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

Abstract

Motivation. 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.

Results. 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 5 gene expression data sets: NCI, Lymphoma, Lung, 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 dataset are listed in http://www.nersc.gov/~cding/MRMR/.

Contact. chqding@lbl.gov, hpeng@lbl.gov

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][14], 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: (1) dimension reduction to reduce the computational cost; (2) reduction of noise 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 informative genes are usually sufficient [12]. There are studies suggestting that only a few genes are sufficient [20][35]. 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 [16][18]. 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 target classes. Simple methods based on mutual information [4], statistical tests (*t*-test, *F*-test) have been shown to be effective [12][6][9][22]. More sophisticated methods are also developed [17][3]. 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 [16] [5][20][35] , 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 [12]. More sophisticated regression models or tests along this line were also developed [29][26][34]. 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 [20]. A deficiency of this simple ranking approach is that the features could be correlated among themselves [15][8]. 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 [20][35] 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 pairwise 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 $p(y)$:

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

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 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 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:

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

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

The 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:

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

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

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 $O(N^{|S|})$ 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.

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	$\max_{i\in\Omega_S}\{F(i,h)/[\frac{1}{\lvert S\rvert}\sum_{j\in S}\lvert c(i,j)\rvert]\}$

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, 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:

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

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

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|·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 [6][9]:

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

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 $\sigma^2 = [\sum_k (n_k - 1)\sigma_k^2] \Big/ (n-K)$ 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:

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

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

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

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 [8], 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,

(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 §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) [21] 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

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

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 [21], on par with many more sophisticated classifiers.

4.2. Support Vector Machine (SVM)

SVM is a relatively new and promising classification method [30]. 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 [13][33]. In this paper, we used the Matlab version of LIBSVM [13].

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.

4.4. Logistic Regression (LR)

LR [31] 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 fives data sets of gene expression profiles. Two expression datasets popularly used in research literature are the Leukemia data of Golub et al [12] 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 # Sample	7070 72		2000 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 (*#S* is the number of samples)

Dataset	NCI		Lung cancer		Lymphoma	
Source	Ross et al (2000) Scherf et al (2000)		Garber et al (2001)		Alizadeh et al (2000)	
# Gene # S # Class	9703 60 9		918 73 7		4026 96 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

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 [27][28], Lung cancer [11] 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.

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

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 LOOCV errors in Tables 4 - 8.

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. First reason is noise reduction because the original readings contain substantial noise. 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 σ (standard deviation) and μ (mean): 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 smaller than $\mu-\sigma/2$ were transformed to state -1. These three states correspond to the over-expression, baseline, and under-expression of genes.

5.4. Results

We applied the MRMR feature selection methods on both continuous and descretized data. The top 60 MRMR genes for each of the 5 datasets are listed in http://www.nersc.gov/~cding/MRMR/. We performed LOOCV using NB, LDA, SVM and LR on all 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,…54,60 for multi-class datasets and

m=1,2,3,…,8,10,…,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 4. Lymphoma data (96 samples for 9 classes) LOOCV errors.

Classifier	Data Type	M 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.

Classifier	Data Type	M 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

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

Classifier	Data Type	M 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.

Classifier	Data Type	M 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.

Classifier	Data Type	M 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

(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.

We list the best performance of MRMR features together with the best baseline performance in Table 9. From this table, we can quantify the improvements due to MRMR feature selection. For the three multi-class datasets, the LOOCV errors are reduced by a factor of 10. 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 Fig. 1 (a) and (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 Fig.1 (c) and (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.

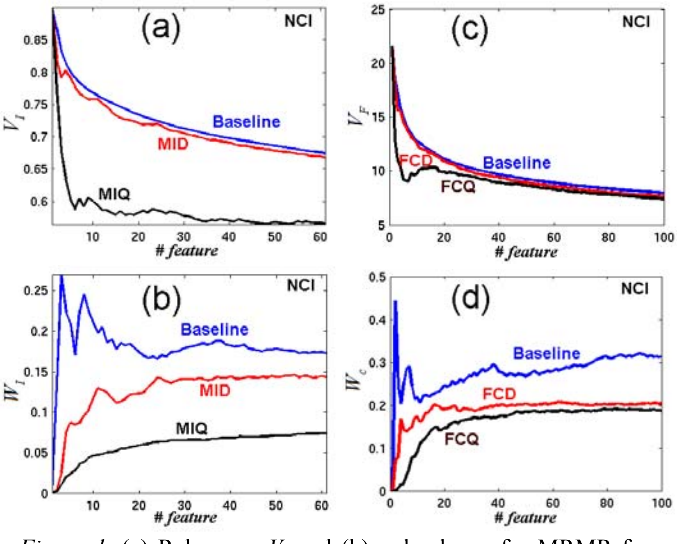

Figure 1. (a) Relevance V_I and (b) redundancy for MRMR features on discretized NCI dataset. (c) relevance V_F and (d) redundancy W_c on the continuous NCI dataset.

5.5 Comparison with Other Work

Results of similar class prediction on microarray gene expression data obtained by others are listed in Table 9. For NCI, our result of LOOCV error rate is 1.67% using NB, whereas Ooi & Tan [25] obtained 14.6% error rate. On the 5-class subset of NCI, Nguyen & Rocke [23] 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 & Rocke [23] obtained 2.4%; on the same 3 classes, our LDA results is 0% 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 [10] obtained 100% accuracy using SVM, and Lee & Lee [19] obtained 1.39% error rate.

For Colon data*, our result is 6.45% error rate, which is the same as Nguyen & Rocke [23] using PLS. The SVM result of [10] is 9.68%.

Table 9. Comparison of the best results (lowest error rates in percentage) of the baseline and MRMR features. Also listed are results in literature. [a] Ooi & Tan, using genetic algorithm [25]. [b] Nguyen and Rocke [23] used only 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 & Rocke used only 3 classes in lymphoma dataset and obtain 2.4% error rate. Using the same 3 classes, our NB lead to zero errors. [d] Furey et al, using SVM [10]. [e] Lee & Lee, using SVM [19]. [f] Nguyen & Rocke, using PLS [24].

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	--	
Leukemia	Baseline	0	1.39	1.39	1.39	0 [d]
	MRMR	0	0	0	0	1.39 [e]
Colon	Baseline	11.29	11.29	11.29	11.29	9.68 [d]
	MRMR	6.45	8.06	9.68	9.68	6.45 [f]

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 5 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 redun-

* Many classification studies have used Leukemia and Colon datasets. Due to space limitation, we only list two for each dataset in Table 9.

dancy 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 ~ 8, 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.

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*, *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, *ISMB 1993*, 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] Ding, C., & Peng, H.C., (2003). Minimum redundancy feature selection from microarray gene expression data, *IEEE Computer Society Bioinformatics Conf.*, 2003.

[9] 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.

[10] Furey,T.S., Cristianini,N., Duffy, N., Bednarski, D., Schummer, M., and Haussler, D. (2000). Support vector machine classification and validation of cancer tissue samples using microarray expression data, *Bioinformatics*,16, 906-914.

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

[12] 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.

[13] 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.

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

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

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

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

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

[19] Lee, Y., and Lee, C.K. (2003). Classification of multiple cancer types by multicategory support vector machines using gene expression data, *Bioinformatics*, 19, 1132-1139.

[20] 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.

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

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

[23] Nguyen, D.V., & Rocke, D.M. (2002). Multi-class cancer classification via partial least squares with gene expression profiles, *Bioinformatics*, 18, 1216-1226.

[24] Nguyen, D.V., & Rocke, D. M. (2002). Tumor classification by partial least squares using microarray gene expression data, *Bioinformatics*, 18, 39-50.

[25] Ooi, C. H., and Tan, P. (2003). Genetic algorithms applied to multi-class prediction for the analysis of gene expression data, *Bioinformatics,* 19, 37-44.

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

[27] 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.

[28] 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.

[29] 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.

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

[31] Cox,D.R., (1970) *Analysis of Binary Data*, Methuen, London.

[32] 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.

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

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

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