Minimum Redundancy Feature Selection from Microarray Gene Expression Data

Chris Ding and **Hanchuan Peng**
NERSC Division, Lawrence Berkeley National Laboratory,
University of California, Berkeley, CA, 94720, USA
Email: chqding@lbl.gov, hpeng@lbl.gov

Abstract

Selecting a small subset of genes 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. Feature sets obtained through the minimum redundancy – maximum relevance framework represent broader spectrum of characteristics of phenotypes than those obtained through standard ranking methods; they are more robust, generalize well to unseen data, and lead to significantly improved classifications in extensive experiments on 5 gene expressions data sets.

1. Introduction

Discriminant analysis is now widely used in bioinformatics, such as distinguishing cancer tissues from normal tissues [2] or one cancer subtype vs another [1], predicting protein fold or super-family from its sequence [7][12], 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 selections: (1) dimension reduction to reduce the computational cost; (2) reduction of noises to improve the classification accuracy; (3) 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 or so such informative genes are usually sufficient [10]. There are even studies which suggest that a few (1 or 2) genes are sufficient [17][28]. Thus computation is reduced while 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 [14][16]. 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 powers with regard to the targeted classes. Simple methods based on mutual information [4], statistical tests (*t*-test, *F*-test) have been shown to be effective [10][6][8][19][25]. More sophisticated methods were also developed [15][3]. They also have the virtue of being easily and very efficiently computed. In filters, 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 very small number of non-redundant features [14][5], which gives high 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 current filter type methods is to simply select the top-ranked genes, say the top 50 [10]. More sophisticated regression models or tests along this line were also developed [23][20][27]. 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 how to more objectively determine m based on certain assumptions on data distributions [17]. A deficiency of this simple ranking approach is that the features could be correlated among themselves [13]. 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. In a number of studies [17][28], it is frequently observed 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 targeted 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 targeted phenotypes, but these could still be narrow regions of

IEEE COMPUTER SOCIETY

the relevant space covering the targeted phenotypes. Thus the generalization ability of the feature set could be limited.

Based on these observations, we propose to expand the space covered by the feature set by requiring that features are maximally dissimilar to each other, for example, their mutual Euclidean distances are maximized, or their pairwise correlations are minimized. These minimum redundancy criteria are of course supplemented by the usual maximum relevance criteria such as maximal mutual information with the targeted phenotypes. We therefore call this approach minimum redundancy – maximum relevance ("MRMR" for short). 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 targeted phenotypes, therefore leading to better generalization property; (2) equivalently, we can use a smaller MRMR feature set to effectively cover the same space that 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 reducing 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. Minimum Redundancy - Maximum Relevance 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 $p(y)$:

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

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 that we are seeking. The minimum redundancy condition is

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

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

To measure the level of discriminant powers of genes when they are differentially expressed for different targeted 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. Thus $I(h,g_i)$ 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:

$$\max V_I, \qquad V_I=\tfrac{1}{|S|}\sum_{i\in S} I(h,i), \qquad (3)$$

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

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

$$\max(V_I - W_I), \qquad (4)$$

$$\max(V_I / W_I). \qquad (5)$$

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 O($N^{|S|}$) search to obtain (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.

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

TYPE	ACRONYM	FULL NAME	FORMULA
DISCRETE	MID	Mutual information difference	$\max_{i\in\Omega_S}[I(i,h)-\frac{1}{\lvert S\rvert}\sum_{j\in S} I(i,j)]$
	MIQ	Mutual information quotient	$\max_{i\in\Omega_S}\{I(i,h)/[\frac{1}{\lvert S\rvert}\sum_{j\in S} I(i,j)]\}$
CONTINUOUS	FCD	*F*-test correlation difference	$\max_{i\in\Omega_S}[F(i,h)-\frac{1}{\lvert S\rvert}\sum_{j\in S} \lvert c(i,j)\rvert]$
	FCQ	*F*-test correlation quotient	$\max_{i\in\Omega_S}\{F(i,h)/[\frac{1}{\lvert S\rvert}\sum_{j\in S} \lvert c(i,j)\rvert]\}$
	FDM	*F*-test distance multiplicative	$\max_{i\in\Omega_S}[F(i,h)\cdot\frac{1}{\lvert S\rvert}\sum_{j\in S} d(i,j)]$
	FSQ	*F*-test similarity quotient	$\max_{i\in\Omega_S}\{F(i,h)/[\frac{1}{\lvert S\rvert}\sum_{j\in S} \frac{1}{d(i,j)}]\}$

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 we already select m features (genes) for the set S, 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:

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

$$\min_{i\in\Omega_S} \tfrac{1}{|S|}\sum_{j\in S} I(i,j)\cdot \qquad (7)$$

The condition in Eq. (6) is equivalent to the condition in Eq. (3), while Eq. (7) is an approximation of the condition of Eq. (2). The two combinations of Eqs. (4) and (5) for relevance and redundancy 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*|·*N*) complexity.

3.2. Minimum Redundancy - Maximum Relevance 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 [6][8]:

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

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 *k*th class, *K* is the number of classes, and $\sigma^2 = [\sum_k (n_k-1)\sigma_k^2]\Big/(n-K)$ is the pooled variance (where n_k and σ_k are the size and the variance of the *k*th 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:

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

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

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

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 $d(i,j) = d(g_i,g_j)$ (we choose the L_1 distance in this paper). The minimum redundancy condition can be specified as

$$\max W_d, \quad W_d = \frac{1}{|S|^2}\sum_{i,j\in S} d(i,j). \qquad (11)$$

Furthermore, instead of using "dissimilarity" or distance, we may use "similarity" or inverse distance to measure redundancy. The minimum redundancy condition is

$$\min W_s, \quad W_s = \frac{1}{|S|^2}\sum_{i,j\in S} \frac{1}{d(i,j)}. \qquad (12)$$

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

(1) FCD: combine *F*-test with correlation using difference, $\max(V_F - W_c)$;

(2) FCQ: combine *F*-test with correlation using quotient, $\max(V_F / W_c)$.

(3) FDM: combine *F*-test with distance using multiplication, $\max(V_F \cdot W_d)$;

(4) FSQ: combine *F*-test with similarity using quotient, $\max(V_F / W_s)$.

We use the same linear incremental search algorithm as in the discrete variable case in §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) [18] is one of the oldest classifiers. It is obtained by using the Bayes rule and assuming features (variables) are independent of each other given the targeted classes. Given a tissue sample *s* with 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

$$p(h_k \mid s) \propto \prod_{i\in S} p(g_i \mid h_k), \qquad (13)$$

where $p(g_i|h_k)$ are conditional tables (or conditional density) learned in training using examples. Despite the independence assumption, NB has been shown to have very good classification performance for many real data sets, especially for documents [18], on par with many more sophisticated classifiers.

4.2. Support Vector Machine (SVM)

SVM is a relatively new and promising classification method [24]. 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 [7]. Another approach is to directly construct a multi-class SVM [11][26]. In this paper, we used the Matlab version of LIBSVM [11].

5. Experiments

5.1. Data Sets

To evaluate the usefulness of the MRMR approach, we carried out experiments on fives data sets of gene expression profiles. Two expression data sets popularly used in research literature are the leukemia data of Golub et al [10] and the colon cancer data of Alon et al [2]. As listed in Table 2, both leukemia and colon cancer data sets have two classes. The colon dataset contains normal tissue samples 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.

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 [21][22], lung cancer [9] and lymphoma [1]. The details of these data sets are summarized in Table 3. 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 five data sets provide a comprehensive test suit.

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

Table 3. Multi-class datasets used in our experiments

DATASET	NCI		LUNG CANCER		LYMPHOMA	
SOURCE	Ross et al (2000) Scherf et al (2000)		Garber et al (2001)		Alizadeh et al (2000)	
# GENE	9703		918		4026	
# SAMPLE (#S)	60		73		96	
# CLASS	9		7		9	
CLASS	CLASS NAME	# S	CLASS NAME	# S	CLASS NAME	# S
C1	NSCLC	9	AC-group-1	21	Diffuse large B cell lymphoma	46
C2	Renal	9	Squamous	16	Chronic Lympho. leukemia	11
C3	Breast	8	AC-group-3	13	Activated blood B	10
C4	Melanoma	8	AC-group-2	7	Follicular lymphoma	9
C5	Colon	7	Normal	6	Resting/ activated T	6
C6	Leukemia	6	Small-cell	5	Transformed cell lines	6
C7	Ovarian	6	Large-cell	5	Resting blood B	4
C8	CNS	5			Germinal center B	2
C9	Prostate	2			Lymph node/tonsil	2

5.2. Assessment Measure

We assess 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 errors in LOOCV in figures and tables.

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. To assess the MRMR features, we used up to 60 genes for NB and up to 100 genes for SVM.

5.3. Results for Discrete Features

We thresholded the observations of each gene expression variable using σ (standard deviation) and μ (mean): any data larger than μ+σ/2 were transformed to state 1; any data between μ–σ/2 and μ+σ/2 were transformed to state 0; any data smaller than μ–σ/2 were transformed to state -1. These three states correspond to the over-expression, baseline, and under-expression of genes. We applied the feature selection methods and performed LOOCV using NB on the 5 datasets. A summary of the results is shown in Table 4, where due to the space limitation we only list results of *m*=3,6,…,60.

For NCI data, with 36 MRMR MIQ features, we attained 1 LOOCV error, or (60–1)/60=98.3% accuracy. In the baseline feature sets, the best case has 10 errors.

For lung cancer dataset, the MRMR feature set also outperformed the baseline substantially. The best MRMR MIQ features leads to 2 errors or (73–2)/73 =97.3% accuracy. The best baseline result is 8 errors.

For lymphoma data, MRMR features out-performed baseline features significantly. The best MIQ feature set leads to 3 LOOCV errors, or an accuracy (96–3)/96 = 96.9%. In contrast, the best baseline feature set leads to 17 errors.

For leukemia data, MRMR performed well for all feature selection methods, with zero LOOCV errors.

For colon data, MRMR features also show clear improvements over baseline features. For example, a 9-gene MRMR MIQ feature set has LOOCV error 4, in contrast to the LOOCV error of 7 in the best case of baseline features.

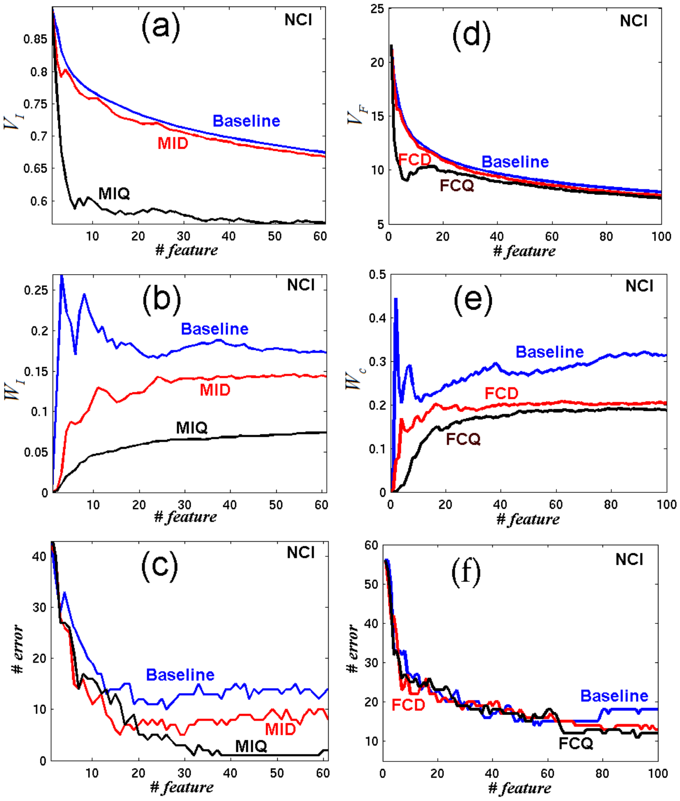

Figure 1. **(a)~(c) for NB**: (a) Relevance V_I, (b) redundancy W_I and (c) LOOCV error using NB for the case of discrete variables. **(d)~(f) for SVM**: (d) relevance V_F, (e) redundancy W_c and (f) the LOOCV error using SVM for the case of continuous variables. Dataset: NCI.

In summary, for multi-class problems with 7~9 classes, MRMR feature sets lead to LOOCV error rates of 2~3%. In contrast, error rates of baseline features are 11~18%. This demonstrates the effectiveness of MRMR feature selection over the baseline method, and NB is an accurate classification method. We emphasize that the features selected according to mutual information in MRMR are independent of NB, and thus do not directly aim at producing the best results in NB. We expect these MRMR feature sets will produce similar good results using different classification methods.

Table 4. LOOCV errors using Naïve Bayes for 5 datasets.

Dataset	m Method	3	6	9	12	15	18	21	24	27	30	36	42	48	54	60
NCI	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
Lung	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
Lymphoma	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
Leukemia	Baseline	1	0	1	0	1	2	2	2	1	1	1	3	3	2	3
	MID	1	0	0	0	0	0	0	1	1	1	1	2	1	1	1
	MIQ	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Colon	Baseline	10	7	8	8	8	9	8	9	8	9	9	9	9	10	9
	MID	8	10	7	7	7	7	8	8	7	7	7	7	7	7	7
	MIQ	12	6	4	5	7	7	7	7	8	8	7	7	7	7	7

Table 5. LOOCV errors for continuous multi-class data using SVM. The sign test statistics are explained in the text.

Dataset	m Method	5	10	20	30	40	50	60	70	80	90	100	All 100 features			
													+	=	–	R
NCI	Baseline	33	27	23	20	17	16	15	15	18	18	18	--	--	--	--
	FCD	37	22	20	19	19	18	16	14	13	14	13	53	20	27	0.26
	FCQ	33	25	23	19	18	16	18	12	12	12	12	61	11	28	0.33
	FDM	33	26	22	19	16	16	14	14	14	14	14	66	22	12	0.54
	FSQ	28	21	20	17	17	14	14	14	14	14	13	79	17	4	0.75
Lung	Baseline	25	18	9	8	9	9	8	7	8	8	8	--	--	--	--
	FCD	15	11	7	7	6	6	7	7	5	6	8	93	6	1	0.92
	FCQ	19	11	7	7	5	6	7	6	5	6	8	90	7	3	0.87
	FDM	26	15	10	8	9	10	8	9	6	6	8	41	34	25	0.16
	FSQ	18	16	11	7	7	9	9	7	6	7	7	57	27	16	0.41
Lymphoma	Baseline	26	16	13	6	7	5	7	6	6	7	6	--	--	--	--
	FCD	21	11	9	8	6	4	5	4	6	6	5	72	24	4	0.68
	FCQ	25	6	9	7	7	6	4	3	3	2	2	80	8	12	0.68
	FDM	16	10	9	8	8	5	6	5	6	6	5	69	15	16	0.53
	FSQ	18	11	7	9	7	6	6	6	6	6	5	65	21	14	0.51

Table 6. LOOCV errors for continuous 2-class datasets using SVM. The sign test statistics are explained in the text.

Dataset	m Method	2	4	6	10	20	30	40	50	First 20 features				All 50 features			
										+	=	–	R	+	=	–	R
Leukemia	Baseline	3	2	3	2	3	3	4	1	--	--	--	--	--	--	--	--
	TCD	3	3	3	2	5	1	1	1	3	4	3	0	12	10	3	0.36
	TCQ	3	2	1	0	1	1	1	1	8	2	0	0.8	18	3	4	0.56
	TDM	3	4	3	3	4	2	2	1	2	3	5	-0.3	6	12	7	-0.04
	TSQ	3	4	3	4	1	2	1	1	5	3	2	0.3	13	10	2	0.44
Colon	Baseline	10	11	9	10	13	10	9	8	--	--	--	--	--	--	--	--
	TCD	10	7	7	8	8	8	13	14	8	2	0	0.8	12	3	10	0.08
	TCQ	10	8	7	10	5	13	12	15	6	3	1	0.5	9	3	13	-0.16
	TDM	10	9	10	9	12	10	10	11	6	3	1	0.5	9	7	9	0
	TSQ	10	7	7	9	11	7	12	13	9	1	0	0.9	15	5	5	0.40

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)) and the LOOCV error, as plotted in Fig. 1 (a)~(c). In Fig.1, the relevance reduces for MID and MIQ, but the respective redundancy also reduces considerably. This is most clearly seen for MIQ. The fact that the MIQ feature set is the most effective illustrates the importance of reducing redundancy, the central theme of this research.

5.4. Results for Continuous Features

We directly classified the continuous features using the SVM classifier. We pre-processed the data so each gene has zero mean value and unit variance. The feature selection methods were applied; based on the obtained feature sets, SVM was trained and LOOCV was performed. Table 5 lists the LOOCV results of the 3 multi-class problems. The relevance, redundancy and LOOCV error results for the NCI data are also plotted in Fig.1 (d)~(f). A quick look at Table 5 indicates the improvement of MRMR feature set over baseline is not as pronounced as that for the discrete cases in Table 4. However, the improvement is still visible, consistent and statistically significant. For NCI data, one can see FCQ results are consistently better than baseline results. FCQ features achieve the best error rate of 12/62, vs the best error rate of 15/62 for baseline features.

To compare the results in a statistically consistent way, we did a sign test based on classification results with the feature set size m=1,2,…,100 (only a limited number of results are shown in Table 5). If MRMR features caused less (equal, or more) errors than the baseline features, we gave it "+" ("=", or "–"). Thus for FCQ features, for 61 different feature set sizes, FCQ is better than baseline; for 11 feature set sizes, FCQ is equally good as baseline; and for 28 feature set sizes, FCQ results are worse than baseline. Therefore, we say FCQ is better than the baseline with a confidence of R=(61–28)/100=0.33. These sign statistics show that MRMR features are better than the baseline significantly and consistently.

For lung cancer data, MRMR features also consistently give lower errors than the baseline features, as the sign tests show. The FCQ feature set of 40 genes achieves an error rate of 5/73, in contrast to the best error rate of 7/73 of baseline features with 70 genes.

For lymphoma data, MRMR features also consistently give lower errors than the baseline features. FCQ features achieve an error rate of 2/96, in contrast to the best error rate of 5/96 of baseline.

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. Table 6 lists the results of LOOCV for these 2-class datasets. We see that MRMR TCQ features improve classification, especially at small number of features. The sign tests for the first 20 features (i.e. 10 pairs of two-sided features) show 6 out of the 8 MRMR feature selection methods outperform the baseline feature selection. Note that in the original paper, Golub et al [10] used a prediction strength feature selection that is close to the two-sided *t*-test. Using their feature set, SVM gives a LOOCV error of 4, whereas our feature sets with 50 genes (i.e. 25 pairs) lead to only 1 error.

6. Discussions

Experiment results suggest that MRMR features are more effective for discrete variables with smaller number of features than for continues variables with larger number of features. This can be seen in the following two observations: (1) Fig. 1 show that the reduction of redundancy in MRMR feature sets is more pronounced for discrete variables than for continuous variables. (2) The effectiveness of MRMR is more pronounced in the region of small feature set sizes. If we use the feature sets of 1000 genes, the difference between the MRMR approach and the baseline approach will not be large. For gene selection, small feature set is of practical importance.

Our extensive tests, as shown in Tables 4 ~ 6, also show that discretization of the gene expressions leads to clearly better classification accuracy than the original continuous data.

Acknowledgements

This work is supported by US Department of Energy, Office of Science (MICS office and LBNL LDRD) under the contract DE-AC03-76SF00098. This paper is a condensed version of a LBNL Tech Report 51985, Jan 6, 2003, submitted to ISMB 2003. We thank a referee for pointing out Ref. [13].

References

[1] Alizadeh, A.A., et al.. (2000). Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling, *Nature*, *403*, 503-511.

[2] Alon, U., Barkai, N., Notterman, D.A., et al. (1999). Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays, *PNAS USA*, *96*, 6745-6750.

[3] Ben-Dor, A., Bruhn, L., Friedman, N., Nachman, I., Schummer, M., & Yakhini, Z. (2000), Tissue classification with gene expression profiles, *J Comput Biol*, *7*, 559--584.

[4] Cheng, J., & Greiner, R. (1999). Comparing Bayesian network classifiers, *UAI'99*.

[5] Cherkauer, K.J., & J. W. Shavlik, J.W. (1993). Protein structure prediction: selecting salient features from large candidate pools, *1st ISMB*, 74-82.

[6] Ding, C. (2002). Analysis of gene expression profiles: class discovery and leaf ordering, *RECOMB 2002*, 127-136.

[7] Ding, C., & Dubchak, I. (2001). Multi-class protein fold recognition using support vector machines and neural networks, *Bioinformatics*, *17*, 349-358.

[8] Dudoit, S., Fridlyand, J., & Speed, T. (2000). Comparison of discrimination methods fro the classification of tumors using gene expression data, *Tech Report 576*, Dept of Statistics, UC Berkeley.

[9] Garber, M.E., Troyanskaya, O.G., et al. (2001). Diversity of gene expression in adenocarcinoma of the lung, *PNAS USA*, *98*(*24*), 13784-13789.

[10] Golub, T.R., Slonim, D.K. et al, (1999). Molecular classification of cancer: class discovery and class prediction by gene expression monitoring, *Science*, *286*, 531-537.

[11] Hsu, C.W., & Lin, C.J. (2002). A comparison of methods for multi-class support vector machines, *IEEE Trans. on Neural Networks*, *13*, 415-425.

[12] Jaakkola, T., Diekhans, M., & Haussler, D. (1999). Using the Fisher kernel method to detect remote protein homologies, *ISMB'99*, 149-158.

[13] Jaeger, J., Sengupta, R., Ruzzo, W.L. (2003) Improved Gene Selection for Classification of Microarrays , *PSB'2003*, pp.53-64.

[14] Kohavi, R., & John, G. (1997). Wrapper for feature subset selection, *Artificial Intelligence*, *97*(*1-2*), 273-324.

[15] Koller D., & Sahami, M. (1996). Toward optimal feature selection, *ICML'96*, 284-292.

[16] Langley, P. (1994). Selection of relevant features in machine learning, *AAAI Fall Symposium on Relevance*.

[17] Li, W., & Yang Y. (2000). How many genes are needed for a discriminant microarray data analysis?, *Critical Assessment of Techniques for Microarray Data Mining Workshop*, 137-150.

[18] Mitchell, T., (1997). *Machine Learning*, McGraw-Hill.

[19] Model, F., Adorján, P., Olek, A., & Piepenbrock, C. (2001). Feature selection for DNA methylation based cancer classification, *Bioinformatics*, 17, S157-S164.

[20] Park, P.J., Pagano, M, & Bonetti, M. (2001). A nonparametric scoring algorithm for identifying informative genes from microarray data, *6th PSB*, 52-63.

[21] Ross, D.T., Scherf, U., et al. (2000). Systematic variation in gene expression patterns in human cancer cell lines, *Nature Genetics*, *24*(*3*), 227-234.

[22] Scherf, U., Ross, D.T., et al. (2000). A cDNA microarray gene expression database for the molecular pharmacology of cancer, *Nature Genetics*, 24(3), 236-244.

[23] Thomas, J.G., Olson, J.M., Stephen J., et al. (2001). An efficient and robust statistical modeling approach to discover differentially expressed genes using genomic expression profiles, *Genome Research*, *11*, 1227-1236.

[24] Vapnik, V. (1995). *The Nature of Statistical Learning Theory*. New York: Springer.

[25] Welsh, J.B., Zarrinkar, P.P., et al. (2001). 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.

[26] Weston, J., & Watkins, C. (1999). Multi-class support vector machines, *ESANN'99*, Brussels.

[27] Xing, E.P., Jordan, M.I., & Karp, R.M. (2001). Feature selection for high-dimensional genomic microarray data, *ICML2001*.

[28] Xiong, M., Fang, Z., & Zhao, J. (2001). Biomarker identification by feature wrappers, *Genome Research*, *11*, 1878-1887.