2	3	4	1	2	2	3	4	4	2	5	4	8
		MIQ	11	6	6	2	0	0	0	0	0	0	0	0	1	1	3
	Continuous	Baseline	9	2	2	2	4	5	5	6	7	6	1	2	7	12	8
		TCD	9	2	3	3	5	4	2	5	5	2	6	3	2	1	7
		TCQ	9	2	3	4	3	2	2	1	0	0	0	1	0	2	3

Table 8. Colon data (62 samples for 2 classes) LOOCV errors.

			M														
Classifier	Data Type	Method	1	2	3	4	5	6	7	8	10	12	15	20	30	40	50
NB	Discrete	Baseline	10	7	10	9	9	7	9	9	7	8	8	8	9	9	10
		MID	10	8	8	8	9	10	9	8	7	7	7	8	7	7	7
		MIQ	10	8	12	8	8	6	6	5	4	5	7	7	8	8	7
LDA	Discrete	Baseline	22	14	10	10	9	9	8	8	8	8	7	9	8	9	8
		MID	22	6	7	7	8	8	9	7	8	7	7	8	8	7	7
		MIQ	22	15	12	9	12	10	7	7	7	8	8	7	8	8	8
	Continuous	Baseline	18	9	7	9	8	7	7	8	8	8	7	7	7	9	9
		TCD	18	9	6	8	6	7	7	7	7	7	6	7	7	7	8
		TCQ	18	9	6	6	7	5	6	7	7	7	7	7	7	7	7
SVM	Discrete	Baseline	10	16	7	7	7	7	11	10	13	12	14	14	15	18	18
		MID	10	6	6	10	8	12	11	12	10	12	8	9	9	13	15
		MIQ	10	10	8	12	15	11	7	7	10	12	10	12	11	12	12
	Continuous	Baseline	14	10	9	11	10	9	9	9	10	10	10	13	10	9	8
		TCD	14	10	8	7	7	7	6	7	8	10	8	8	8	13	14
		TCQ	14	10	8	8	7	7	9	9	10	11	10	5	13	12	15
LR	Discrete	Baseline	10	7	8	10	11	11	8	9	11	12	14	18	17	23	21
		MID	10	6	9	7	7	11	10	11	11	13	13	15	16	17	15
		MIQ	10	10	8	12	12	13	8	8	10	13	14	14	18	22	27
	Continuous	Baseline	15	7	8	8	9	9	8	9	11	11	12	9	19	24	16
		TCD	15	7	7	9	9	10	9	10	9	11	14	14	13	18	13
		TCQ	15	7	7	7	8	9	9	9	11	10	14	10	13	20	21

smaller than $\mu - \sigma/2$ were transformed to state -1. These three states correspond to the over-expression, baseline, and under-expression of genes. We also compared different discretization schemes; partial results are summarized in Table 10.

5.4. *Results*

We applied the MRMR feature selection methods on both continuous and descretized data. We performed LOOCV using NB, LDA, SVM and LR on the first 5 datasets. The results of the LOOCV errors are shown in Tables 4–8. Due to the space limitation we only list results of $m = 3, 6, 9, \ldots, 54, 60$ for multi-class datasets and $m = 1, 2, 3, \ldots, 8, 10, \ldots, 50$ for 2-class datasets. From these comprehensive test results, we have following observations.

(1) For discrete datasets, the MRMR MIQ features outperform the baseline features. This is consistent for all the classification methods and for all 5 datasets. Several examples: For lymphoma dataset, using LDA, MIQ leads to 1 errors while baseline leads to 9 errors (see Table 4); using SVM, MIQ leads to 1 errors while baseline leads to 8 errors. For NCI data, using Naïve Bayes, MIQ leads to 1 LOOCV error while baseline leads to 11 errors (we quote the best performance for a given case).

Table 9. Child leukemia data (7 classes, 215 training samples, 112 testing samples) testing errors.

			M														
Classifier	Data Type	Method	3	6	9	12	15	18	24	30	40	50	60	70	80	90	100
NB	Discrete	Baseline	52	47	44	43	36	33	84	86	82	85	79	83	82	82	88
		MID	48	44	36	31	31	30	28	88	85	76	78	87	89	87	101
		MIQ	43	32	30	24	21	18	15	54	70	69	76	83	90	88	91
LDA	Discrete	Baseline	55	47	46	38	34	27	19	28	22	19	15	14	11	8	8
		MID	50	43	32	29	30	29	22	15	13	10	10	9	7	8	9
		MIQ	43	43	34	27	23	21	18	16	11	11	6	4	6	6	4
	Continuous	Baseline	70	69	55	54	54	54	42	31	24	17	15	13	13	10	11
		FCD	55	41	35	34	37	32	34	29	19	15	13	8	4	3	3
		FCQ	66	62	52	42	40	41	24	22	11	10	10	9	8	9	8
SVM	Discrete	Baseline	56	55	49	37	33	33	27	35	29	30	23	20	18	14	13
		MID	45	42	33	33	25	25	29	25	26	22	20	13	10	12	9
		MIQ	38	30	34	33	27	26	24	21	14	15	17	10	7	11	9
	Continuous	Baseline	61	55	54	49	53	59	39	38	33	29	27	21	17	18	19
		FCD	46	44	39	41	48	46	37	35	28	27	29	24	21	25	24
		FCQ	49	46	39	38	27	32	26	29	33	29	26	28	29	26	25

Table 10. LOOCV testing results (#error) for binarized NCI and Lymphoma data using SVM classifier.

		M														
Data Sets	Method	3	6	9	12	15	18	21	24	27	30	36	42	48	54	60
NCI	Baseline	34	25	23	25	19	17	18	15	14	12	12	12	13	12	10
	MID	34	29	23	20	17	19	15	10	12	12	10	10	9	8	10
	MIQ	35	22	22	16	12	11	10	8	5	3	4	4	2	2	3
Lymphoma	Baseline	58	52	44	39	44	17	17	14	16	13	11	10	13	10	12
	MID	27	14	6	10	11	9	9	10	4	5	4	4	4	4	4
	MIQ	24	17	7	8	4	2	1	2	4	3	2	2	2	2	2

(2) For continuous datasets, FCQ features outperform baseline features. This is consistent for LDA and SVM for all three multi-class datasets, and for LDA, SVM and LR for both 2-class datasets (here FCQ is replaced by TCQ). Examples: For lymphoma, using LDA, FCQ leads to 6 errors while baseline leads to 11 errors. For Lung, using SVM, FCQ leads to 5 errors while baseline leads to 8 errors.

(3) Discretization of gene expression data consistently leads to better prediction accuracy. Examples: For lymphoma, using LDA, the best continuous features (selected by FCQ) leads to 6 errors while the best discretized features (selected by MIQ) lead to 1 error. Using SVM, the discrete features also outperform the continuous features. The same conclusions can be drawn for all other 4 datasets. Note that if we restrict to baseline features, this conclusion is not true. In other words, MRMR can make full use of the noise reduction due to discretization.

(4) Naïve Bayes performs better than LDA, SVM, LR. For the multi-class datasets NCI and Lung, NB clearly outperforms other methods. For the 2-class datasets, NB also performs better than other methods. However, for lymphoma, using discrete MIQ features, LDA and SVM performs better than NB.

(5) With MRMR, for discrete data, MIQ outperforms MID; for continuous data, FCQ (or TCQ) is better than FCD (TCD). Both MIQ and FCG use the divisive combination of Eq. (5) while both MID and FCD use the difference combination of Eq. (4). Thus the divisive combination of relevance and redundancy is preferred.

To test the case that features are selected using only the training set and then tested on a separate testing set, we considered the fourth multi-class data set, child leukemia.[22,40] As shown in Table 9, for the first 100 features selected, in most cases the MRMR features lead to significant less errors than baseline features, especially for the LDA and SVM classifiers. For the NB classifier, the better performance of MRMR features can be seen clearly for less than 30 features (note: for this data set the non-robustness of NB to extra-features turns out to be significant for more than 30 features, it is only faithful to compare less than 30 features using NB in this case).

We list the best performance of MRMR features together with the best baseline performance in Table 11. From this table, we can quantify the improvements due to MRMR feature selection. For the first three multi-class datasets, the LOOCV

Table 11. Comparison of the best results (lowest error rates in percentage) of the baseline and MRMR features. Also listed are results in literature (the best results in each paper).

Data	Method	NB	LDA	SVM	LR	Literature
NCI	Baseline	18.33	26.67	25.00	—	14.63[a]
	MRMR	1.67	13.33	11.67	—	5-class: 0,[b] 0[b]
Lymphoma	Baseline	17.71	11.46	5.21	—	3-class: 2.4,[c] 0[c]
	MRMR	3.13	1.04	1.04	—	
Lung	Baseline	10.96	10.96	10.96	—	—
	MRMR	2.74	5.48	5.48	—	
Child leukemia	Baseline	29.46	7.14	11.61	—	5.36[d]
	MRMR	13.39	2.68	6.25	—	
Leukemia	Baseline	0	1.39	1.39	1.39	0[e]
	MRMR	0	0	0	0	1.39[f]
Colon	Baseline	11.29	11.29	11.29	11.29	9.68[e]
	MRMR	6.45	8.06	9.68	9.68	6.45[g]

[a] Ooi and Tan used a genetic algorithm.[28] [b] Nguyen and Rocke[27] used a 5-class subset of NCI dataset and obtained 0% error rate; using the same 5-class subset, our NB achieves also 0% error rate. [c] Nguyen and Rocke used 3-class subset in lymphoma dataset and obtain 2.4% error rate. Using the same 3 classes, our NB leads to zero error. [d] Li *et al.*, using prediction by collective likelihood.[22] [e] Furey *et al.*, using SVM.[11] [f] Lee and Lee, using SVM.[21] [g] Nguyen and Rocke, using PLS.[26]

errors are reduced by a factor of 10. For the child leukemia data, the testing error is reduced by several times, too. For the 2-class datasets, the improvements are also significant, although not as dramatic as for the multi-class datasets.

To better understand the effectiveness of the MRMR approach, we calculated the average relevance V_I and average redundancy W_I [see Eqs. (3) and (2)], as plotted in Figs. 1(a) and 1(b). Although for MID and MIQ the relevance reduces as compared to baseline, the redundancy also reduces considerably. This is most clear for MIQ. The fact that the MIQ feature set is the most effective as seen from Tables 4–8 illustrates the importance of reducing redundancy, the central theme of this research.

The relevance and redundancy for the continuous NCI data are also plotted in Figs. 1(c) and 1(d). For continuous data, the relevance of FCD and FCQ features is reduced slightly from that of baseline, while the redundancy of FCD/FCQ reduce significantly.

It is also interesting to examine how the feature sets selected via different methods intersect. For example, in Fig. 2, we plot the rates of intersecting features for

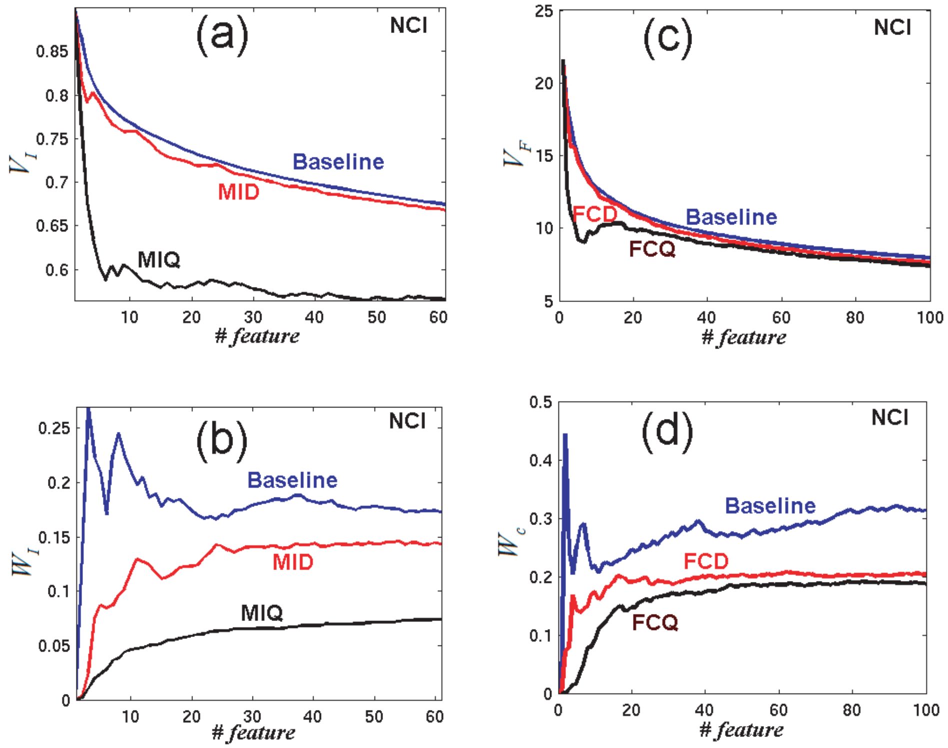

Fig. 1. (a) Relevance V_I, and (b) redundancy for MRMR features on discretized NCI dataset. (c) Relevance V_F, and (d) redundancy W_c on the continuous NCI dataset.

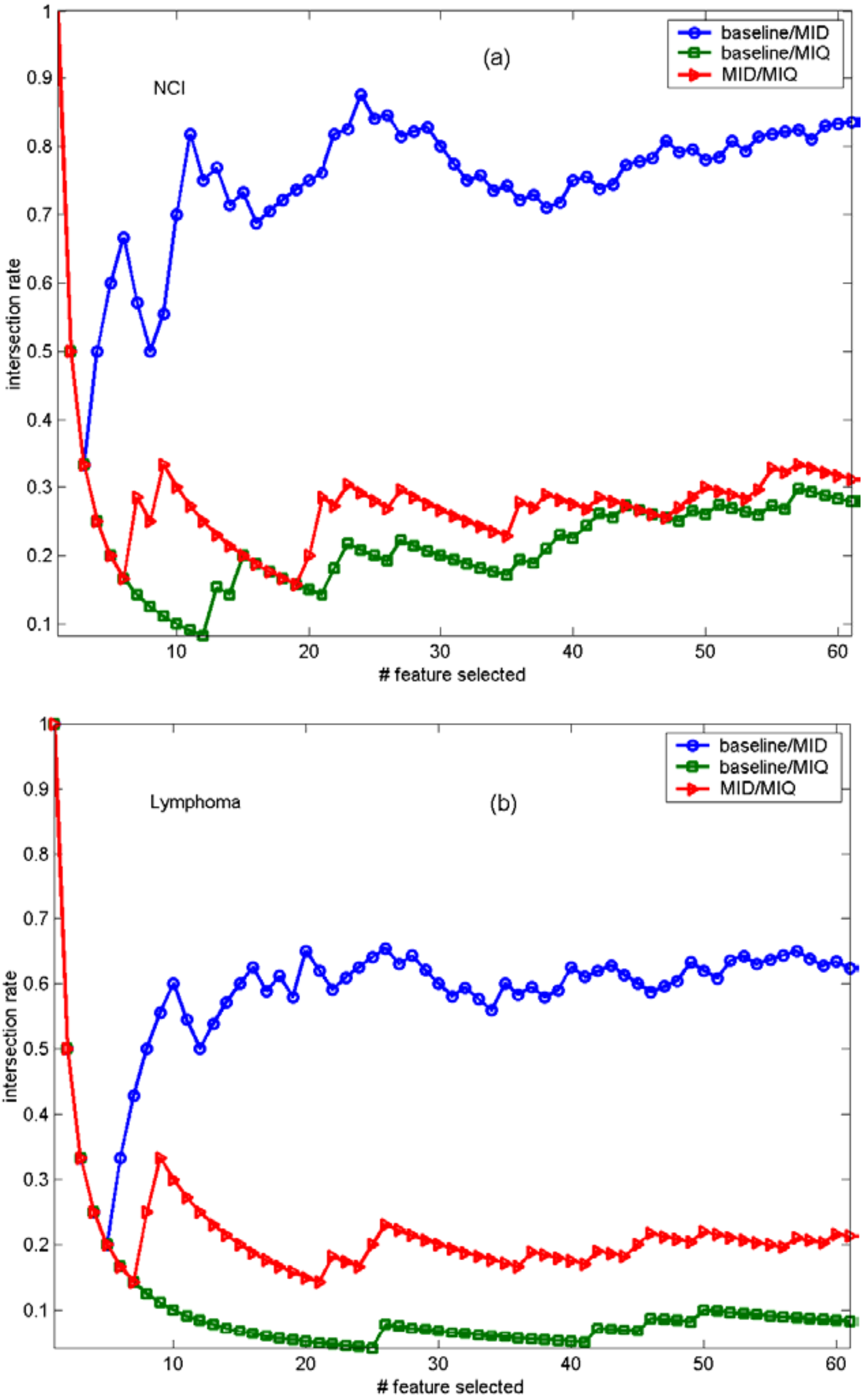

Fig. 2. Intersection of features selected using different methods. (a) NCI data results, and (b) lymphoma data results.

the top m ($1 \leq m \leq 61$) features selected for the NCI and lymphoma data sets. It is clear that the features selected via MID have some chance ($>50\%$) to be also selected by the baseline method. In contrast, features selected using MIQ have much less overlap with those selected using baseline method or MID. This is because the quotient-combination of the MRMR scheme often has a much greater penalty on the redundant features than the difference-combination of MRMR. We note that the results in Fig. 2 are consistent with those in Fig. 1.

It is also of concern how the discretization method will influence the feature selection results. We tested many different discretization parameters to transform the original continuous gene sample data to either 2-state or 3-state variables. The features consequently selected via MRMR always outperform the respective features selected using baseline methods. For simplicity, we only show two exemplary results for the NCI and lymphoma data sets using the SVM classifier. The data are binarized using the mean value of each gene as the threshold of that gene's samples. As illustrated in Table 10, we see that MRMR features always lead to better prediction accuracy than the baseline features. For example, for NCI data, 48 baseline features lead to 13 errors, whereas MIQ features lead to only 2 errors (3% error rate). For lymphoma data, the baseline error is never less than 10, whereas the MIQ features in most cases lead to only 1 or 2 errors (1∼2% error rate). These results are consistent with those shown in Tables 4 and 5. This shows that under different discretization schemes the superiority of MRMR over conventional feature selection schemes is prominent.

5.5. *Comparison with other work*

Results of similar class prediction on microarray gene expression data obtained by others are listed in Table 11. For NCI, our result of LOOCV error rate is 1.67% using NB, whereas Ooi and Tan[28] obtained 14.6% error rate. On the 5-class subset of NCI, Nguyen and Rocke[27] obtained 0% rate, which is the same as our NB results on the same 5-class subset.

For lymphoma data (Table 4), our result is LOOCV error rate of 1%. Using 3 classes only, Nguyen and Rocke[27] obtained 2.4%; on the same 3 classes, our LDA results is 0% error rate.

For child leukemia data, Li *et al.*[22] obtained 5.36% error rate using collective likelihood. In our best case, the MRMR features lead to the 2.68% error rate.

The leukemia data* is a most widely studied dataset. Using MRMR feature selection, we achieve 100% LOOCV accuracy for every classification methods. Furey *et al.*[11] obtained 100% accuracy using SVM, and Lee and Lee[21] obtained 1.39% error rate.

*Many classification studies have used leukemia and colon datasets. Due to space limitation, we only list two for each dataset in Table 11.

For colon data, our result is 6.45% error rate, which is the same as Nguyen and Rocke[27] using PLS. The SVM result from Furey *et al.*[11] is 9.68%.

6. Discussions

In this paper we emphasize the redundancy issue in feature selection and propose a new feature selection framework, the minimum redundancy — maximum relevance (MRMR) optimization approach. We studied several simple forms of this approach with linear search algorithms, and performed experiments on 6 gene expression datasets. Using Naïve Bayes, linear discriminant analysis, logistic regression and SVM class prediction methods, we computed the leave-one-out cross validation accuracy. These experiment results clearly and consistently show that the MRMR feature sets outperform the baseline feature sets based solely on maximum relevance. For discrete features, MIQ is the better choice; for continuous features, FCQ is the better choice. The divisive combination of relevance and redundancy of Eq. (5) appears to lead features with the least redundancy.

The main benefit of MRMR feature set is that by reducing mutual redundancy within the feature set, these features capture the class characteristics in a broader scope. Features selected within the MRMR framework are independent of class prediction methods, and thus do not directly aim at producing the best results for any prediction method. The fact that MRMR features improve prediction for all four methods we tested confirms that these features have better generalization property. This also implies that with fewer features the MRMR feature set can effectively cover the same class characteristic space as more features in the baseline approach.

Our extensive tests, as shown in Tables 4–9, also show that discretization of the gene expressions leads to clearly better classification accuracy than the original continuous data.

For biologists, sometimes the redundant features might also be important. A Bayesian clustering method[14,30,31] can be developed to identify the highly correlated gene clusters. Then, representative genes from these clusters can be combined to produce good prediction results. We find that our MRMR approach is essentially consistent with the variable selection method in other papers.[14,30,31]

Acknowledgments

This work is mainly supported by US Department of Energy, Office of Science (MICS office and LBNL LDRD) under the contract DE-AC03-76SF00098. Hanchuan Peng is also supported by National Institutes of Health (NIH), under contract No. R01 GM70444-01. We thank the reviewers for the helpful suggestions.

References

1. Alizadeh AA *et al.*, Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling, *Nature* **403**:503–511, 2000.

2. Alon U, Barkai N, Notterman DA *et al.*, Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays, *PNAS* **96**:6745–6750, 1999.
3. Ben-Dor A, Bruhn L, Friedman N, Nachman I, Schummer M, Yakhini Z, Tissue classification with gene expression profiles, *J Comput Biol* **7**:559–584, 2000.
4. Cheng J, Greiner R, *Comparing Bayesian Network Classifiers*, UAI'99. 1999.
5. Cherkauer KJ, Shavlik JW, Protein structure prediction: Selecting salient features from large candidate pools, *ISMB 1993*, pp. 74–82, 1993.
6. Cox DR, *Analysis of Binary Data*, Methuen, London, 1970.
7. Ding C, Analysis of gene expression profiles: class discovery and leaf ordering, in *RECOMB 2002*, pp. 127–136, 2002.
8. Ding C, Dubchak I, Multi-class protein fold recognition using support vector machines and neural networks, *Bioinformatics* **17**:349–358, 2001.
9. Ding C, Peng HC, Minimum redundancy feature selection from microarray gene expression data, in *IEEE Computer Society Bioinformatics Conf 2003*, pp. 523–528, 2003.
10. Dudoit S, Fridlyand J, Speed T, Comparison of discrimination methods fro the classification of tumors using gene expression data, in *Tech Report 576*, Dept of Statistics, UC Berkeley, 2000.
11. Furey TS, Cristianini N, Duffy N, Bednarski D, Schummer M, Haussler D, Support vector machine classification and validation of cancer tissue samples using microarray expression data, *Bioinformatics* **16**:906–914, 2000.
12. Garber ME, Troyanskaya OG *et al.*, Diversity of gene expression in adenocarcinoma of the lung, *PNAS USA* **98**(24):13784–13789, 2001.
13. Golub TR, Slonim DK *et al.*, Molecular classification of cancer: class discovery and class prediction by gene expression monitoring, *Science* **286**:531–537, 1999.
14. Herskovits E, Peng HC, Davatzikos C, A Bayesian morphometry algorithm, *IEEE Transactions on Medical Imaging* **24**(6):723–737, 2004.
15. Hsu CW, Lin CJ, A comparison of methods for multi-class support vector machines, *IEEE Trans Neural Networks* **13**:415–425, 2002.
16. Jaakkola T, Diekhans M, Haussler D, Using the Fisher kernel method to detect remote protein homologies, in *ISMB'99*, pp. 149–158, 1999.
17. Jaeger J, Sengupta R, Ruzzo WL, Improved gene selection for classification of microarrays, in *PSB'2003*, pp. 53–64, 2003.
18. Kohavi R, John G, Wrapper for feature subset selection, *Artificial Intelligence* **97**(1–2):273–324, 1997.
19. Koller D, Sahami M, Toward optimal feature selection, in *ICML'96*, pp. 284–292, 1996.
20. Langley P, Selection of relevant features in machine learning, in *AAAI Fall Symposium on Relevance*, 1994.
21. Lee Y, Lee CK, Classification of multiple cancer types by multicategory support vector machines using gene expression data, *Bioinformatics* **19**:1132–1139, 2003.
22. Li J, Liu H, Downing JR, Yeoh A, Wong L, Simple rules underlying gene expression profiles of more than six subtypes of acute lymphoblastic leukemia (ALL) patients, *Bioinformatics* **19**:71–78, 2003.
23. Li W, Yang Y, How many genes are needed for a discriminant microarray data analysis?, in *Critical Assessment of Techniques for Microarray Data Mining Workshop*, pp. 137–150, 2000.
24. Mitchell T, *Machine Learning*, McGraw-Hill, 1997.
25. Model F, Adorján P, Olek A, Piepenbrock C, Feature selection for DNA methylation based cancer classification, *Bioinformatics* **17**:S157–S164, 2001.

26. Nguyen DV, Rocke DM, Tumor classification by partial least squares using microarray gene expression data, *Bioinformatics* **18**:39–50, 2002.
27. Nguyen DV, Rocke DM, Multi-class cancer classification via partial least squares with gene expression profiles, *Bioinformatics* **18**:1216–1226, 2002.
28. Ooi CH, Tan P, Genetic algorithms applied to multi-class prediction for the analysis of gene expression data, *Bioinformatics* **19**:37–44, 2003.
29. Park PJ, Pagano M, Bonetti M, A nonparametric scoring algorithm for identifying informative genes from microarray data, in *6th PSB*, pp. 52–63, 2001.
30. Peng HC, Long FH, A Bayesian learning algorithm of discrete variables for automatically mining irregular features of pattern images, in *Proc of Second International Workshop on Multimedia Data Mining (MDM/KDD'2001), in conjunction with ACM SIG/KDD2001*, San Francisco, CA, USA, pp. 87–93, 2001.
31. Peng HC, Herskovits E, Davatzikos C, Bayesian clustering methods for morphological analysis of MR images, in *Proc of 2002 IEEE Int Symposium on Biomedical Imaging: From Nano to Macro*, pp. 485–488, Washington, DC, USA, July, 2002.
32. Ross DT, Scherf U *et al.*, Systematic variation in gene expression patterns in human cancer cell lines, *Nat Genet* **24**(3):227–234, 2000.
33. Scherf U, Ross DT *et al.*, A cDNA microarray gene expression database for the molecular pharmacology of cancer, *Nat Genet* **24**(3):236–244, 2000.
34. Thomas JG, Olson JM, Stephen J *et al.*, An efficient and robust statistical modeling approach to discover differentially expressed genes using genomic expression profiles, *Genome Res* **11**:1227–1236, 2001.
35. Vapnik V, *The Nature of Statistical Learning Theory*, New York: Springer, 1995.
36. Welsh JB, Zarrinkar PP *et al.*, Analysis of gene expression profiles in normal and neoplastic ovarian tissue samples identifies candidate molecular markers of epithelial ovarian cancer, *PNAS USA* **98**:1176–1181, 2001.
37. Weston J, Watkins C, Multi-class support vector machines, in *ESANN'99*, Brussels, 1999.
38. Xing EP, Jordan MI, Karp RM, Feature selection for high-dimensional genomic microarray data, *ICML2001*.
39. Xiong M, Fang Z, Zhao J, Biomarker identification by feature wrappers, *Genome Res* **11**:1878–1887, 2001.
40. Yeoh A, ..., Wong L, Downing J, Classification, subtype discovery, and prediction of outcome in pediatric acute lymphoblastic leukemia by gene expression profiling, *Cancer Cell* **1**:133–143, 2002.

Chris Ding is a staff computer scientist at Berkeley Lab. He obtained a Ph.D. from Columbia University and worked in California Institute of Technology and Jet Propulsion Laboratory. He started work on biomolecule simulations in 1992 and computational genomics research in 1998. He is the first to use Support Vector Machines for protein 3D structure prediction. He's written 10 bioinformatics papers and also published extensively on machine learning, data mining, text and Web analysis. He's given invited seminars at Stanford, Carnegie Mellon, UC Berkeley, UC Davis, and many conferences, workshops and panel discussions. He has given a tutorial on Bioinformatics and Data Mining in ICDM'03 and a tutorial on Spectral

Clustering in ICML'04. He is on the program committee in ICDM'03,'04, SDM'04. He is a referee on many journals and also served on NSF panels. More details: http://crd.lbl.gov/∼cding.

Hanchuan Peng is a scientist with the Lawrence Berkeley National Laboratory, University of California at Berkeley. His research interests include image data mining, bioinformatics and medical informatics, pattern recognition and computer vision, signal processing, artificial intelligence and machine learning, biocomputing. His most recent work focuses on computational reconstruction of gene regulatory networks based on analysis of *in situ* gene expression patterns. He is a member of the Berkeley Drosophila Transcriptional Network Project. He has published about 50 peer-reviewed research papers and won several awards including the champion of national computer software competition in 1997, co-awarded by Ministry of Education, China, and Ministry of Electronic Industry, China. He received a Ph.D. degree in Biomedical Engineering from Southeast University, China, in 1999. His web pages are http://miracle.lbl.gov and http://www.hpeng.net.