

Minimum Redundancy Feature Selection for Microarray Gene Expression Data

Chris Ding and Hanchuan Peng

NERSC Division, Lawrence Berkeley National Laboratory

University of California, Berkeley, CA, 94720, USA

chqding@lbl.gov, hpeng@lbl.gov

Keywords: feature selection, discriminant analysis, Naïve Bayes, SVM

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 class predictions in extensive experiments on 5 gene expression datasets.

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 [15][27]. Thus computation is reduced while accuracy is increased via effective feature selection. Of the small number selected genes, their biological relationship with the target diseases is more easily identified. These "marker" genes thus provide additional scientific understanding to the problem.

Selecting an effective and more representative feature set is the subject of this paper.

There are two general approaches of feature selection: filters and wrappers [13][14]. 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] [18] [24]. They also have the virtue of being easily and very efficiently computed. In filters, the biases in the feature selection are not correlated to the biases in the learning methods, therefore they have better generalization property, i.e., the selected features from training data generalize well to new data that are not used in training. 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 of very small number of features, which gives high accuracy, because the characteristics of the features match well with the characteristics of the learning method. Because of this strong interplay between features and the learning method, the features such selected could sometimes overfit the classifier model and not generalize well to unseen data [13]. Wrapper methods typically require extensive computation to search the best features.

In this paper, we will focus on the filter type methods, mainly because they are widely adopted in practice – they can be efficiently computed for large datasets, their characteristics are relatively clear, and their implementations are relatively simple.

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 [22][19][26]. 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 [15]. A deficiency of this simple ranking approach is that the features could be correlated among themselves. 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 [15][27], 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 the 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 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 the 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 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) approach. The benefits of this more complex 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 than the conventional approach; (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 this minimum redundancy in gene selection and provide a comprehensive study. To our knowledge, this work is the first study on this issue.

Note that feature selection was widely used in other areas such as information retrieval, image analysis, machine learning, etc. The minimum redundancy feature selection can be adopted in all these areas without modifications.

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 [5] based on their joint probabilistic distribution $p(x,y)$ and the respective marginal probabilities $p(x)$ and $p(y)$:

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

$I(x,y)$ attains the maximum when x and y are fully dependent, i.e. $p(x,y)=p(x)p(y)$, and reduces to the minimum (=0) when x and y are independent of each other, i.e. $p(x,y)=p(x)p(y)$.

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

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

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 $h = \{h_1, h_2, \dots, h_K\}$ (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 :

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

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 eqns. (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:

$$\max(V_i - W_i), \quad (4)$$

$$\max(V_i / W_i). \quad (5)$$

Our goal here is to see whether the MRMR approach is effective in its simplest form. 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.

In our algorithm, the first feature is selected according to eqn. (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:

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

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

The condition in eqn. (6) is equivalent to the condition in eqn. (3), while eqn. (7) is an approximation of the condition of eqn. (2). The two combinations of eqns. (4) and (5) for relevance and redundancy lead to the selection criteria of a new feature:

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

Type	Acronym	Full Name	Formula
Discrete	MID	Mutual information difference	$\max_{i \in \Omega_S} [I(i, h) - \frac{1}{ S } \sum_{j \in S} I(i, j)]$
	MIQ	Mutual information quotient	$\max_{i \in \Omega_S} \{I(i, h) / [\frac{1}{ S } \sum_{j \in S} I(i, j)]\}$
Continuous	FCD	F -test correlation difference	$\max_{i \in \Omega_S} [F(i, h) - \frac{1}{ S } \sum_{j \in S} c(i, j)]$
	FCQ	F -test correlation quotient	$\max_{i \in \Omega_S} \{F(i, h) / [\frac{1}{ S } \sum_{j \in S} c(i, j)]\}$
	FDM	F -test distance multiplicative	$\max_{i \in \Omega_S} [F(i, h) \cdot \frac{1}{ S } \sum_{j \in S} d(i, j)]$
	FSQ	F -test similarity quotient	$\max_{i \in \Omega_S} \{F(i, h) / [\frac{1}{ S } \sum_{j \in S} \frac{1}{d(i, j)}]\}$

(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 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]:

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

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 σ^2 is the pooled variance:

$$\sigma^2 = [\sum_k (n_k - 1) \sigma_k^2] / (n - K), \quad (9)$$

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 the following:

$$\max V_F, \quad V_F = \frac{1}{|S|} \sum_{i \in S} F(i, h). \quad (10)$$

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

$$\min W_c, \quad W_c = \frac{1}{|S|^2} \sum_{i, j} |c(i, j)|, \quad (11)$$

where we have assumed that both high positive and high negative correlation mean redundancy, and thus take the absolute value of the 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

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

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

$$\min W_s, \quad W_s = \frac{1}{|S|^2} \sum_{i, j \in S} \frac{1}{d(i, j)}. \quad (13)$$

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

$$p(h_k | s) \propto \prod_{i \in S} p(g_i | h_k), \quad (14)$$

where $p(g_i | h_k)$ are conditional tables learned in the training using examples. Despite the independence assumption, NB has been shown to have very good classification performance for many real datasets, especially for documents [17], on par with many more sophisticated classifiers. For implementation simplicity, feature values need to be discretized to use NB.

4.2 Support vector machine (SVM)

SVM is a relatively new and promising classification method [23][3]. 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.

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 con-

struct a multi-class SVM [11] [25]. In this paper, we use the latter approach.

5. Experiments

5.1 Datasets

To evaluate the usefulness of the MRMR approach, we carried out experiments on five datasets of gene expression profiles. Two expression datasets 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 datasets 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

Table 2. Two-class datasets used in our experiments

Dataset	Leukemia		Colon cancer	
Source	Golub et al [10]		Alon et al [2]	
# 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 [20] Scherf et al [21]		Garber et al [9]		Alizadeh et al [1]	
# 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 lymphocytic 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

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 datasets: NCI [20] [21], lung cancer [9] and lymphoma [1]. The details of these datasets 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 datasets provide a comprehensive test suit.

5.2 Assessment measure

There are two types of accuracy in classification experiments. One is the training accuracy, also called in-sample accuracy, in which the samples tested against the classifier are also used in training the classifier. Another is the Cross Validation (CV) accuracy, in which the samples tested against the classifier are not used during the training of the classifier. All samples in the dataset take turns to be left out of the training and are tested against such trained classifiers. A popular CV method is the "Leave-One-Out CV" (LOOCV). CV accuracy provides more realistic assessment of classifiers since the classifiers have not seen the test samples. Training accuracy does not generalize well to unseen data samples because the classifier is tested against only those used in training. It is not uncommon that a classifier has good training accuracy but poor CV accuracy. In this paper, we used the LOOCV results throughout. For presentation purpose, we used 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.

5.3 Feature set sizes

In gene expression analysis, especially in clinical tests and diagnosis, the number of marker genes should be small for practical reasons. Golub et al [10] used 50 genes, while others suggested much smaller numbers [27][15]. To assess the MRMR features, we used up to 60 genes for NB and up to 100 genes for SVM.

The classification results change for different feature sizes even for a fixed feature selection method. Thus fixing feature size to a single number and then com-

paring different selection methods is not adequate or reliable. For this reason, we let feature size m take multiple values, $m=1,2,3,\dots,60$ for NB, and $m=1,2,\dots,100$ for SVM. For each feature set size, we selected the features according to various criteria (as listed in Table 1), trained the classifiers, and did the LOOCV. This was repeated for all five datasets. The LOOCV errors are tabulated in Tables 4, 6, and 7, and are also shown in Fig. 1. With these extensive experiments, the comparisons between different feature selection methods are comprehensive and reliable.

Another reason for these comprehensive experiments is to search for the optimal feature set for each of the gene expression datasets. For example, for NCI data we obtained a 36-gene set that leads to 98% LOOCV accuracy for the diagnosis problem of 9 different types of cancer. For leukemia data, we obtained a 6-gene set with the zero LOOCV error. These optimal feature sets are useful for research comparisons and are of biological interests on related problems.

5.4 Results for discrete features

Although gene expression values are continuous, we may discretize them into discrete states. Discretization

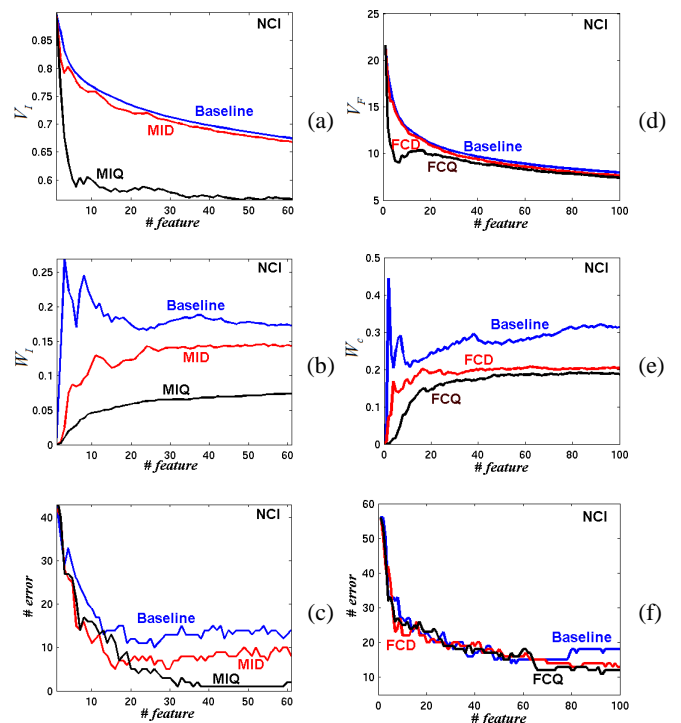


Figure 1. (a)~(c) for NB: (a) Relevance V_I , (b) redundancy W_I and (c) LOOCV error using NB for the case of discrete variables. (d)~(f) for SVM: (d) relevance V_F , (e) redundancy W_C and (f) the LOOCV error using SVM for the case of continuous variables. Dataset: NCI.

is necessary to utilize the NB classifiers. For simplicity and ease to replicate the experiments, in this paper we used a simple method to threshold 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. (For a pure Gaussian distribution, the percentages of these three regions are about 31%, 38%, and 31%). These three states correspond to the over-expression, baseline, and under-expression of genes. Discretization into a few categories also reduces noises incurred in the experiments.

We applied the feature selection methods and performed LOOCV using NB on the 5 datasets. A sum-

mary of the results is shown in Table 4, where due to the space limitation we only list results of $m=3,6,\dots,60$.

For NCI data, with 36 MRMR MIQ features, we attained 1 LOOCV error, or $(60-1)/60=98.3\%$ classification 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 feature sets outperformed baseline features significantly. The best MIQ feature set leads to 3 LOOCV errors, or an accuracy of $(96-3)/96=96.9\%$. In contrast, the best base-

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
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. MRMR MIQ feature sets. For Leukemia, the 6-gene set leads to 100% LOOCV accuracy. Note that only the gene U46499 is present in the 50-gene set of Golub et al [10]. For Colon cancer dataset, the 9-gene set leads to $(62-4)/62=93.6\%$ LOOCV accuracy. MIQ feature sets for other datasets are listed in www.nersc.gov/~cding/MRMR.

Dataset	Probe number	Gene Description
Leukemia	U46499	Glutathione S-Transferase, Microsomal
	X66867_cds1	H.Sapiens max gene
	J04182	Lysosome-Associated Membrane Protein 1
	U01062	Inositol 1,4,5-Triphosphate Receptor, Type 3
	U82759	Homeodomain Protein HoxA9 mRNA
	HG2846-HT2983	Dihydrofolate Reductase, Alt. Splice 6
Colon	M76378	Gene 1 Human CRP Gene
	M84739	Gene 1 Autoantigen Calreticulin mRNA
	Y00345	Gene 1 mRNA For Poly Binding Protein
	R52081	3' Utr Transcriptional Activator Gen5
	T60155	3' Utr Actin, Aortic Smooth Muscle (Human)
	M26383	Gene 1 Human MONAP mRNA
	R54097	3' Utr Translational Initiation Factor
	R39209	3' Utr Human Immunodeficiency Virus
	H77348	3' Utr 5-Lipoxygenase Activating Protein

line feature set leads to 17 errors.

In summary, for multi-class problems with 7~9 classes, MRMR feature sets lead to LOOCV error rates of 2~3%. In contrast, the error rates of baseline features are 11~18%.

For leukemia dataset, NB performed surprisingly well for all feature selection methods. We got zero errors out of the 72 tissue samples in quite a number of feature sets. But still, the MRMR MIQ feature sets lead to perfect classification consistently. The MRMR MIQ 6-gene feature set is listed in Table 5. For colon cancer data, MRMR features show clear improvements over baseline features. For the 9-gene MRMR MIQ feature set, the LOOCV error is 4, in contrast to the LOOCV error of 7 in the best case of baseline features.

These results clearly demonstrate the effectiveness of MRMR feature selection method 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.

To better understand the effectiveness of the MRMR approach, we calculated the average relevance V_l and average redundancy W_l (see eqns.(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.

5.5. 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 6 lists the LOOCV results of

Table 6. LOOCV errors for continuous multi-class data using SVM. 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 7. LOOCV errors for continuous 2-class datasets using SVM.

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

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 6 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 and FSQ 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 $m=1,2,\dots,100$ (only a limited number of results are shown in Table 6). If MRMR features caused less (equal, or more) errors than the baseline features, we gave it "+" ("=", or "-"). Thus for FSQ features, for 79 different feature sizes, FSQ is better than baseline; for 17 feature sizes, FSQ result are equally good as baseline; and for 4 feature sizes, FSQ results are worse than baseline. Therefore, we say FSQ is better than the baseline with a confidence of $R = (79-4)/100 = 0.75$. 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. All 4 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), TCQ (vs FCQ), TDM (vs FDM) and TSQ (vs FSQ) schemes. Table 7 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 fea-

tures) 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 and summary

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 for discrete variables and SVM for continuous variables, 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 eqn. (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, therefore they 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.

Experiment results suggest that MRMR features are more effective for discrete variables with smaller number of features than for continuous 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.

MRMR feature selections are filter-type methods; features selected are independent of discriminant methods and are not directly aimed at producing the

best results in NB or SVM. Thus we expect these MRMR feature sets produce good results for many different classification methods.

Given the MRMR criterion functions, we use the simple linear incremental search to obtain feature set in this paper. A more complex search exploring broader combinatorial space may be pursued to see if the effectiveness of MRMR approach can be further improved. The obtained feature set should lead to more reduction of redundancy.

References:

- [1] Alizadeh, A.A., et al., "Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling," *Nature*, vol.403, pp.503-511, 2000.
- [2] Alon, U., Barkai, N., Notterman, D.A., et al, "Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays," *PNAS. USA*, vol.96, pp.6745-6750, 1999.
- [3] Burges, C.J.C., "A tutorial on support vector machines for pattern recognition," *Knowledge Discovery and Data Mining*, vol.2, pp.1-43, 1998.
- [4] Cheng, J., and Greiner, R., "Comparing Bayesian network classifiers," *UAI'99*, 1999.
- [5] Cover, T., and Thomas, J., *Elements of Information Theory*, New York: Wiley, 1991.
- [6] Ding, C., "Analysis of gene expression profiles: class discovery and leaf ordering," *RECOMB 2002*, pp.127-136, 2002.
- [7] Ding, C., and Dubchak, I., "Multi-class protein fold recognition using support vector machines and neural networks," *Bioinformatics*, vol.17, pp.349-358, 2001.
- [8] Dudoit, S., Fridlyand, J., and Speed, T., "Comparison of discrimination methods from the classification of tumors using gene expression data," *Technical Report 576*, Department of Statistics, UC Berkeley, 2000.
- [9] Garber, M.E., Troyanskaya, O.G., et al "Diversity of gene expression in adenocarcinoma of the lung," *Proc. Natl. Acad. Sci. USA*, vol.98, no.24, pp.13784-13789, 2001.
- [10] Golub, T.R., Slonim, D.K., et al, "Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring," *Science*, vol. 286, pp.531-537, 1999.
- [11] Hsu, C.W., and Lin, C.J., "A comparison of methods for multi-class support vector machines," *IEEE Transactions on Neural Networks*, vol.13, pp.415-425, 2002.
- [12] Jaakkola, T., Diekhans, M., and Haussler, D., "Using the Fisher kernel method to detect remote protein homologies," *ISMB'99*, pp.149-158, 1999.
- [13] Kohavi, R., and John, G., "Wrapper for feature subset selection," *Artificial Intelligence*, vol. 97, no.1-2, pp.273-324, 1997.
- [14] Langley, P., "Selection of relevant features in machine learning," *AAAI Fall Symposium on Relevance*, 1994.
- [15] Li, W., and Yang Y., "How many genes are needed for a discriminant microarray data analysis?," *Critical Assessment of Techniques for Microarray Data Mining Workshop*, Dec 2000. pp.137-150.
- [16] Maron, M.E., and Kuhns, J.L., "On relevance, probabilistic indexing and information retrieval," *J. ACM*, vol.7, pp.216-224, 1960.
- [17] Mitchell, T., *Machine Learning*. McGraw-Hill. 1997.
- [18] Model, F., Adorján, P., Olek, A., and Piepenbrock, C., "Feature selection for DNA methylation based cancer classification," *Bioinformatics*, vol.17, pp. S157-S164, 2001.
- [19] Park, PJ, Pagano, M, and Bonetti, M., "A nonparametric scoring algorithm for identifying informative genes from microarray data," *6th PSB*, pp.52-63, 2001.
- [20] Ross, D.T., Scherf, U., et al, "Systematic variation in gene expression patterns in human cancer cell lines," *Nature Genetics*, vol.24, no.3, pp.227-234, 2000.
- [21] Scherf, U., Ross, D.T., et al, "A cDNA microarray gene expression database for the molecular pharmacology of cancer," *Nature Genetics*, vol.24, no.3, pp.236-244, 2000.
- [22] Thomas, J.G., Olson, J.M., Stephen J., et al, "An efficient and robust statistical modeling approach to discover differentially expressed genes using genomic expression profiles," *Genome Research*, vol. 11, pp.1227-1236, 2001.
- [23] Vapnik, V., *The Nature of Statistical Learning Theory*, Springer: New York, 1995.
- [24] Welsh, J.B., Zarrinkar, P.P., et al, "Analysis of gene expression profiles in normal and neoplastic ovarian tissue samples identifies candidate molecular markers of epithelial ovarian cancer," *PNAS. USA*, vol.98, pp.1176-1181, 2001.
- [25] Weston, J., and Watkins, C., "Multi-class support vector machines," *ESANN'99*, Brussels, 1999.
- [26] Xing, E.P., Jordan, M.I., and Karp, R.M., "Feature selection for high-dimensional genomic microarray data," *ICML2001*, 2001.
- [27] Xiong, M., Fang, Z., and Zhao, J., "Biomarker identification by feature wrappers," *Genome Research*, vol.11, pp.1878-1887, 2001.