CROSS-WEIGHTED FISHER DISCRIMINANT ANALYSIS FOR VISUALIZATION OF DNA MICROARRAY DATA

Xinying /Zhang, Chad L. Myers and S.Y. Kung

Princeton University

ABSTRACT

Fisher's Discriminant Analysis has recently shown promise in dimensionality reduction of high dimensional DNA data. However, the one-dimensional projection provided by this method is an optimal Bayesian classifier only when the intraclass data patterns are purely Gaussian distributed. Unfortunately, it has been well recognized that most DNA expression data are much more realistically represented by a Gaussian mixture model (GMM), which allows for multiple cluster centroids per class. When a data set from such a GMM is projected onto a one-dimensional subspace, its inherent multi-modal nature may be partially or completely obscured. Consequently, traditional Fisher DA is quite inadequate when higher dimensional visualization (e.g. 2-D or 3-D) is necessary. The proposed technique addresses this problem and makes use of combined supervised and unsupervised learning techniques for several DNA microarray signal processing functions, including intraclass cluster discovery, optimal projection, and identification/selection of responsible gene groups. In particular, a cross-weighted Fisher Discriminant Analysis is proposed and its abilities to reduce dimensionality and to visualize data sets are evaluated.

1. INTRODUCTION

For cancer treatment, it is now well recognized that cancers with histopathologically similar appearance may follow significantly different clinical courses and show different responses to therapy. It is therefore important to target specific therapies to pathogenetically distinct cancer types or subtypes so as to maximize efficacy and minimize toxicity. Recent developments in gene microarray technology provide us with large sets of high-dimensional gene expression data useful in such classification and prediction tasks. Briefly, DNA microarray experiments yield a number of gene expression intensity values (scalars) for each of a set of tissues in question. In general, the resulting gene expression levels are numerous (on the order of 10^4 or higher) and have very high resolution. The number of samples (i.e. patient dimension), however, is relatively small (e.g. < 100).

In this paper, our goal is to develop and test effective machine learning computational and visualization tools to reveal and interpret the rich information derived by microarrays about underlying cancer biology. Furthermore, we wish to facilitate molecular classification/prediction of known cancer types and the discovery of new types. Given a set of DNA microarray data and corresponding class knowledge, we identify two distinct objectives pivotal to our understanding of class prediction and discovery:

1. Gene Pre-selection: find a subset of the genes that provides the most information regarding sample classification.
2. Optimal Visualization Projection: choose the projection of this subset of genes that yields the most distinction between classes.

1.1. Data Clustering Models

The formulation of a learning mechanism that achieves a visualization or class prediction goal is inherently tied to the assumptions on the underlying data distribution. For instance, optimality of a particular classification/visualization criterion for one type of distribution may not necessarily correspond to optimality under the assumption of a different distribution. Thus, choosing a criterion is closely linked to choosing the correct cluster model for a particular data set. Note that the terms *class* and *cluster* are used with very distinctive meaning in this paper. We use *class* to refer to a particular pathological type, and *cluster* to refer to an arrangement of samples (e.g. patient tissues) within a single *class*. Two common models for distributions of clusters within a given class are:

1. **Single-Cluster Gaussian Model:** Within a class, the simplest model is a single-cluster Gaussian distribution. The advantage of such a simple statistic model is twofold: (1) It is straightforward to estimate the density parameters (centroid and covariance) for each class. (2) Such a simple cluster model leads to a closed-form optimal Bayesian classifier known as the Fisher classifier [4].
2. **Multiple-Cluster Gaussian Mixture Model:** Most DNA gene expression data do not fit the aforementioned single-cluster Gaussian model. Indeed, they can be much more accurately modelled via a GMM (or SFNM) model [7, 8], i.e. the distribution of the data set within each class is modelled as a mixture of multiple normal distributions (with possibly different means and covariances).

For a multi-cluster GMM distribution, the Fisher classifier is no longer an optimal Bayesian classifier. Therefore, novel pre-selection and projection techniques must be developed. In general, three subtasks will be required for both:

- Estimation of density parameters and overall mixture: A popular clustering method can be applied for this purpose, e.g. the k-means or EM algorithms. Introducing user interaction into the clustering algorithm is a more practical approach, which greatly reduces both computational complexity and local optimum likelihood [5].
- Selection of an optimal number of clusters within a class: The following two approaches are worthy considerations.
    1. Information theoretic metrics, such as AIC or MDL [9, 7, 8].

2. Robustness metric: The number of clusters is also tightly coupled with the cluster mass, i.e. the larger number of clusters the smaller the average cluster mass. The choice of cluster mass depends on the desired classification objectives. If the strategy is to achieve a robust classification, then the Fisher's DCA based on the single-cluster Gaussian model is superior. In contrast, multiple-cluster GMM will be preferred if the strategy is to display the fine cluster structure and e.g. yield a minimum error rate at the expense of robustness.

- Design of an effective pre-selection/projection criteria reflecting new GMM assumption as the traditional DCA is no longer a statistically optimal solution.

Without loss of generality, we focus on the two-class case, with multiple clusters per class.

1.2. Supervised vs. Unsupervised Learning

Throughout our analysis, we assume that we are able to obtain class information (known phenotypes from biological experimental setting). However, within-class structure is treated as unsupervised learning as is suggested in the previous section. This hybrid supervised/unsupervised learning strategy resembles that proposed in [7, 8], where the (supervised) DCA approach is shown to be very effective on hierarchical sub-levels when some supervision is passed down from upper-level processing. While those results suggest that a small amount of supervision complements a mostly unsupervised approach, the present paper will demonstrate that intra-class unsupervised clustering (e.g. k-mean or EM methods) can be instrumental in revealing fine cluster structure for the supervised problem. Both suggest that such hybrid approaches provide useful insight that may otherwise be missed.

1.3. Pairwise Fisher Linear Discriminant

For future reference, we introduce the notion of a pairwise Fisher Linear Discriminant. Given a pair of clusters[1], say, cluster α (from class 1) and cluster β (from class 2) with means $\underline{m}_\alpha$, $\underline{m}_\beta$, and covariances $\mathbf{S}_\alpha$, $\mathbf{S}_\beta$ respectively, the goal of Fisher's discriminant is to find the linear projection, $\underline{w}$, that maximizes the contrast criterion $J(\underline{w})$, the ratio of inter-cluster distance to intra-cluster variance:

$$\begin{aligned} J_{\alpha,\beta}(\underline{w}) &= \underline{w}^T \mathbf{M}(\alpha,\beta)\underline{w}; \\ \mathbf{M}(\alpha,\beta) &\equiv \mathbf{S}_{\alpha\beta}^{-1}(\underline{m}_\alpha - \underline{m}_\beta)(\underline{m}_\alpha - \underline{m}_\beta)^T, \end{aligned} \quad (1)$$

where $\|\underline{w}\| = 1$ and $\mathbf{S}_{\alpha\beta} \equiv \frac{P_\alpha S_\alpha + P_\beta \mathbf{S}_\beta}{P_\alpha + P_\beta}$.

The quantity $J_{\alpha,\beta}(\underline{w})$ represents the pairwise Fisher discriminant power of clusters α and β along the projection direction $\underline{w}$. The *pairwise Fisher discriminant power*, defined as the maximum value of $J_{\alpha,\beta}(\underline{w})$, is well known to be

$$\max_{\underline{w}} J_{\alpha,\beta}(\underline{w}) = \lambda_{\mathbf{M}(\alpha,\beta)}, \quad (2)$$

[1]Note that Fisher's Linear Discriminant has been traditionally applied to a pair or set of multiple *classes*. Here, we propose such a discriminant be defined between pairs of *clusters*, where each cluster is a member of particular *class*.

where $\lambda_{\mathbf{M}(\alpha,\beta)}$ denotes the largest eigenvalue of the matrix $\mathbf{M}(\alpha,\beta)$. The corresponding optimal projection is therefore along the direction of the principal eigenvector:

$$\underline{w}_{opt} = \mathbf{S}_{\alpha\beta}^{-1}(\underline{m}_\alpha - \underline{m}_\beta). \quad (3)$$

2. GENE PRE-SELECTION

Before attempting visualization or class prediction for a particular data set, we first want to identify those genes that are most informative with regard to the sample classification (gene pre-selection). Without class information, gene selection can be done by either simply choosing those genes with highest minimal intensity across the samples, or pursuing more sophisticated data separability criteria via projection pursuit (PP) or independent component analysis (ICA) [7, 8]. In contrast, when class information is known, Fisher's discriminant criterion is commonly recognized to be most effective in measuring the interclass separability.

2.1. Individual Gene Pre-selection by Fisher Discriminant

We propose that Fisher's discriminant power (see 2) is not only useful in obtaining a 1-D projection, but is a convenient metric for the separability offered along a single gene dimension. Our pre-selection approach entails computing $J_{\alpha,\beta}(\underline{w})$ (a scalar) along each gene dimension (i.e. $w = [0,\ldots,0,1,0,\ldots]$) where α,β now refer to class 1 and class 2 (no within-class cluster information is used). The resulting values for each gene can then be used as a basis for selecting those genes that are most indicative of class distinction.

2.2. Cross-Weighted Ambiguity Function

Because Fisher's discriminant is not optimal for a GMM assumption, we also propose a variation of the aforementioned pre-selection scheme. We need to first perform unsupervised clustering (e.g. k-mean or EM algorithm) for each gene's 1-D expression space. It should be obvious that class separability depends most critically on the pair(s) of closest cross-class clusters. Thus, just like SVM, for maximizing class separability, special attention should be paid to the border clusters. Note that, for the 1-D case, $J_{\alpha,\beta}(w) = \mathbf{M}(\alpha,\beta)$ degenerates into a scalar value. For a pair of clusters, $\mathbf{M}(\alpha,\beta)^{-1}$ offers as an effective measure of pairwise ambiguity. (In other words, the ambiguity is inversely proportional to the distance of any cross-class cluster-pair.) Accordingly, we propose to use the following total ambiguity metric:

$$\text{ambiguity} = \sum_{\alpha\epsilon C_1} \sum_{\beta\epsilon C_2} P_\alpha P_\beta \mathbf{M}^{-1}(\alpha,\beta) \quad (4)$$

which can be shown to place emphasis on closer cluster-pair(s). The genes with the least total ambiguity are chosen in the pre-selection process.

2.3. Hybrid Selection Scheme

In general, the Fisher DCA pre-selection approach offers more robustness while the cross-weighted ambiguity scheme places heavier weights on the closer inter-class cluster-pairs. A safe strategy is to choose the genes which exhibit good behavior in both accounts. More specifically, choose those genes which yield maximum

$$J_{1,2}(g_1) + \sum_{\alpha\epsilon C_1} \sum_{\beta\epsilon C_2} P_\alpha P_\beta \mathbf{M}(\alpha,\beta) \quad (5)$$

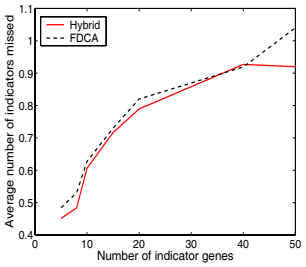

Fig. 1. Pre-selection results via Fishers discriminant on individual genes and hybrid with Cross-weighted ambiguity function (average of 150 independent simulations).

2.4. Pre-Selection Simulation Results

For testing the three gene selection methods, we generate a synthetic data set consisting of two classes, each with 100 samples and 4 and 5 Gaussian mixtures respectively. Each sample includes 200 gene expression values, with the number of indicator genes varying as plotted in Fig. 1. For each instance, the number of indicator genes omitted by the pre-selection stage is recorded. While all schemes are relatively successful, we find that the cross-weighted ambiguity metric alone (not shown here) is unable to compete with the robustness of simple Fisher DCA. The hybrid approach, however, seems to offer a slight improvement.

3. OPTIMAL PROJECTION FOR VISUALIZATION

The purpose of developing discriminatory data projection tools is to maximally discover hidden (global and fine) cluster structure in the data space. The huge dimensionality ($500 \sim 8000$) of microarray data introduces a new challenge in the revelation of data structure into a 3-D (or lower dimensional) space to support effective human interactions. In order to find an optimal view for cluster structure discovery, we have to search the space of possible gene vectors. We shall first review the traditional Fisher DCA and then point out why and how it could be modified to better represent the GMM nature of the DNA gene expression data. In particular, we have extended a notion of weighted Fisher criteria [6, 7] to a so-called Cross-Weighted DCA which is meant for class separability under a supervised learning assumption.

3.1. Fisher DCA and the Fundamental Deficiency

The classification of data samples can be visualized by adopting the traditional Fisher projection. If class information is known, then it can be effectively used to provide better guidance for discovering cluster structure. To obtain the best separability between the two classes, the optimal pairwise Fisher vector $\underline{w}$ can be derived from (3) with the clusters α and β now representing the two classes to be identified. As the pairwise Fisher matrix $\mathbf{M}(\alpha, \beta)$ for two classes (or clusters) in (1) has rank one, the separation can be only visualized on a 1-D space.

The consideration of using multiple-dimensional data projection tools is primarily based on the fact that most gene expression microarray data are a mixture of samples of cancer and non-cancer, or a mixture of samples of various types of cancers. As a result, the GMM (or SFNM) model may be the best approach for describing such multi-modal data structure [3]. However, the Bayesian optimality claims for pairwise Fisher is no longer valid for the more general multi-cluster GMM distribution. Another critical deficiency of the traditional Fisher DCA is that it does not suggest any projection direction beyond one-dimensional subspace.

Example 1 (Illustrative Example)
Let us use a simple two-gene, two-class, and two-clusters-per-class example to show why the Fisher DCA is not optimal. The centers are $[1, 10]$, $[1, 0]$ *for the two clusters belonging to the first class and* $[-1, 0]$, $[-1, -10]$ *for the second class. If the* 2*-D data is projected to* 1*-D space with projection angle* θ*, the centers will be* $\cos\theta + 10\sin\theta$, $\cos\theta$, $-\cos\theta$, *and* $-\cos\theta - 10\sin\theta$, *leading to a classification error rate:* $0.25[erfc(\frac{\cos\theta+10\sin\theta}{\sqrt{2}\sigma}) + erfc(\frac{\cos\theta}{\sqrt{2}\sigma})]$, *where* $\sigma = 3$ *denotes the cluster variance. It can be verified that the Fisher's linear projection angle is* 53° *with an error of* 21%, *while the best projection angle should be* 38° *and yield a lower error rate* 20%.

3.2. Cross-Weighted DCA

In short, the GMM distribution implies that the data points are most likely forming multi-modal structure. When a GMM data set is projected onto a one-dimensional subspace, its inherent multi-modal nature may be partially or completely obscured according to Cover's theorem on the separability of patterns [1]. Fortunately, the concealed fine cluster structure is often quite visible via a higher dimensional (2-D or 3-D) display. The deficiency of traditional Fisher based DCA motivates the cross-weighted DCA for cluster visualization.

First, k-mean or EM methods are applied to all the preselected genes to yield an estimate of the center and covariance of each cluster. In order to find projections that maximizes the cluster-separability (as opposed to class-separability), we propose a modified linear discriminant. Again let us adopt the same pairwise Fisher's scatter matrix, cf. (1). To highlight the inter-class cluster separation, a *cross-weighted Fisher matrix* is introduced:

$$\mathbf{M}_{cw} = \sum_{\alpha \epsilon C_1} \sum_{\beta \epsilon C_2} \gamma(\Delta_{\alpha\beta}) P_\alpha P_\beta \mathbf{M}(\alpha, \beta), \tag{6}$$

where $\Delta_{\alpha\beta} = \sqrt{(\underline{m}_\alpha - \underline{m}_\beta)^T \mathbf{S}_{\alpha\beta}^{-1} (\underline{m}_\alpha - \underline{m}_\beta)}$ is the weighted distance between two clusters and $\gamma(\Delta) = \frac{1}{2\Delta^2}\text{erf}(\frac{\Delta}{2\sqrt{2}})$ is a proper weighting assigned to place more emphasis on closer inter-class cluster-pairs [6]. This naturally yields the cross-weighted discriminant function:

$$J_{cw}(\mathbf{W}) = Trace[\mathbf{W}^T \mathbf{M}_{cw} \mathbf{W}] \tag{7}$$

where W is a $K \times N$ projection matrix and N is the number of genes. It is straightforward to show that the best K (say, $K = 3$) projection vectors for optimization of $J_{cw}(\mathbf{W}$ are the K principal eigenvectors of $\mathbf{M}_{cw}$.

3.3. Hybrid Scheme Using DCA and CW-DCA

In general, the DCA offers more robustness while CW-DCA tends to reveal more fine structure of the data clusters. Fortunately, DCA will consume only one of three display dimensions in 3-D visualization. Therefore, a safe (and highly recommended) strategy is to adopt the following hybrid scheme.

1. Determine the optimal DCA vector, denoted by $\underline{w}_f$.

2. Then, we obtain the two best complementary vectors from the 2 principal eigenvectors of the following deflated CW-DCA matrix:

$$\mathbf{M}_h = [\mathbf{I} - \underline{w}_f \underline{w}_f^T]\mathbf{M}_{cw}[\mathbf{I} - \underline{w}_f \underline{w}_f^T] \quad (8)$$

3. The combination of the one DCA vector and two CCW-DCA vectors will be used as the projection vectors for the final 3-D display.

Of course, it is also possible that we reverse the order by finding two CW-DCA vectors first and then find a complementary DCA solution as the third vector.

3.4. Simulation Results

1. **GMM Synthetic Data:** A demonstration of the capability of finding cluster structure by Fisher DCA, Cross-Weighted DCA, and Hybrid DCA/CW-DCA is first done on the simulated data set discussed in Sec. 2.4 with 500 total genes per sample, of which 30 are actually class indicators. The results are illustrated in Figure 2. Note the trade-off between class separation robustness and fine cluster structure identification evident in the Fisher DCA and CW-DCA plots. Also, note the apparent compromise offered by the hybrid approach.
2. **MIT Acute Leukemia Data:** To verify each approach with real DNA microarray data, we obtain leukemia tissue sample data with class knowledge from [2]. Simulation results are illustrated in Figure 2. Again, note the additional cluster structure information provided by both CW-DCA and the Hybrid DCA/CW-DCA approaches, even when no knowledge of in-class clusters is given.

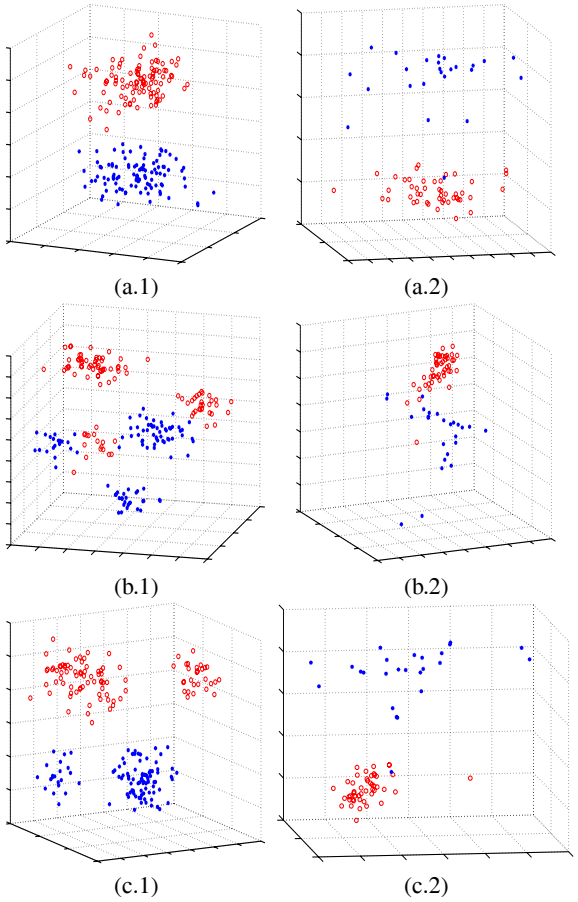

Fig. 2. Visualization results for synthetic (1) and MIT leukemia (2) data sets. (a) Fisher DCA projection (b) CW-DCA 3-D Projection and (c) Hybrid DCA/CW-DCA 3-D Projection.

4. CONCLUSION

Our approach offers a new data visualization method for the revelation of high dimensional and multi-modal DNA microarray data. The technique optimizes the class separability while revealing local cluster structure more effectively. In summary, the DNA microarray data mining and visualization represents an enormous challenge and opportunity to information scientists.

5. ACKNOWLEDGEMENTS

The authors wish to thank Dr. Joseph Yue Wang at Virginia Tech. and Dr. Zuyi Wang at Childrens National Medical Centre for insightful discussions.

6. REFERENCES

[1] S. Haykin, *Neural Networks: A Comprehensive Foundation,* 2nd ed., Prentice-Hall, Inc., Upper Saddle River, New Jersey, 1999.

[2] T. R. Golub, D. K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. P. Mesirov, H. Coller, M. L. Loh, J. R. Downing, M. A. Caligiuri, C. D. Bloomfield, and E. S. Lander, "Molecular classification of cancer: class discovery and class prediction by gene expression monitoring," *Science*, vol. 286, pp. 531-537, Oct. 1999.

[3] D. M. Titterington, A. F. M. Smith, and U. E. Markov, *Statistical analysis of finite mixture distributions*. New York: John Wiley, 1985.

[4] K. Fukunaga, *Introduction to Statistical Pattern Recognition*, 2nd ed., New York: Academic Press, 1990.

[5] Y. Wang, L. Luo, M. T. Freedman, and S-Y Kung, "Probabilistic principal component subspaces: A hierarchical finite mixture model for data visualization," *IEEE Trans. on Neural Networks*, vol. 11, no. 3, pp. 625-636, May 2000.

[6] M. Loog, R.P.W. Duin, and R. Haeb-Umbach, "Multiclass linear dimension reduction by weighted pairwise fischer criteria, *IEEE Trans. on Pattern Analysis and Machine Intelligence*, Vol. 23, No. 7, pp. 762-766, July 2001.

[7] Z. Wang, S.-Y. Kung, J. Zhang, J. Khan, J. Xuan and Y. Wang, "Computational intelligence approach for gene expression data mining and classification", *Proc. IEEE International Conference on Multimedia & Expo,* , July 2003.

[8] Z. Wang, Y. Wang, J. Lu,S.Y. Kung, J. Zhang, and , et al., "Discriminative mining of gene microarray data", *Journal of VLSI Signal Processing*, pp. 255-272, Nov. 2003.

[9] J. Rissanen, "Modeling by shortest data description," *Automatica*, vol. 14, pp. 465-471, 1978.