IEEE Multimedia Magazine, Special Issue on Multimedia Signal Processing and Systems in Health Care and Life Science, in press, August 2007.

Learning Microarray Gene Expression Data by Hybrid Discriminant Analysis

Yijuan Lu, Qi Tian[1], *Senior Member, IEEE,* Maribel Sanchez, Jennifer Neary, Feng Liu, Yufeng Wang[1], *Member, IEEE*

***Abstract*— Microarray technology offers a high throughput means to study expression networks and gene regulatory networks in cells. The intrinsic nature of high dimensionality and small sample size in microarray data calls for effective computational methods. In this paper, we propose a novel hybrid dimension reduction technique for classification - hybrid PCA (principal component analysis) and LDA (linear discriminant analysis) analysis. This technique effectively solves the singular scatter matrix problem caused by small training samples and increases the effective dimension of the projected subspace. It offers more flexibility and a richer set of alternatives to LDA and PCA in the parametric space. In addition, a boosted hybrid discriminant analysis is also proposed, which provides a unified and stable solution to find close to the optimal PCA-LDA prediction result and reduces computational complexity. Extensive experiments on the yeast cell cycle regulation data set show the superior performance of the hybrid analysis.**

***Keywords*— LDA, PCA, Dimension Reduction, Microarray analysis**

I. INTRODUCTION

Microarray technology provides a sizeable number of high dimensional gene expression data on different patterns. Genes of similar function yield similar expression patterns in microarray hybridization experiments, so analyzing these data and discovering their expression patterns becomes fundamental when studying networks of expression and gene regulation.

Generally, the computational analysis of gene expression data can be approached in two ways: unsupervised and supervised. A learning method is considered unsupervised if no *prior* label or class information is given. In this case, gene expression patterns are grouped by clustering algorithm based on a measure of distance (or similarity) between genes or samples. The commonly used clustering methods [1-2] in gene expression space are hierarchical clustering [3], *K*-means clustering [4] and self-organizing maps (SOMs) [5]. The unsupervised method has certain disadvantages - it cannot utilize some prior information about which samples or genes are expected to group together or construct a classifier and use the classifier to predict some unknown. Thus, supervised methods designed for classification and prediction are becoming most commonly used in microarray analysis. Supervised approaches have labeled output. They can be used to construct a robust classifier, which accurately recognizes patterns from given training samples and classifies test samples into known phenotypes based on the trained classifier.

Representative supervised classification algorithms previously used in classifying gene expression data [2] include Fisher linear discriminant analysis [6], *k* nearest neighbor [7], decision tree, multi-layer perceptron [8] and support vector machines (SVM) [9]. Although these methods have achieved some useful classification results, there is still an inevitable problem plaguing efforts to analyze high throughput microarray data - high

[1] Contact Authors: Dr. Qi Tian, Department of Computer Science, One UTSA Circle, San Antonio, TX 78249. {qitian@cs.utsa.edu}; Dr. Yufeng Wang, Department of Biology, One UTSA Circle, San Antonio, TX 78249. {yufeng.wang@utsa.edu}.

Yijuan Lu is with Department of Computer Science, University of Texas at San Antonio, TX, USA {email: lyijuan@cs.utsa.edu}

Maribel Sanchez, Jennifer Neary are with Department of Biology, University of Texas at San Antonio, TX, USA {email: msanchez@lonestar.utsa.edu, jingraha@lonestar.utsa.edu}

Feng Liu is with Department of Pharmacology, University of Texas Health Science Center at San Antonio, TX, USA {email: liuf@uthsca.edu}

dimensionality. The dimension of the genomic data is usually very high (typically from tens to hundreds) compared to the limited sample size. Machine learning is afflicted by the *curse of dimensionality*, the search space grows exponentially with the dimension. Despite the widely held view that high throughput approaches are swamping us with data, in fact much of the time *high dimensionality* obscures the details in the data.

The problem of high dimensionality can be alleviated by dimension reduction. Principal component analysis (PCA) [10, 11] and linear discriminant analysis (LDA) [12, 13, 14] are both well-known techniques for feature dimension reduction. PCA, considered as one of the simplest and best known data analysis techniques, has found various applications in many fields. LDA also plays a key role in areas of science and engineering, including face recognition [15, 16], image retrieval [17, 18], and bio-informatics [19].

LDA constructs the most discriminant features, by attempting to minimize the Bayes error through selection of the feature vectors $\boldsymbol{w}$ which maximizes $\frac{|w^T S_B w|}{|w^T S_W w|}$, where S_B measures the variance between the class means, and S_W measures the variance of the samples in the same class. It is a simple algorithm that is used for both dimension reduction and classification.

Alternatively, PCA captures the most descriptive features with respect to packing the most "energy". PCA is a useful statistical technique that has found various applications in many fields [10, 11]. It is considered as one of the simplest and best-known *Data Analysis* techniques. Its goal is to replace the original (numerical) variables with new numerical variables called "Principal Components" that have the following properties: (1) They can be ranked by decreasing order of "importance" (this term can be given a precise meaning). The first few most "important" Principal Components account for most of the information in the data. In other words, one may then discard the original data set, and replace it with a new data set with the same observations, but fewer variables, without throwing away too much information. (2) These new variables are uncorrelated.

In supervised learning, when choosing LDA and PCA, there is a tendency to prefer LDA over PCA, because, as intuition would suggest, the former deals directly with discrimination between classes, while the latter pays no particular attention to the underlying class structure. When the data for each class is represented by a single Gaussian distribution and shares a common covariance matrix, LDA will outperform PCA. However, PCA might outperform LDA when the number of samples per class is small, or when the training data samples the underlying distribution non-uniformly [20, 21]. Additionally, LDA cannot classify small sample data effectively because a singular scatter matrix problem occurs when the number of the feature dimensions is large compared to the number of training examples. Unfortunately, the training sample sizes of microarray data are often relatively small. Figure 1 shows several examples, where LDA outperforms PCA in Fig.1 (a) and where PCA outperforms LDA in Fig.1 (d) respectively.

Since both LDA and PCA have pros and cons, it is important and valuable to find a computational method that can effectively exploit their favourable attributes while simultaneously avoiding their unfavourable ones. In this paper, we propose a novel hybrid feature dimension reduction scheme (hybrid discriminant analysis) to merge LDA and PCA in a unified framework. In this technique, PCA compensates LDA for singular scatter matrix caused by small training samples and increases the effective dimension of the projected subspace. Alternatively, LDA compensates PCA for dealing directly with discrimination between classes. The hybrid PCA and LDA analysis offers more flexibility and a richer set of alternatives to LDA and PCA in the parametric space. Extensive experiments on the yeast cell cycle regulation data set demonstrate the superior performance of this hybrid analysis.

The rest of paper is organized as follows: hybrid discriminant analysis and boosting multiple classifiers constructed by hybrid discriminant analysis are illustrated in section II. Then we apply hybrid discriminant analysis on gene classification and analyze the results. Finally, we present our contribution and future work in the discussion and conclusion section.

II. Hybrid Feature Dimension Reduction

A. Hybrid Discriminant Analysis

It is common practice to preprocess microarray data by extracting linear and non-linear features. In many feature extraction techniques, one has a criterion assessing the

quality of a single feature, which ought to be optimized. Often one has prior information available that can be used to formulate quality criteria, or probably even more common, the features are extracted for a certain purpose, *e.g.*, for subsequently training some classifier. What one would like to obtain is a feature, which is as invariant to transformation (i.e. rotation, scale) as possible while still covering as much of the information necessary for describing the data's properties of interest.

A classical and well-known technique that solves this type of problem, considering only one linear feature, is the maximization of the *Rayleigh* coefficient [14].

$$J(W) = \frac{|W^T S_1 W|}{|W^T S_2 W|} \qquad (1)$$

Here, *W* denotes the weight vector of a linear feature extractor (*i.e.*, for an example **x**, the feature is given by the projections ($W^T \cdot \mathbf{x}$) and S_1 and S_2 are symmetric matrices designed such that they measure the desired information and the undesired noise along the direction *W*. The ratio in equation (1) is maximized when one covers as much as possible of the desired information while avoiding the undesired.

If we look for discriminating directions for classification, we can choose S_B to measure the separation between class centers (between-class variance), *i.e.*, S_1 in equation (1), and S_W to measure the within-class variance, *i.e.*, S_2 in equation (1). In this case, we recover the Fisher discriminant [12], where S_B and S_W are given by:

$$S_B = \sum_{j=1}^{C} N_j \cdot (m_j - m)(m_j - m)^T \qquad (2)$$

$$S_W = \sum_{j=1}^{C} \sum_{i=1}^{N_j} (x_i^{(j)} - m_j)(x_i^{(j)} - m_j)^T \qquad (3)$$

We use $\{x_i^{(j)}, i = 1, \ldots, N_j\}$, $j = 1, \ldots, C$ to denote the feature vectors of training samples. *C* is the number of classes ($C = 2$ for Fisher discriminant analysis (FDA) and C>2 for multiple discriminant analysis (MDA)), N_j is the number of the samples of the j^{th} class, $x_i^{(j)}$ is the i^{th} sample from the j^{th} class, m_j is mean vector of the j^{th} class, and *m* is grand mean of all examples.

If S_1 in equation (1) is the covariance matrix S_Σ of all the samples

$$S_\Sigma = \frac{1}{C} \sum_{j=1}^{C} \frac{1}{N_j} \sum_{i=1}^{N_j} (x_i^{(j)} - m)(x_i^{(j)} - m)^T \qquad (4)$$

and S_2 is an identity matrix, we recover standard principal component analysis (PCA) [10, 11].

We design our optimal function as

$$W_{opt} = \arg\max_W \frac{|W^T[(1-\lambda) \cdot S_B + \lambda \cdot S_\Sigma]W|}{|W^T[(1-\eta) \cdot S_W + \eta \cdot I]W|} \qquad (5)$$

where λ, η are two regularization parameters [22], S_Σ is the covariance matrix of all the training samples, and *I* is an identity matrix. The range of the parametric pair (λ, η) is from $(0,0)$ to $(1,1)$.

TABLE 1
SPECIAL CASES OF HYBRID DISCRIMINANT ANALYSIS

(λ, η)	Hybrid Discriminant Analysis	Note
(0, 0)	$W_{opt} = \arg\max_W \frac{\|W^T S_B W\|}{\|W^T S_W W\|}$	Case 1 (LDA)
(0, 1)	$W_{opt} = \arg\max_W \frac{\|W^T S_B W\|}{\|W^T \cdot I \cdot W\|}$	Case 2
(1, 0)	$W_{opt} = \arg\max_W \frac{\|W^T S_\Sigma W\|}{\|W^T S_W W\|}$	Case 3
(1, 1)	$W_{opt} = \arg\max_W \frac{\|W^T S_\Sigma W\|}{\|W^T \cdot I \cdot W\|}$	Case 4 (PCA)
$(\frac{1}{2}, \frac{1}{2})$	$W_{opt} = \arg\max_W \frac{\|W^T (S_B + S_\Sigma) W\|}{\|W^T (S_W + I) W\|}$	Case 5

With different (λ, η) values, the equation (5) provides a rich set of alternatives to PCA and LDA: $(\lambda = 0, \eta = 0)$ reduces to the full LDA; $(\lambda = 1, \eta = 1)$ recovers the full PCA; $(\lambda = 0, \eta = 1)$ gives a subspace that is mainly defined by maximizing the scatters among all the classes with minimal effort on clustering each class; $(\lambda = 1, \eta = 0)$ gives a subspace that mainly preserves the most energy while minimizing the scatter matrices of within-classes; $(\lambda = \frac{1}{2}, \eta = \frac{1}{2})$ gives a subspace that is discriminative while preserving as much energy as possible, a trade-off between LDA and PCA. Table 1

summarizes the five special cases of such a hybrid analysis. All these five cases fit certain gene feature distributions and have correspondence with some scenarios as illustrated in Fig. 1.

The difference of the proposed hybrid analysis from the existing formulae for linear discriminant analysis is subtle, but critical. Two points are worth mentioning: one is the *regularization*; and the other is the *effective dimension.* These two differences are responsible for the robust performance of the hybrid PCA and LDA analysis.

Regularization

It is well known that sample-based plug-in estimates of the scatter matrices based on equations (2-5) will be severely biased for small number of training samples. If the number of the feature dimensions is large compared to the number of training examples, the problem becomes ill-posed, *i.e.*, $|W^T S_W W| = 0$ in equation (1). A compensation or regularization can be simply done by adding quantities to the diagonal of the scatter matrices [22]. It is denoted as simple regularization scheme. If we examine the denominator of (5), by adding the part $\eta \cdot I$, the denominator will not become 0 even when the number of the feature dimensions is large compared to the number of training examples. It is equivalent to simple regularization scheme, which has been shown to significantly improve the classification accuracy in average by 15% ~ 40% in [23]. It effectively solves the singular scatter matrix problem caused by small training samples.

Effective dimension

In LDA, W maps the original d_1-dimensional data space X to a d_2-dimensional space Δ. The maximum dimension of the projected subspace is $C-1$, where C is the number of the classes [13], while in PCA there is no such limitation. Due to the full rank of the $(1-\eta)\cdot S_W + \eta \cdot I$, the hybrid discriminant analysis has *effective dimension* up to d_1, while for FDA it is only 1 and for MDA it is at most $C-1$ ($d_1 >> C$ usually). This gives the hybrid approach significantly higher capacity for informative density modeling, for which FDA has virtually none.

In order to show the advantages of hybrid discriminant analysis over PCA or LDA, we use synthetic data to simulate different sample distributions as shown in Fig. 1, which correspond to five special cases in Table 1. Original data are simulated in 2-D feature space, and positive examples are marked with "+" s and negative examples are marked with "o" s as showed in the figure. In each case, we apply PCA, LDA and hybrid discriminant analysis (HDA) to find the best projection direction by their own criterion functions. The resulting projection lines are drawn in dotted, dash-dotted and solid lines, respectively. In addition, the distributions of the examples along these projections are also drawn like bell-shaped curves along projection line, assuming Gaussian distribution for each class. The thicker curves represent the distribution of projected positive examples and the thinner curves denote the distribution of projected negative examples.

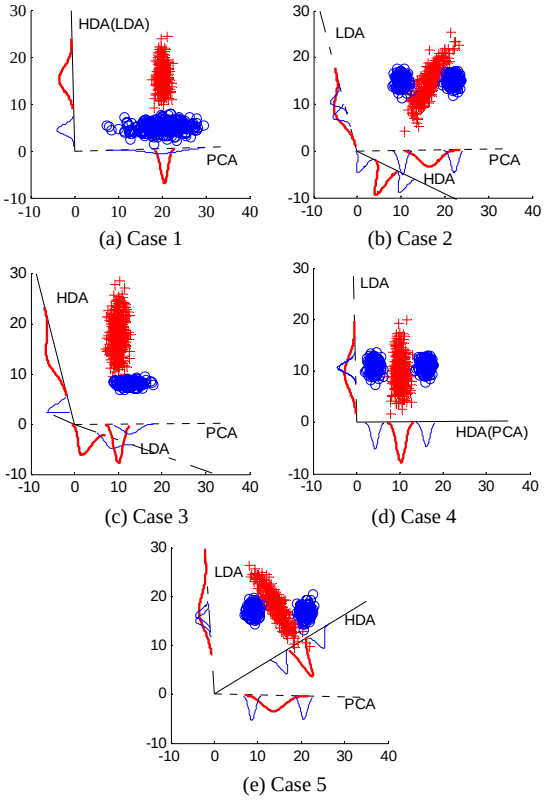

Fig. 1. Comparison of PCA, LDA and HDA for dimension reduction from 2-D to 1-D on synthetic data. (a) In case 1, HDA and LDA yield projection with good class separation. But PCA fails; In (b), (c) and (e), HDA finds good projection with class separated while PCA and LDA produce relative bad projection in case 2, case 3 and case 5; (d) In case 4, HDA and PCA give better class separation. But LDA fails.

From Fig. 1, we can see these five cases actually represent several typical data distribution scenarios. Case 1 is the scenario where the major descriptive directions of positive genes and negative genes are upright (Fig. 1 (a)). Case 2, Case 4 and Case 5 may correspond to the scenario that one gene function contains multiple subcategories (Fig. 1 (b), (d), (e)). They best fit the distribution where all positive gene expressions are alike while negative ones may be irrelevant (functional dissimilar) to each other and from different distributions. Case 3 represents the imbalanced data set. In Case 3, the size of positive genes is much larger than that of negative genes and the negative genes may be from different smaller classes (Fig. 1 (c)).

From projection results, we can see LDA treats positive and negative samples equally. It tries to cluster the positive samples and decrease the scatter of the negative samples, although some positive (negative) samples maybe come from different sub-classes. This makes it a bad choice in Case 2, Case 4 and Case 5. Similarly, since PCA captures the most descriptive features with respect to packing the most "energy", it fails in Case 1, Case 2, and Case 5. In Case 3, PCA and LDA are found not applicable for imbalanced data sets, especially the number of positive samples is much larger than that of the negative samples. The reason is that LDA or PCA tends to severely bias to cluster the positive genes or pack the most "energy" since they dominate the data set.

In all five cases, hybrid discriminant analysis yields good projection with positive samples and negative samples well separated and outperforms PCA or LDA alone. Note in Case 2, Case 3 and Case 5, both PCA and LDA totally fail while HDA still produces a good projection. It clearly demonstrates that no matter whether it is an imbalanced data set or samples are from different sub-class clusters, HDA can fit into different distributions of samples and find a balance between clustering and separating, which are embedded in the criterion function. Here, we only show five special cases of HDA. Since (λ,η) values can be any number from $(0,0)$ to $(1,1)$, more accurate data model fitting could be achieved by fine parameter tuning.

B. Boosted Hybrid Discriminant Analysis

Given the data distribution and classification task, the optimal projection of HDA that offers the best classification performance could lie outside of the line of PCA and LDA in the parametric space of (λ,η). But it is hard to tell which parameter is the best one. Searching the whole parametric space will result in extra computational complexity. It is also true that the best pair we found for one particular dataset is often different from that of another dataset and therefore this cannot lead to a generalization.

AdaBoost [30] developed in the computational machine learning area has emerged as a competitive technique that has a theoretically justified ability to improve the performance of any weak classification algorithm in terms of bounds on the generalization error.

The basic idea of boosting is to iteratively re-weight the training examples based on the outputs of some weak learners. The intention is to increase the weights of the incorrectly classified examples and decrease the weights of the correctly classified examples. This forces the classifier to focus more on the incorrectly classified examples in the next iteration. The final prediction is the combination of the prediction from each classifier weighted by its classification performance, that is, the smaller the training error rate the larger the weight.

Therefore AdaBoost provides a general way of combining and enhancing a set of PCA-LDA classifiers in the parametric space. With affordable computational cost, AdaBoost can provide a unified and stable solution to find close to optimal PCA-LDA prediction result. The re-weight and re-training mechanism is expected to enhance each classifier's performance. Unlike most of the existing approaches that boost individual features to form a composite classifier, our scheme boosts both the individual features and a set of weak classifiers.

Our algorithm is shown below:

Algorithm AdaBoost with HDA

Given: Training Sample set X and corresponding label Y
K HDA classifiers with different (λ,η)

Initialization: weight $w_{k,t=1}(x) = 1/|\mathrm{X}|$

AdaBoost:

For $t = 1,\ldots,M$

For each classifier $k = 1,\ldots,K$ do

Train the classifier on weighted mean for all the samples, positive samples and negative samples and weighted scatter

matrices in the following way. Note that $\sum_{x \in X} w_{k,t}(x) = 1$.

(a) Update weighted mean μ_{all}, μ_p, and μ_n

$$\mu_{all} = \sum w_{k,t}(x) \cdot x / \sum w_{k,t}(x)$$

$$\mu_p = \sum_{x \in p} w_{k,t}(x) \cdot x / \sum_{x \in p} w_{k,t}(x)$$

$$\mu_n = \sum_{x \in n} w_{k,t}(x) \cdot x / \sum_{x \in n} w_{k,t}(x)$$

(b) Update within-class and between-class scatter matrices and co-variance matrix

$$S_w = \sum_{x \in p} (x - \mu_p) w_{k,t}(x) (x - \mu_p)^T / \sum_{x \in p} w_{k,t}(x) + \sum_{x \in n} (x - \mu_n) w_{k,t}(x) (x - \mu_n)^T / \sum_{x \in n} w_{k,t}(x)$$

$$S_b = (\mu_p - \mu_{all}) \cdot (\mu_p - \mu_{all})^T \cdot \sum_{x \in p} w_{k,t}(x) + (\mu_n - \mu_{all}) \cdot (\mu_n - \mu_{all})^T \cdot \sum_{x \in n} w_{k,t}(x)$$

$$S_\Sigma = \sum_{x \in X} (x - \mu_{all}) w_{k,t}(x) (x - \mu_{all})^T / \sum_{x \in X} w_{k,t}(x)$$

Get the confidence-rated prediction on each sample $h_{k,t}(x) \in (-1, 1)$

Suppose the probability of a samples x belongs to positive and negative class is denoted as $P(x \in p)$ and $P(x \in n)$, respectively.

If $P(x \in p) \geq P(x \in n)$

$$h(x) = \frac{P(x \in p)}{P(x \in p) + P(x \in n)}$$

else

$$h(x) = \frac{-P(x \in n)}{P(x \in p) + P(x \in n)}$$

Compute the weight of one classifier $\alpha_{k,t}$:

$$r_{k,t} = \sum_{x \in X} w_{k,t}(x) \cdot h_{k,t}(x) \cdot y;$$

$$\alpha_{k,t} = \frac{1}{2} \ln\left(\frac{1 + r_{k,t}}{1 - r_{k,t}}\right)$$

Update the weight of each sample

$$w_{k,t+1}(x) = w_{k,t}(x) \exp(-\alpha_{k,t} \cdot h_{k,t}(x) \cdot y) / Z_t$$

where Z_t is chosen such that $\sum_{x \in X} w_{k,t}(x) = 1$.

End for each classifier

End for t

The final prediction $H(x) = sign\left(\sum_{k=1..K} \sum_{t=1..M} \alpha_{k,t} \cdot h_{k,t}(x)\right)$

III. Experiments and Results

A. *Hybrid discriminant analysis on yeast cell cycle regulation data set*

In order to evaluate our hybrid discriminant analysis on gene expression data, we apply hybrid discriminant analysis, classic dimension reduction methods such as PCA and Support Vector Machine (SVM) *etc*. on the same data set and compare their results.

1) Data set

At first, we chose to use the baker's yeast (*Saccharomyces cerevisiae)* cell cycle expression data [24] as our benchmark test data set, which contained expression vectors from a total of 80 different DNA microarray hybridization experiments on 6,221 yeast ORFs (open reading frames). We chose this data set because sequencing and functional annotation of the entire *S. cerevisiae* genome has been completed, making it an ideal test bed for estimating the accuracy of our proposed methods. According to the Comprehensive Yeast Genome Database (CYGD), a repertoire of molecular structures and functional networks in the yeast genome, 4449 out of a total of 6221 genes have annotated functions (CYGD).

The 80 microarray experiments covered a wide spectrum of conditions for cell cycle synchronization and regulation, including α factor-based synchronization, Cdc15-based synchronization, elutriation synchronization, Cln3 and Clb2 experiments, and the conditions under nitrogen deficiency and glucose depletion. The microarray data also included spotted array samples in mitotic cell division, spore morphogenesis and diauxic shift. It has been shown that combining multiple microarray studies can improve functional classification [25]. This data set has been used in numerous microarray studies and is publicly available at http://rana.lbl.gov/EisenData.htm.

To compare the performance of classification techniques, we focused on five representative functional classes: TCA cycle, respiration, cytoplasmic ribosomes, proteasomes and histone/chromosome that have been previously analyzed and demonstrated to be learnable by Brown *et al.* [26] and Mateos *et al.* [27]. Biologically, they represent categories of genes expected to exhibit similar expression profiles [25].

Out of the 4449 annotated yeast genes, genes with incomplete expression data were filtered to assure accurate evaluation. The resulting data set included 2324 annotated genes for our comprehensive evaluations. Among them, 385 genes (TCA cycle: 18, Respiration: 68, Cytoplasmic ribsome: 171, Proteasome: 77 and Histone/Chromosome: 51) belong to the above five functional classes and the remaining 1939 genes have other different functions.

2) Experiments

A well-cited microarray classification study [26] has investigated the use of SVM, two decision tree learners (C4.5 and MOC1) and Parzen windows *etc*., in gene classification to the same data set. The congruent results are that SVM, especially SVM with kernel functions, significantly outperformed the other algorithms in the functional classification. Therefore in our experiments, we focused on comparing hybrid analysis with SVM using polynomial and radial basis kernel (RBF) functions. Here polynomial kernel functions were $K(\boldsymbol{X},\boldsymbol{Y}) = (\boldsymbol{X} * \boldsymbol{Y} + 1)^d$, with $d = 1,2,3,4$ and RBF functions used were $K(\boldsymbol{X},\boldsymbol{Y}) = \exp(-\| \boldsymbol{X} - \boldsymbol{Y} \|^2 / 2\alpha^2)$. In this work, α was set to be a widely used value, the median of the Euclidean distances from each positive example to the nearest negative example [26]. Besides, for the sake of showing the hybrid discriminant analysis outperforms the single method such as single PCA *etc*, we also compared the performance of single PCA, LDA with HDA.

In order to compare to SVM, we performed a two-class classification with positive genes from one functional class and the negative genes from the remaining classes. Each gene could be classified in one of the four ways: *true positive* (TP), *true negative* (TN), *false positive* (FP) and *false negative* (FN), according to the CYGD annotation and classifier results. The yeast gene data set is an imbalanced data set - the number of negative genes is much larger than the number of positive genes. For example, with the TCA cycle the number of positive instances was only 18, whereas the number of negative instances reached 2306. In the case of an imbalanced set such as this, accuracy and single *precision* are not good evaluation metrics because FN is more important than FP [26]. Thus, we chose to use *f_measure*=2*(*Recall***Precision*)/(*Recall*+*Precision*) to measure the overall performance of each classifier, taking both *Precision* and *Recall* factors into account [28]. By definition, *Precision*= (number of TP instances)/(number of TP + FP predictions), and *Recall*=(number of TP instances)/(number of TP + FN instances). *Recall* is a measure of the completeness of the retrieved set, *i.e.*, the percentage of retrieved objects in the correct answer set. *Precision*, on the other hand, measures the purity of the retrieved set, *i.e.*, the percentage of relevant objects among those retrieved. Usually, a trade-off must be made between these two measures because improving one will sacrifice the other. In the case of imbalanced data where negative instances are dominant, *recall* is the more important measure because it focuses more on FN predictions.

In our experiments, each method classified the genes in the test set to the above five learnable functional classes and their performance was compared. When classifying one class, we set all the genes belonging to that class positive and the remaining negative. For each class, we also randomly selected 2/3 positive genes and 2/3 negative genes as a training set to do training and the remaining gene data as a testing set to do the classification. This training and testing procedure was repeated 100 times. For hybrid discriminant analysis, we searched (λ,η) from (0, 0) to (1, 1) with step size 0.1. Therefore each λ and η could have 11 options: 0, 0.1, 0.2, …, 0.9 and 1. Since each pair of (λ,η) corresponds to one feature dimension reduction schemes in-between PCA and LDA (on diagonal line from (0,0) to (1,1)) or beyond PCA and LDA (non diagonal lines), we could get a total of 121 different classifiers, called parameterised classifiers. Hence, we tested SVM, PCA, LDA and all parameterised classifiers constructed by hybrid discriminant analysis on the same data sets. Finally, we obtained the average values of *recall*, *precision* and *f_measure* for 100 rounds of each method.

3) Results

Precision, *recall* and *f_measure* for five different classifiers on the yeast five functional classes and their 95%-level confidence interval [29] are listed in Table 2. The first five methods are SVM using different polynomial kernel and RBF kernel. Here, *D-p* 1 to *D-p* 4 represents four kinds of polynomial kernel functions with

TABLE 2. COMPARISON OF *PRECISION*, *RECALL*, *F_MEASURE* FOR VARIOUS CLASSIFICATION METHODS ON YEAST CELL CYCLE REGULATION DATA SET (INCLUDING THEIR 95%-LEVEL CONFIDENCE INTERVALS)

Class	method	SVM (%)					PCA (%)	LDA (%)	HDA (%)
		D-p 1	D-p 2	D-p 3	D-p 4	RBF			
TCA cycle	*Precision*	0.0	**60.56**± 16.3	65± 15.6	28.89± 15.4	3.33± 6.2	0.0	35.24± 5.1	38.42± 4.4
	Recall	0.0	13.33± 4.1	16.67± 4.5	5.56± 3.1	0.56± 1.0	0.0	50.56± 5.5	**52.78**± 4.9
	f_measure	0.0	21.15± 6.0	25.64± 6.5	9.10± 4.9	0.95± 1.8	0.0	40.38± 4.6	43.43± 3.6
Respiration	*Precision*	0.0	71.21± 4.6	61.64± 4.2	47.31± 6.5	**90.17**± 6.7	0.0	40.73± 3.3	47.52± 2.6
	Recall	0.0	20.28± 1.8	22.22± 1.5	11.39± 1.9	11.94± 1.9	0.0	33.33± 2.0	**42.64**± 2.6
	f_measure	0.0	31.13± 2.2	32.26± 1.8	17.84± 2.7	20.68± 3.0	0.0	36.37± 2.1	**44.74**± 2.4
Cytoplasmic ribsome	*Precision*	88.27± 2.0	89.06± 2.0	86.12± 1.9	85.97± 2.1	**96.28**± 1.6	26.9± 2.6	68.89± 2.4	71.17± 1.8
	Recall	47.84± 2.0	46.55± 2.0	45.85± 2.0	43.27± 2.0	45.67± 2.0	4.44± 2.1	56.67± 1.8	**59.36**± 2.0
	f_measure	61.8± 1.9	60.89± 1.9	59.6± 1.8	57.31± 1.9	61.72± 1.9	7.56± 1.8	62.00± 1.7	**64.60**± 1.7
Proteasome	*Precision*	0.0	1.667± 3.1	0.0	0.83± 1.6	**72.5**± 15.2	0.0	37.74± 3.9	47.45± 4.2
	Recall	0.0	0.123± 0.2	0.0	0.12± 0.2	5.56± 1.5	0.0	15.06± 1.8	**16.05**± 1.6
	f_measure	0.0	0.23± 0.4	0.0	0.22± 0.4	10.17± 2.8	0.0	21.04± 2.3	**23.68**± 2.2
Histone/ Chromosome	*Precision*	10± 10.3	**90.28**± 8.9	65.29± 10.2	52.93± 11.1	86.67± 11.8	0.0	26.54± 5.2	32.32± 4.3
	Recall	0.59± 0.6	11.57± 1.9	11.18± 1.7	8.824± 1.8	9.02± 1.72	0.0	15.88± 3.1	**16.08**± 2.4
	f_measure	1.11± 1.2	20.2± 3.0	18.52± 2.7	14.66± 2.9	16.16± 3.0	0.0	19.44± 3.7	**20.99**± 2.9

d from 1 to 4. Others are PCA, LDA, and the best parameterised classifier constructed by HDA.

Table 2 exhibits the favourable and stable performance of HDA. From this table, we can clearly see that HDA outperformed all other methods for total of five classes using *recall* or *f_measure* as criteria, which are more important evaluation factors than *precision* when working with an imbalanced data set

SVM failed for most classes with small sample size and yielded very low *f_measure*. For example, for the Proteasome class, most SVM methods have almost zero *precision*, *recall* and *f_measure*, which indicates that SVM method is nearly helpless for this class. The reason being that given a small sample size, SVM could hardly find sufficient labeled data to train classifiers well. By contrast, hybrid analysis substantially improved the performance of classification, especially on *recall* and *f_measure*. In the class TCA cycle, the *recall* and *f_measure* of SVM method were 16.67% and 25.64%, while the PCA-LDA achieved 52.78% for *recall* and 44.48% for *f_measure*.

Not surprisingly, the SVM method showed fairly good and stable performance on all the five classes, especially its relative high *precision* value, but some exceptions were still observed. For example, among five classes, the *D-p* 1 SVM achieved zero *precision* and zero *recall* for three classes, which means all the positive instances recognized were wrong. The reason is that in the transformed space produced by *D-p* 1, it is hard to find a maximum-margin hyperplane to separate data. Despite the fact that higher-dimensional dot product kernel seems to have better classification, it is very hard to tell which dimension *i.e.*, *d* can give the best result.

Compared to HDA, any single analysis method performed much worse than the hybrid analysis on the total five classes. For example, PCA failed for most classes and gave zero *precision* and zero *recall*, because PCA cannot determine most discriminant features for these classes. The results for all of our yeast experiments show that the hybrid analysis has its own capability of emphasizing different aspects for the alternative schemes, and offers more flexibility than any single

method does.

B. Boosting Hybrid Discriminant Analysis on yeast cell cycle regulation data set

In previous experiments, hybrid discriminant analysis has shown promising performance. As we can imagine, for simply searching the parametric space, the larger the searched space, the better is the performance of the best single classifier. However, the exhaustive search means more computational costs. Table 3 shows the boosted HDA classifier and the best single classifier of HDA analysis in the different search space. Due to the space limit, only the performance on Cytoplasmic ribsome class of yeast cell cycle regulation data set is shown. Similar results are obtained on other classes. The range of (λ, η) is between $(0,0)$ and $(1,1)$. The searching step size of λ and η is 0.25, 0.2, 0.167, and 0.1 resulting in the searching space size 16, 25, 36, and 100, respectively.

Although Boosted HDA did not provide big performance boost in this experiment, such a minor performance enhancement may be important for performance-sensitive tasks such as gene classification. Most importantly, from Table 3, we find the boosted HDA classifier is not sensitive to the size of the search space, e.g., the boosted HDA classifier from a weak set of 16 single classifiers achieves the better performance (i.e., 62.09%) than the best single classifier (i.e., 61.80%) of search space size 100 after three iterations. Therefore, instead of searching a large parametric space to find the best single classifier, the boosted HDA classifier provides a more efficient way to combine a small set of classifiers into a more powerful one.

TABLE 3. COMPARISON OF THE BOOSTED HDA CLASSIFIER AND BEST SINGLE CLASSIFIER OF HDA PAIR ON CYTOPLASMIC RIBSOME CLASS OF YEAST CELL CYCLE REGULATION DATA SET.

Search space size	*f_measure* (%) of the best single classifier (λ^*, η^*)	Boosted HDA		
		t=1	t=2	t=3
16	61.45 (0.33,0)	61.27	61.79	62.09
25	61.44 (0.25, 0)	61.22	61.48	61.76
36	61.46 (0, 0.6)	61.59	61.90	62.47
100	61.80 (0, 0.5)	61.61	61.74	62.04

IV. DISCUSSIONS AND CONCLUSIONS

This paper proposed a novel hybrid discriminant analysis method for classification. This approach addresses the high dimensionality problem by applying hybrid discriminant analysis in an optimal linear discriminant subspace. In order to reduce the computational complexity and combine multiple classifiers into a powerful one, the boosted hybrid analysis is also proposed. The proposed approach is applied on gene classification of yeast cell cycle regulation data. Its demonstrated superior performance indicates that hybrid discriminant analysis is a promising and efficient approach to microarray data analysis.

The main contributions of this work are:

(1) Hybrid discriminant analysis provides a richer set of alternatives to single method such as LDA. As a result, it not only compensates for regularization that is afflicted by all sample-based estimation methods, but also increases the effective dimension of the projected subspace. The various tests on yeast gene expression data have shown its superior performance.

(2) In order to reduce the searching time of parameter space, we propose the boosted hybrid discriminant analysis. It boosts not only the individual features but also a set of weak classifiers. The weighted training scheme in AdaBoost adds indirect non-linearity and adaptivity to the linear methods and thus enhances it by iterations. With affordable computational cost, AdaBoost can provide a unified and stable solution to find close to optimal PCA-LDA prediction result.

(3) Hybrid discriminant analysis provided insights on transcriptomic data into the dynamics of gene networks, which could shed light on as yet unrecognized network components and interactions [31]. One main limitation on the use of genomic data to better understand cellular networks in infectious agents is the inability to assign gene functionality. This study may offer an effective solution to circumvent this problem: classifying co-expressed genes in developmental cycle helps us to identify what could conceivably be network modules.

ACKNOWLEDGMENT

This work is support in part by San Antonio Life Science Institute (SALSI) to Q. Tian and F. Liu, ARO grant W911NF-05-1-0404 to Q. Tian, and San Antonio

Area Foundation Biomedical Research Funds, NIH RCMI grant 2G12RR013646-06A1, and UTSA Faculty Research Awards to Y. Wang. J. Neary is supported by NIH MBRS-RISE (Minority Biomedical Research Support Research Initiative for Scientific Enhancement-grant GM-60655).

REFERENCES

[1] Ringner M, Peterson C, Khan J: **Analyzing array data using supervised methods.** *Pharmacogenomics*, 2002, 3, pp: 403-15.

[2] Cho SB, Won HH: **Machine learning in DNA microarray analysis for cancer classification.** *Proceedings of the First Asia-Pacific bioinformatics conference on Bioinformatics*, 2003, 19, pp: 189-198.

[3] Eisen MB, Spellman PT, Brown PO, Bostein D: **Cluster analysis and display of genome-wide expression patterns.** *Proc. of the Natl. Acad. of Sci.*, 1998, pp: 14863-14868.

[4] 4. Tavazoie S, Hughes JD, Campbell MJ, Cho RJ, Church GM: **Systematic determination of genetic network architecture.** *Nat. Genet*, 1999, 22, pp: 281-285.

[5] 5. Tamayo P, Slonim D, Mesirov J. *et al:* **Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation.** *Proc. Natl. Acad. Sci*., 1999, pp: 2907-2912.

[6] Dudoit S, Fridlyand J, Speed TP: **Comparison of discrimination methods for the classification of tumors using gene expression data.** *Technical Report 576*, 2000.

[7] Li L, Weinberg CR, Darden TA, Pedersen LG: **Gene selection for sample classification based on gene expression data: study of sensitivity to choice of parameters of the GA/KNN method.** *Bioinformatics*, 17(12), 2001, pp: 1131-1142.

[8] Khan J, Wei JS, Ringner M, Saa LH, Ladanyi M *et al*: **Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks.** *Nature Medicine*, 2001, **7**(6), pp:673-679.

[9] Brown MPS, Grundy WN, Lin D, Cristianini N *et al*: **Knowledge-based analysis of microarray gene expression data by using support vector machines.** *Proc. of the Natl. Acad. of Sci.* 2000, pp: 262-267.

[10] Jolliffe IT: **Principal Component Analysis.** 2nd edition, New-York: Springer-Verlag, 2002.

[11] Diamantaras KI, Kung SY: **Principal Component Neural Networks.** New York: Wiley, 1996.

[12] Fisher RA: **The use of multiple measurement in taxonomic problems.** *Annals of Eugenics*, vol. 7, pp.179-188, 1936.

[13] Fisher RA: **The statistical utilization of multiple measurements.** *Annals of Eugenics*, vol. 8, pp.376-386, 1938.

[14] Duda R, Hart P, Stork D: **Pattern Classification.** 2nd edition, John Wiley & Sons, Inc., 2001.

[15] Belhummeur PN, Hespanha JP, Kriegman DJ: **Eigenfaces vs. Fisherfaces: recognition using class specific linear projection.** *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(7): 711-720, 1997.

[16] Etemad K, Chellapa R: **Discriminant analysis for recognition of human face images.** *Journal of Optical Society of American*, 14(8): 1724-1733, 1997.

[17] Swets D, Weng J: **Hierarchical discriminant analysis for image retrieval.** *IEEE Trans. Pattern Analysis and Machine Intelligence*, 21(5), 396-401, 1999.

[18] Wu Y, Tian Q, Huang TS: **Discriminant EM algorithm with application to image retrieval.** *Proc. of IEEE Conf. Computer Vision and Pattern Recognition*, South Carolina, June 13-15, 2000.

[19] Ewans WJ, Grant GR: **Statistical methods in bioinformatics.** Springer-Verlag, 2001.

[20] Beveridge R, She K, Draper B, Givens G: **A nonparametric statistical comparison of principal component and linear discriminant subspaces for face recognition.** *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, Kauai, Hawaii, vol. 1, pp. 535-542, 2001.

[21] Zhu M, Martinez AM, Tan H: **Template-based recognition of sitting postures.** *Proc. IEEE Workshop on Computer Vision and Pattern Recognition for Human-Computer Interaction*, Madison, WI, 2003.

[22] Friedman H. Jerome: **Regularised Discriminant Analysis.** *Journal of the American Statistical Association*, 84, pp. 165-175, 1989.

[23] Tian Q, Yu J, Rui T, Huang TS: **Parameterized discriminant analysis for image classification.** *Proc. IEEE Int'l Conf. on Multimedia and Expo (ICME 2004)*, June 27-30, Taipei, Taiwan, 2004.

[24] Eisen MB, Spellman PT, Brown PO, Botstein D: **Cluster analysis and display of genome-wide expression patterns.** *Proc. Natl. Acad. Sci.* USA, 95(25), pp. 14863-14868, 1998.

[25] Ng S, Tan S, Sundararajan V.S: **On combing multiple microarray studies for improved functional classification by whole-data set feature selection.** *Genome Informatics* 14, pp. 44-53, 2003.

[26] Brown MP, Grundy WN, Lin D, *et al*: **Knowledge-based analysis of microarray gene expression data by using support vector machines.** *Proc. Natl. Acad. Sci*. USA, 97(1), pp. 262-267, 2000.

[27] Mateos A, Dopazo J, Jansen R, *et al*: **Systematic learning of gene functional classes from DNA array expression data by using multiplayer perceptrons.** *Genomes*. *Res*., 12(11), pp. 1703-1715, 2002.

[28] Van Rijsbergen C: **Information retrieval.** *Second edition*. Butterworths, 1979.
[29] Jain Raj: **The Art of Computer Systems Performance Analysis.** *John Wiley*, 1991.
[30] Freund Y: **Boosting a weak learning algorithm by majority,** *Information and Computation*, 121(2), pp: 256-285, 1995.
[31] Bowers PM, Cokus SJ, Eisenberg D, Yeates TO: **Use of logic relationships to decipher protein network organization.** *Science.* 306:2246-2249, 2004.

Yijuan Lu is a Ph.D. candidate in Computer Science of the University of Texas at San Antonio. She received her B.S. degree from Auhui University in 2001. She was a summer Intern Researcher at Microsoft Research Asia, MIAS Center in UIUC, and Pittsburgh Super Computing Center. She received 2007 HEB Dissertation Fellowship. Her research interests include pattern recognition and bioinformatics.

Qi Tian is an Assistant Professor in CS Department of the University of Texas at San Antonio. He received his Ph.D. in ECE from the University of Illinois at Urbana-Champaign in 2002. His research interests include Multimedia Information Retrieval, Computer Vision, and Bioinformatics. He is a Senior Member of IEEE and a Member of ACM.

Maribel Sanchez received dual B.S. degrees in Biology and Computer Science at the University of Texas at San Antonio in 2004. She was a recipient of the NIH Minority Biomedical Research Support – Research Initiative in Science Enhancement (MBRS-RISE) and Minority Access to Research Careers – Undergraduate Student Training for Academic Research (MARC-U*STAR) fellowships. Currently, she is a Systems Analyst at UTSA.

Jennifer Neary is a graduate student at the University of Texas at San Antonio. She earned a bachelors degree in biochemistry, a master's degree in biology, and is now pursuing a Ph.D. in bioinformatics. Jennifer is supported by MBRS-RISE, a federally funded research training program for minorities and disadvantaged students.

Feng Liu received his Ph.D. degree in Biochemistry in 1990 from Iowa State University. Currently he is a Professor in Department of Pharmacology and Biochemistry at the University of Texas Health Science Center at San Antonio. His research interests include insulin signal transduction pathway and molecular mechanisms regulating aging. He has published over 50 journal and conferences papers.

Yufeng Wang received her Ph.D. degree in Bioinformatics and Computational Biology in 2001 from Iowa State University. Currently she is an assistant professor with the Department of Biology and the South Texas Center for Emerging Infectious Diseases at the University of Texas at San Antonio. Her research interests include comparative genomics, systems biology, and molecular evolution of infectious diseases.