SENRA Academic Publishers, Burnaby, British Columbia
Vol. 3, No. 1, pp. 759-763, 2009
ISSN: 1715-9997

Canadian Journal of pure&applied sciences

MICROARRAY CLASSIFICATION WITH HYBRID APPROACHES

M Arif Wani
Computer Science Department, California State University Bakersfield, CA, USA

ABSTRACT

The work presented in this paper describes hybrid approaches that employ principal component analysis (PCA) and multiple discriminant analysis (MDA) methods for microarray classification. The paper first describes a hybrid approach that incorporates PCA and Fisher linear discriminant analysis (FDA) for microarray classification. This hybrid approach effectively solves the singular scatter matrix problem caused by small training samples. To increase the effective dimension of the projected subspace the use of MDA instead of FDA is explored. The performance of the system is evaluated by projecting data to several subspaces incrementally. The resulting incremental hybrid system improves the accuracy of classification. The paper discusses a comprehensive evaluation of the hybrid systems. The hybrid systems were tested on a dataset of 62 samples (40 colon tumor and 22 normal colon tissues). The results show that the use of incremental hybrid system increased the accuracy of classification of microarray data which will lead to better diagnosis of cancer and other diseases.

Keywords: Microarray classification, hybrid incremental algorithm, multiple discriminant analysis.

INTRODUCTION

One of the major applications of DNA microarray technology is to perform sample classification analyses between different disease phenotypes, for diagnostic and prognostic purposes. The classification analyses involve a wide range of algorithms such as differential gene expression analyses, clustering analyses and supervised machine learning. Machine learning algorithms are most frequently used to complete this task. Two of the most important and hard problems in microarray data analysis relate to the dimensionality of the data and to noise. Because many data analysis techniques involve exhaustive search over the object space, they are very sensitive to the size of the data in terms of time complexity. In case of microarrays, the solution is to reduce the search space vertically (in terms of genes) by using a feature selection method. The other problem is that errors occur during actual data collection and they are referred as noise in the data.

A comparative study of gene selection methods for multi-class classification of microarray data is presented by Chai and Domeniconi (2004). The authors compare several feature ranking techniques, including new variants of correlation coefficients, and Support Vector Machine (SVM) method based on Recursive Feature Elimination (RFE). A study by Hori *et al*. (2001) shows that an independent component analysis (ICA) based method can effectively and blindly classify a vast amount of gene expression data into biologically meaningful groups. Specifically, they show i) that genes, whose expression data are sampled at different times, can be classified into several groups, based on the correlation of each gene with independent component curves over time, and ii) that these classified groups by ICA based method have a good match with the classified groups that are determined by use of domain knowledge and considered to be a benchmark. These results suggested that the ICA based method can be a powerful approach to discover unknown gene functions. The authors also examine classification by principal component analysis (PCA). Then they compared the classification using PCA and ICA methods. PCA only takes into account the second-order statistics and restricts itself to orthogonal transformation to obtain principal components. On the other hand, independent component analysis (ICA) can take into account higher order statistics and can utilize non orthogonal transformation for de-mixing. Zhnag and Deng (2007) discussed the gene selection for classification using DNA microarray data. They select a compact subset of discriminative genes from thousands of genes, which is a critical step for accurate classification of phenotypes. Several widely used gene selection methods often select top-ranked genes according to their individual discriminative power in classifying samples into distinct categories, without considering correlations among genes. A limitation of these gene selection methods is that they may result in gene sets with some redundancy and yield an unnecessary large number of candidate genes for classification analyses. Another study, Rapoport *et al*. (2007) proposed a general mathematical formalism to include a priori the knowledge of a gene network for the analysis of gene expression data. The method is independent of the nature of the network, although they focus on the gene metabolic network. It is based on the hypothesis that genes close on the network are likely to be co-expressed, and consequently a biologically relevant signal can be extracted from noisy gene expression measurement by removing the "high-frequency" components of the gene

*Corresponding author email: awani@csub.edu

expression vector over the gene network. The extraction of the low-frequency component of a vector is a classical operation in signal processing that can be adapted to their problem using discrete Fourier transforms and spectral graph analysis. Wall *et al*. (2001) describes the gene expression analysis by Singular Value Decomposition (SVD), emphasizing initial characterization of the data. They described SVD methods for visualization of gene expression data, representation of the data using a smaller number of variables, and detection of patterns in noisy gene expression data. In addition, they described the precise relation between SVD analysis and Principal Component Analysis (PCA) where PCA is calculated using the covariance matrix.

Pique-Regil *et al*. (2005) propose a novel sequential DLDA (sequential Diagonal Linear Discriminant Analysis) technique that combines gene selection and classification. At each iteration, one gene is sequentially added and the linear discriminate (LD) recomputed using the DLDA model (i.e., a diagonal covariance matrix). Classical DLDA will add the gene with highest t-test score without checking the resulting model. In contrast, SeqDLDA will find the one gene that better improves class separation after recomputing the model parameters using a robust t-test score. They evaluate the new method in several 2-class datasets (Neuroblastoma, Prostate, Leukemia, and Colon) using 10-fold cross-validation and report better results. A generalized output-coding scheme has been applied to multiclass microarray classification by Shen and Tan (2006). With this, different coding strategies and decoding functions can be put into one single framework. The validity of various combinations has been verified. Support Vector Machine (SVM) was chosen as the binary classifier. Kim and Cho (2006) proposed two different correlation methods for the generation of feature sets to learn ensemble classifiers. Each ensemble classifier combines several other classifiers that learn from different features to classify cancer precisely. They adopted several feature selection methods. These feature selection methods included the Pearson's and Spearman's correlation coefficients, the Euclidean distance, the cosine coefficient, information gain, mutual information and signal-to-noise ratio. Experimental results show that two ensemble classifiers whose components are learned from different feature sets that are negatively or complementarily correlated with each other produce the good recognition rates on the chosen datasets.

A data-dependent kernel for microarray data classification was presented by Xiong *et al*. (2007). This kernel function is engineered so that the class reparability of the training data is maximized. A bootstrapping-based resampling scheme is introduced to reduce the possible training bias. Wang *et al*. (2007) use a hybrid huberized support vector machine (HHSVM). The HHSVM uses the huberized hinge loss function to measure misclassification and the elastic-net penalty to control the complexity of the model. They develop an efficient algorithm that computes the entire regularized solution path for HHSVM. They have applied their method to real microarray data and achieved promising results on both classification and gene selection.

In this work a different approach is used to solve the microarray classification problem. The approach is based on a Hybrid PCA (principal component analysis) and FDA (Fisher linear discriminant analysis) classification. The details of this approach are discussed in the next section. To increase the effective dimension of the projected subspace, the use of MDA (multiple discriminant analysis) instead of FDA (Fisher linear discriminant analysis) is explored in this work. The use of several subspaces, where data is incrementally projected, is proposed in this work. The resulting incremental hybrid PCA (principal component analysis) and MDA (multiple discriminant analysis) approach helped in enhancing the classification accuracy of the microarrays.

Hybrid Approach for Microarray Classification

A hybrid feature dimension reduction scheme that merges PCA and FDA algorithms in a unified framework was used. This hybridization of approach exploits the favorable attributes of these two methods while simultaneously avoiding their unfavorable attributes.

Principal component analysis (PCA) is a widely-used statistical technique. It works by replacing the original (numerical) variables with new numerical variables called "Principal Components", PCA captures the most descriptive features with respect to packing most "energy".

Fisher linear discriminant analysis (FDA) is a simple algorithm that is used for both dimension reduction and classification. In either case, FDA attempts to minimize the Bayes error by selecting the most discriminant feature vectors. It plays a key role in many research areas in science and engineering such as face recognition, image retrieval, and bioinformatics.

PCA and FDA, each has its own pros and cons. FDA deals directly with discrimination between classes, whereas PCA does not pay particular attention to the underlying class structure. When the data of each class can be represented by a single Gaussian distribution and share a common covariance matrix, FDA will outperform PCA. By contrast, when the number of samples per class is small or when the training data non-uniformly sample the underlying distribution, PCA might outperform FDA. In addition, FDA 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 sample sizes of microarray data are often relatively small.

A well-known technique that extracts invariant but descriptive features is the maximization of the formula [12]:

$$J(w) = \frac{|W^t S_1 W|}{|W^t S_2 W|}$$

W is the weight vector of a linear feature extractor 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.

We can choose S_B to measure the separability of class centers (between-class variance), i.e., S_1, and S_W to measure the within-class variance, i.e., S_2. In this case, we recover the well-known FDA, where S_B and S_W are given by:

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

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

Where { $x_i^{(j)}$, i=1,…,N_j},j=1,…C are feature vectors of training samples, C is the number of classes, N_j is the number of the samples of the jth class, $x_i^{(j)}$ is the ith sample from the jth class, m_j is mean vector of the jth class, and m is grand mean of all examples.

We use S_1 as the covariance matrix S_Σ of all the samples and S_2 as the identity matrix. In this case, we recover the well-known PCA, where:

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

Our optimal function will be:

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

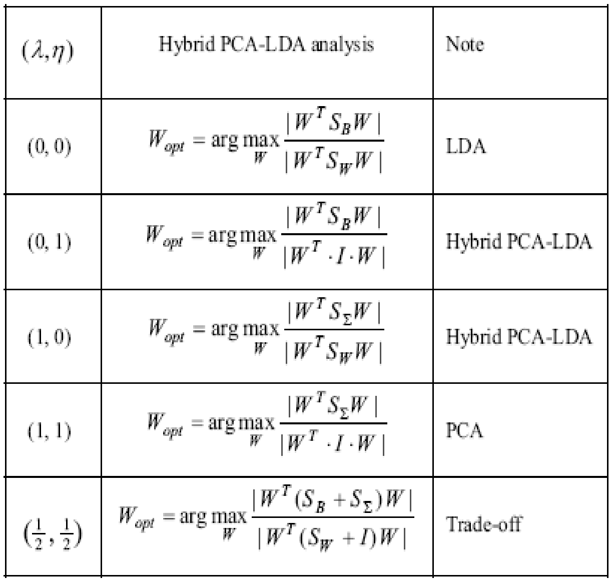
Where λ, η are two parameters, S_Σ is the covariance matrix of all the training samples, and I is the identity matrix. The range of the parametric pair (λ, η) is from (0,0) to (1, 1).

With different (λ, η) values, the last equation provides a rich set of alternatives to PCA and FDA: (λ=0, η=0) reduces to the full FDA; (λ=1, η=1) recovers the full PCA. Clearly, FDA and PCA are the special cases in the hybrid PCA and FDA analysis. (λ=0, η=1) gives a subspace that is mainly defined by maximizing the scatters among all the classes with minimal effort on clustering each class; (λ=1, η=0) gives a subspace that mainly preserves the most energy while minimizing the scatter matrices of within-classes; (λ=1/2, η=1/2) gives a subspace that is discriminative while preserving as much energy as possible, a trade-off between FDA and PCA. Table 1 summarizes these five special hybrid cases.

One approach to improve the accuracy of classification of the PCA-FDA algorithm is to project the given data to a higher dimension space. The use of Multiple Discriminant Analysis (MDA) instead of FDA can help to project data to a higher dimensional space. We will use this modification to result in a hybrid scheme that employs PCA and MDA in a unified framework.

Multiple discriminant analysis is an extension of discriminant analysis and a cousin of multiple analysis of variance (MANOVA), sharing many of the same assumptions and tests. MDA is used to classify a categorical dependent which has more than two categories, using as predictors a number of interval or dummy independent variables. MDA is a generalization of linear discriminant analysis (LDA). MDA is sometimes also called discriminant factor analysis or canonical discriminant analysis (Table 1).

Table 1. Special cases of PCA-FDA.



(λ,η)	Hybrid PCA-LDA analysis	Note
(0, 0)	$W_{opt} = \arg\max_{W} \frac{\lvert W^T S_B W \rvert}{\lvert W^T S_W W \rvert}$	LDA
(0, 1)	$W_{opt} = \arg\max_{W} \frac{\lvert W^T S_B W \rvert}{\lvert W^T \cdot I \cdot W \rvert}$	Hybrid PCA-LDA
(1, 0)	$W_{opt} = \arg\max_{W} \frac{\lvert W^T S_\Sigma W \rvert}{\lvert W^T S_W W \rvert}$	Hybrid PCA-LDA
(1, 1)	$W_{opt} = \arg\max_{W} \frac{\lvert W^T S_\Sigma W \rvert}{\lvert W^T \cdot I \cdot W \rvert}$	PCA
$(\frac{1}{2},\frac{1}{2})$	$W_{opt} = \arg\max_{W} \frac{\lvert W^T(S_B + S_\Sigma)W \rvert}{\lvert W^T(S_W + I)W \rvert}$	Trade-off

Multiple discriminant analysis adopts a perspective similar to Principal Components Analysis, but PCA and MDA are mathematically different in what they are maximizing. MDA maximizes the difference between values of the dependent, whereas PCA maximizes the variance in all the variables accounted for by the factor.

In this modification, the same equations stated in the PCA-FDA algorithm were used. Instead of projecting the data into a 1D space, we projected the data into a 2D space. This was done by using the two eigenvectors that corresponds to the largest two eigenvaluse to classify the samples in the dataset. Note that only one eigenvector was used to classify data in the PCA-FDA method.

As shown in the results section, the accuracy of classification of the proposed PCA-MDA method was better than PCA-FDA approach but it was not satisfactory enough. This was mainly due to the reason that the projected data representing the various classes overlapped. This problem can be solved by projecting the data into several subspaces using an incremental approach that is described below.

The incremental PCA and MDA approach projects data into several subspaces using various values of eigen values. Each eigen value results into a space with a particular orientation. The steps to obtain various subspaces are summarized below:

Initialize λ to 0.2 and $\Delta\lambda$ to 0.2.

1. Project data into a subspace with the current value of λ. Identify ranges of values in the projected subspace that discriminate positive and negative examples correctly.
2. Update λ to $\lambda + \Delta\lambda$.
3. Terminate the procedure if $\lambda >= 1$. Otherwise go to step 1.

The several subspaces obtained incrementally are used to classify the given data. This procedure proved to be more efficient than the hybrid approaches described above.

RESULTS AND DISCUSSION

The dataset chosen in this work is the same that was used by Alon *et al.* (1999). The data set is composed of 40 colon tumor and 22 normal colon tissue samples which were analyzed with an Affymetrix oligonucleotide array complementary to more than 6,500 human genes. A two-way clustering algorithm was applied to both the genes and the tissues, revealing broad coherent patterns that suggest a high degree of organization underlying gene expression in these tissues. 2000 genes were chosen to be the features for each sample. The tissues were taken from 40 patients. The training data set consist of 40 samples (26 tumor and 14 normal) and the testing data set consists of 22 samples (14 tumor and 8 normal).

In the hybrid PCA-FDA method, different combinations of λ, η have been used to find out the best combination. With λ=0 and η=1, 35 out of 40 samples were correctly classified in the training data set and 15 out of 22 samples were correctly classified in the testing data set. With λ=1 and η=0, 24 out of 40 samples were correctly classified in the training data set and 16 out of 22 samples were correctly classified in the testing data set. With λ=1/2, η=1/2, 40 out of 40 samples were correctly classified in the training data set and 15 out of 22 samples were correctly classified in the testing data set.

The accuracy of classification of the Hybrid PCA-FDA method is summarized below:

In the hybrid PCA-MDA method, we have also tried the same combinations of λ, η as we have done in the hybrid PCA-FDA method. With λ=0 and η=1, 36 out of 40 samples were correctly classified in the training data set and 16 out of 22 samples were correctly classified in the testing data set. With λ=1 and η=0, 25 out of 40 samples were correctly classified in the training data set and 16 out of 22 samples were correctly classified in the testing data set. With λ=1/2 and η=1/2, 40 out of 40 samples were correctly classified in the training data set, 15 out of 22 samples were correctly classified in the testing data set.

The accuracy of classification of the Hybrid PCA-MDA method is summarized in table 2 and 3.

Table 2. PCA-FDA accuracy.

	$\lambda = 0, \eta = 1$		$\lambda = 1, \eta = 0$		$\lambda = ½, \eta = ½$	
	Train	Test	Train	Test	Train	Test
Accuracy	87.5%	68.18%	60%	72.7%	100%	68.18%

Table 3. PCA-MDA accuracy.

	$\lambda = 0, \eta = 1$		$\lambda = 1, \eta = 0$		$\lambda = ½, \eta = ½$	
	Train	Test	Train	Test	Train	Test
Accuracy	90%	72.73%	62.5%	72.7%	100%	68.18%

The accuracy of the PCA-MDA method is better than the PCA-FDA method in some cases and is the same in other. This increase in accuracy in PCA-MDA is due to the projection of data into a 2D space which had helped in separating the data in a way that most of the samples of the same class are closer together.

The accuracy of classification was further improved by using the incremental hybrid approach. The results of classifying the training set with this approach were 100% while as that of test data set was 91%.

CONCLUSION

In this paper, incremental hybrid approaches for microarray data classification were employed. First the paper discussed PCA (principal component analysis) and FDA (Fisher linear discriminant analysis) hybrid approach for classification and evaluated the approach by noting its accuracy on different values of λ and η. The results were improved by modifying the above approach that enabled projecting the data to a higher dimensional space. This modification was based on a hybrid PCA (principal component analysis) and MDA (Multiple discriminant analysis) method. The modified method is shown to improve classification performance. The results were further improved by employing incremental hybrid approach. These results guide the development of a software system that will fully automate cancer diagnostic model. In future this will be used in clinics and health care facilities to achieve better treatment for cancer patients.

REFERENCES

Alon, U., Barkai, N., Notterman, DA., Gish, K., Ybarra, S., Mack, D. and Levine, AJ. 1999. Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. Proceedings of the National Academy of Sciences. 96(12):6745-6750.

Chai, H. and Domeniconi, D. 2004. An Evaluation of Gene Selection Methods for Multi-class Microarray Data Classification. Proceedings of the Second European Workshop on Data Mining and Text Mining in Bioinformatics. 3-10.

Hori, G., Inoue, M., Nishimura, S. and Nakahara, H. 2001. Blind gene classification based on ICA of microarray data. Proceedings of 3rd International Conference on Independent Component Analysis and Blind Signal Separation. San Diego. 332-336.

Kim, K. and Cho, S. 2006. Ensemble classifiers based on correlation analysis for DNA microarray classification. Neurocomputing . 70:187-199.

Pique-Regi1, R., Ortega, A. and Asgharzadeh, S. 2005. Sequential Diagonal Linear Discriminant Analysis (SeqDLDA) for Microarray Classification and Gene Identification. Proceedings of the 2005 IEEE Computational Systems Bioinformatics Conference–Workshop. 112-116.

Rapaport, F., Zinovyev, A., Dutreix, M., Barillot, E. and Vert, J. 2007. Classification of microarray data using gene networks. BMC Bioinformatics. 8:35.

Shen, L. and Tan, E. 2006. A Generalized Output-Coding Scheme With SVM For Multiclass Microarray Classification. Proceedings of 4th Asia-Pacific Bioinformatics Conference.

Wall, ME., Dyck, PA. and Brettin, TS. 2001. SVDMAN: Singular Value Decomposition Analysis of Microarray Data. Bioinformatics. 17:566-68.

Wang, L., Zhu, J. and Zou, H. 2007. Hybrid Huberized Support Vector Machines for Microarray Classification. Proceedings of the 24th International Conference on Machine Learning. Corvallis, Oregon. 983-990.

Xiong, H., Zhang, Y. and Chen, X. 2007. Data-dependent Kernel Machines for Microarray Data Classification. IEEE/ACM Trans. Comput. Biology Bioinform. 583-595.

Zhang, J. and Deng, H. 2007. Gene selection for classification of microarray data based on the Bayes error. BMC Bioinformatics. 8:370.