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

Xinying Zhang, Chad L. Myers, 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

It is now well recognized that cancers with histopathologically similar appearance may follow significantly different clinical courses and show different responses to therapy. There is a need, for cancer treatment, to target specific therapies to pathogenetically distinct cancer types or subtypes, to maximize efficacy and minimize toxicity.

The recent development of gene microarrays provides an opportunity to take a genome-wide approach to predict clinical heterogeneity in cancer treatment. In a DNA microarray experiment, a tumor mixture is prepared with fluorescent tags. If it contains the complementary RNAs of an oligo, they will bind and result in a attached fluorescent hybrid. With laser scan, the gene expression levels can be derived from the hybridized tissue samples. Each sample case is described as a point in a d-dimensional gene expression space in which each axis represents the expression level of one gene. Gene expression levels derived from tissue samples can be assayed for thousands of genes simultaneously. This technology facilitates monitoring the whole genome on a single chip so that researchers can have a better (and bigger) picture of the interactions among thousands of genes simultaneously. In short, it revolutionizes the gene analysis tools. It has two important applications: (1) Target Gene Selection for Drug Design: Microarrays provide pharmaceutical firms with a means to identify drug targets. (2) Classification of Disease: E.g. different leukemia (ALL/AML) or breast-cancer (ER+/ER-), can be identified by different patterns of gene expression.

In this paper, our goal is to develop and test effective machine learning computational tools to reveal and interpret the rich information derived by microarrays about underlying cancer biology and to facilitate molecular classification/prediction of cancer and response to therapy. Ultimately, it can help reveal critical distinguishing patterns among gene expression profiles and identify genes which are most responsible for the cell's phenotypes.

1.1. Data structure

1. 1. Spatial Dimension: Usually large size. A DNA array may contain up to 100,000 probe oligomers (50-80 bp) with spot diameter as small as 150. (A particular oligo attaches a particular gene.)
2. 2. Case/Patient Dimension: Usually small size. Intensity resolution: The intensity level of image reflecting the copies of gene made during the process - has very high resolution.

As an example, for the MIT leukemia data set, the data set contains intensities of 7,129 genes in 50 ALL and 22 AML tissues. The genes chosen are the ???? genes with highest minimal intensity across the samples. The vertical axis corresponds to genes, and the horizontal axis to tissues. The color code used is indicated in the adjoining scal

***

The main challenge lies in that the microarray data high-dimensional, multi-modal, and lacking in prior knowledge.

It will require a very discriminative visualization tool, which will in turhn involve a very rich information interplay between two cluster structures: gene clustering structures and case/ clustering structures.

Figure 1: Expanded view of biologically distinct gene expression signatures designed by hierarchical clustering.

1.1.1. Major Application Tasks:

Cluster Discovery applies hierarchical cluster strategy to detect/validate previously unrecognized tumor subtypes.

Gene Selection identifies the most relevant gene subset involving the biological process that generates the patterns.

Phenotype Prediction (in vitro) assigns unknown tumor sample to known tumor classes.

Change Profile (in vivo) identifies the effects of drug on the expression levels of genes and provides initial information for finding the gene regulatory network.

****

Data clustering is a process of grouping input data points with similar features in the multidimensional space; the algorithms are being investigated for long time. The most common hierarchical

clustering method often used by biologists for data clustering is dendrogram [**?**]. At the end of the analysis the data points are arranged into a phylogenetic tree, the level of similarity of two pairs is represented by the length of the branch. However, even though the hierarchical clustering is simple and straight forward, it is designed to reflect true hierarchical tree structure and that is not the way in which microarray data is generated. It is very important to include more biological information rather than rigidly clustering data points. Hierarchical clustering may fail to group data points in the right way because it is greatly influenced by local condition and has no opportunity of evaluating the global structure. Support Vector Machines????? (SOM) attempts to search for relevant patterns by first imposing structure on the data with nodes that are expected to eventually move to the center of each cluster, and then updating the structure map in each iteration based on a data point randomly selected from the data set [**?**]. The result ends up gathering similar samples in the same cluster. Studies of using SOM to cluster genes have been done by Whitehead Institute/MIT [**?**]. Under unsupervised situation, the success of this approach partially depends on the initialization of the map structure, e.g. number of nodes and different geometries. Without data modelling, SOM lacks criteria for validation of cluster structure, e.g. whether the number of clusters is optimal.

2. OPTIMAL PROJECTION FOR 2-D OR 3-D VISUALIZATION

The purpose of cluster analysis is to determine whether there are certain number of well-defined data sets within the entire data distribution and/or derive most rational and optimal grouping scheme to partition data into a specified number of clusters.

Since a gene expression microarray data set is a mixture of samples of cancer and non-cancer, or a mixture of samples of various types of cancers, the SFNM model may be the best approach for describing such multi-modal data structure [**?**]. In the case that k clusters exist in the data set, a mixture model with k normal distributions can be used to describe the overall distribution of the data. We will also estimate the density parameters of each cluster and the overall mixture.

Given knowledge of patient classification,

(1) Preprocessing Phase: which subset of the 7129 genes provides the most accurate/distinct classification?

(2) Postprocessing Phase: which projection of this subset of genes yields the most separation between classes? How to group genes effectively.

2.1. Preprocessing Step:

in unsupervised approach, Fishers Linear Discriminant has been adopted for finding the most discrminant projection.

However, for higher dimensional (2-D or 3-D) visualization

the traditional Fisher approach suffers from the fact that

is only one-dimensional

there is a need to establish a new method for

can be conveniently computed given the classification information.

In contrast to the unsupervised approach, Fishers Linear Discriminant can be conveniently computed given the classification information.

Given two classes, K1 and K2 , find the projection, wf, that maximizes J(wf), the ratio of inter-cluster distance to intra-cluster variance.

wopt = Sw-1 (m1 -m2)

Fisher Discriminant by Individual Gene

$$w = [00..10..0]^T$$

2.2. PostProcessing Step: Optimal Projection

Given n genes (found in Step I) and 72 patients with class labels, find best projection ? Fishers method

Traditional Scheme:

1. (projection onto top 3 ind. Fisher genes)
2. (Projection onto top 1-D DCA-selected genes)

Projection onto top DCA-selected representation, a representative feature stands for a closely related gene groups. (See double clustering section.)

(Projection onto top 2-D or 3-D what??? space Fisher pseudo-genes)

Separability vs. Number of Genes Cross-validation results

1. Fisher criterion
2. Classification Results (Cross-Validation)

MIT Data, Princeton Data

3. FIND OPTIMAL PROJECTION FOR 2-D OR 3-D VISUALIZATION

Typical approaches to grouping the genes adopt either the hierarchical VQ method or topologically sensitive clustering approach such as SOM.

The super high dimensionality (500 $\sim$ 8000) of microarray data introduce difficulties in the revelation of data structure,

We believe that the consideration of introducing user interaction into the clustering algorithm is a more practical approach, which greatly reduces both computational complexity and local optimum likelihood [**?**][**?**]. A user-friendly graphical interface for data visualization purpose is developed to allow the user to select initial centers of the data clusters. To visualize data, we further developed data projection methods based on the current methods used in [**?**] in order to maximize the revelation of cluster structure.

We proposed statistically-principled, GMM model-supported, and visually-insightful technique.

Notations:

for each cluster, m and C are the mean vector and covariance matrix, respectively, m is the relative mass, g is the gaussian kernel, K1 is the cluster number identifiable at top level, K2, k is the cluster number identifiable at second level.

3.1. Review of PCA and DCA projection methods

3.1.1. PCA

PCA is an effective unsupervised method for achieving dimensionality reduction [**?**, **?**, **?**]. For a set of observed d-dimensional data vectors $\{\mathbf{t}_i\}$, $i \in \{1, ..., N\}$, the q principal axes $\mathbf{w}_m$, $m \in \{1, ..., q(\leq d)\}$, are those orthogonal axes onto which the retained variances under projection are maximal. It can be shown that the

principal axes $\mathbf{w}_m$ are given by the q dominant eigenvectors (i.e., q maximal eigenvalues) of the sample covariance matrix

$$\mathbf{C_t}=\frac{1}{N}\sum_{i=1}^{N}(\mathbf{t}_i-\mu_\mathbf{t})(\mathbf{t}_i-\mu_\mathbf{t})^{\mathrm{T}} \quad (1)$$

such that

$$\mathbf{C_t}\mathbf{w}_m=\lambda_m\mathbf{w}_m \quad (2)$$

where $\mu_\mathbf{t}$ is the sample mean and λ_m is the eigenvalue. The vector $\mathbf{x}_i=\mathbf{W}^{\mathrm{T}}(\mathbf{t}_i-\mu_\mathbf{t}$, where $\mathbf{W}=\mathbf{w_1},\mathbf{w_2},...,\mathbf{w_q}$, is thus a q dimensional new representation of the observed vector $\mathbf{t}_i$. Two issues contribute to the limitations of the conventional PCA: its global linearity without incorporating data structure; and its optimality based on reconstruction error rather than pattern separability.

3.1.2. DCA

Discriminatory Data Projection

The purpose of developing discriminatory data projection tools is to maximally discover hidden cluster structure in the data space. The consideration of using multiple data projection tools is primarily based on the fact that the performance of the individual projection scheme tends to be case-dependent due to limited number of data samples in nearly all existing microarray data.

It is insufficient to use the PCA projection tool, which can not effectively taken the teacher's information in to account. The discriminatory projection tools presented in this paper is an extension of DCA,

****

If class information is known, the search of directions in data space for discovering cluster structure is under better guidance. There are two types of class information we may be able to obtain: known phenotypes from biological experimental setting, and subcluster information resulting from cluster decomposition based on an unsupervised projection (PCA or PPM). For the top level projection, DCA is a supervised process by using the known phenotype(class) information in the search of projection. However, DCA can also be used in an unsupervised situation that is on the sub-levels by using the second type of class information discussed above. Demonstrations of different applications of DCA are shown in the Result Section.

When confronting a multi-modal data set, however, is to emphasize the inter-cluster separation by replacing the total covariance matrix with the Fisher's scatter matrix [**?**, **?**], i.e , to find the eigenvectors of $\mathbf{S}_w{}^{-1}\mathbf{S}_b$.

$$\mathbf{S}_w^{-1}\mathbf{S}_b\mathbf{w}_m=\lambda_m\mathbf{w}_m \quad (3)$$

where the *within-cluster scatter matrix* ($\mathbf{S}_w$) is the *joint* scatter of data point $\mathbf{t}_i$ around the conditional mean vector $\mu_{\mathbf{t}k}$ of K_D classes (on the top level) or sub-clusters (on the sub-levels)

$$\mathbf{S}_w=\sum_{k=1}^{K_D}\pi_k\mathbf{C}_{\mathbf{t}k} \quad (4)$$

with cluster conditioned covariance matrix

$$\mathbf{C}_{\mathbf{t}k}=\frac{\sum_{i=1}^{N}z_{ik}(\mathbf{t}_i-\mu_{\mathbf{t}k})(\mathbf{t}_i-\mu_{\mathbf{t}k})^{\mathrm{T}}}{\sum_{i=1}^{N}z_{ik}} \quad (5)$$

where

$$z_{ik}=\frac{\pi_k g(\mathbf{t}_i|\mu_{\mathbf{t}k},\mathbf{C}_{\mathbf{t}k})}{p(\mathbf{t}_i)}, \quad (6)$$

and the *between-cluster scatter matrix* ($\mathbf{S}_b$) is the scatter of the cluster conditional mean vector $\mu_{\mathbf{t}k}$ around the overall data center $\mu_\mathbf{t}$

$$\mathbf{S}_b=\sum_{k=1}^{K_D}\pi_k(\mu_{\mathbf{t}k}-\mu_\mathbf{t})(\mu_{\mathbf{t}k}-\mu_\mathbf{t})^{\mathrm{T}} \quad (7)$$

such that the separability of patterns is maximized, that is

$$\mathbf{W}=\arg\max_{\mathbf{W}_0}\{Trace(\mathbf{W}_0^{\mathrm{T}}\mathbf{S}_w^{-1}\mathbf{S}_b\mathbf{W}_0)\}. \quad (8)$$

This is termed as Discriminatory Component Analysis.

The original vectors $\{\mathbf{t}_i\}$ are linearly transformed by $\mathbf{W}$, a $d\times 2$ matrix, through $\mathbf{x}=\mathbf{W}^{\mathrm{T}}(\mathbf{t}-\mu_\mathbf{t})$ into a two-dimensional projection space $\mathbf{x}=(x_1,x_2)^{\mathrm{T}}$. For a pure normal distribution over the data space:

$g(\mathbf{t}|\mu_\mathbf{t}k,\mathbf{C}_{\mathbf{t}k})$

This is optimal under such stochastic single cluster Gausssian modeling for each class of data.

However, the conventional Fisher DCA is an optimal Bayesian classifier only for such statistical cluster model.

However, most DNA gene expression data are much better modelded via GMM (or SFNM) model, [**?**, **?**]

When such 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 [**?**]. In other words, even though the cluster structure of a data set may be evident from the higher dimensional space, it is quite conceivable to have the finer cluster patterns concealed after a single linear projection, leading to an unidentifiable correspondence between Eq. (**??**) and Eq. (**??**) [**?**].

A possible approach is to model high-dimensional multi-modal data set with a hierarchical or non-hierarchical mixture model and accordingly with a collection of probabilistic principal discriminatory subspaces [**?**, **?**, **?**, **?**], namely the exploratory cluster discovery.

Assume a top-level model consisting of a single *Radon* transform $\mathbf{W}$ and a mixture of $K_1(<K_0)$ normal distributions $p(\mathbf{t})=\sum_{k=1}^{K_1}\pi_k g(\mathbf{t}|\mu_{\mathbf{t}k},\mathbf{C}_{\mathbf{t}k})$ which is identifiable in **x**-space, *i.e.*, $f(\mathbf{x})=\sum_{k=1}^{K_1}\pi_k g(\mathbf{x}|\mu_{\mathbf{x}k},\mathbf{C}_{\mathbf{x}k})$, we can form a two-level hierarchy by associating a group of SFNM sub-models with each model k at top-level

$$p(\mathbf{t})=\sum_{k=1}^{K_1}\pi_k\sum_{j=1}^{K_{2,k}}\pi_{j|k}g(\mathbf{t}|\mu_{\mathbf{t}(k,j)},\mathbf{C}_{\mathbf{t}(k,j)}) \quad (9)$$

where $\pi_{j|k}$ again corresponds to a set of mixing proportions, one for each k, with $0\le\pi_{j|k}\le 1$ and $\sum_j\pi_{j|k}=1$, and $\sum_{k=1}^{K_1}K_{2,k}=K_0$. To reveal the hidden cluster pattern within each model k at top-level, *i.e.*, $g(\mathbf{t}|\mu_{\mathbf{t}k},\mathbf{C}_{\mathbf{t}k})=\sum_{j=1}^{K_{2,k}}\pi_{j|k}g(\mathbf{t}|\mu_{\mathbf{t}(k,j)},\mathbf{C}_{\mathbf{t}(k,j)})$, an associated probabilistic principal discriminative subspace is constructed that focuses on the separability of patterns within the data portion defined by model k, where the opaque degree of a data point in the subspace plot is proportional to its posterior probability belonging to this model, *i.e.*, z_{ik} determined at top-level.

The further cluster discovery is a two-stage procedure: a soft partitioning of each model k into $K_{2,k}$ sub-clusters followed by a construction of corresponding subspace. Instead of assigning each given data point exclusively to one subspace, the contribution to its

generation is shared among all the subspaces. The subspaces for the sub-models at second-level are generated by the probabilistic DCA such that

$$\mathbf{S}_{k,w}^{-1}\mathbf{S}_{k,b}\mathbf{w}_{k,m} = \lambda_{k,m}\mathbf{w}_{k,m} \tag{10}$$

where $\mathbf{S}_{k,w} = \sum_{j=1}^{K_{2,k}} \pi_{j|k}\mathbf{C}_{\mathbf{t}(k,j)}$ with subcluster conditioned covariance matrix

$\mathbf{C}_{\mathbf{t}(kj)} = \sum_{i=1}^{N} z_{i(k,j)}(\mathbf{t}_i - \mu_{\mathbf{t}(kj)})(\mathbf{t}_i - \mu_{\mathbf{t}(kj)})^{\mathrm{T}} / \sum_{i=1}^{N} z_{i(k,j)}$,
$z_{i(k,j)} = z_{ik}\pi_{j|k} g(\mathbf{t}_i|\mu_{\mathbf{t}(kj)},\mathbf{C}_{\mathbf{t}(kj)}) / g(\mathbf{t}_i|\mu_{\mathbf{t}k},\mathbf{C}_{\mathbf{t}k})$,
and $\mathbf{S}_{k,b} = \sum_{j=1}^{K_{2,k}} \pi_{j|k}(\mu_{\mathbf{t}(kj)} - \mu_{\mathbf{t}k})(\mu_{\mathbf{t}(kj)} - \mu_{\mathbf{t}k})^{\mathrm{T}}$.

4. CONCLUSION

The user-friendly graphical interface facilitates the data visualization purpose, which allows the user to select initial centers of the data clusters.

The technique makes every subcluster/subspace be discriminatively explored individually so that local cluster structure is effectively revealed.

Although the final SFNM model can be estimated, the pathways of achieving cluster decomposition may be multiple. This user-driven nature of the current algorithm is also highly appropriate for the visualization context.

GenomeChip could be naturally extended to also include the increasingly important and feasible protein chip technology. In summary, the DNA microarray technology represents an enormous business opportunity and at the same time a great challenges in novel methods for DNA gene expression data analysis.

5. ACKNOWLEDGEMENTS

The authors wish to thank Dr. Joesph Yue Wang Virginia Tech. and Dr. Zuyi Wang Childrens National Medical Centre for insightful discussions, and thank Whitehead Institute, MIT, and Princeton Oncology dept. for the leukemia and colon cacer microarray data published in their web site.

6. REFERENCES

[1] D. J. Duggan, M. L. Bittner, Y. Chen, P. Meltzer, and J. M. Trent, "Expression profiling using cDNA microarrays," *Nature Genetics*, vol. 21, pp. 10-14, Jan. 1999.

[2] U. Scherf, D. T. Ross, M. Waltham, L. H. Smith, J. K. Lee, L. Tanabe, K. W. Kohn, W. C. Reinhold, T. G. Myers, D. T. Andrews, D. A. Scudiero, M. B. Eisen, E. A. Sausville, Y. Pommier, D. Botstein, P. O. Brown, and J. N. Weinstein, "A gene expression database for the molecular pharmacology of cancer," *Nature Genetics*, vol. 24, pp. 236-244, Mar. 2000.

[3] M. Bittner, P. Meltzer, Y. Chen, Y. Jiang, E. Seftor, M. Hendrix, M. Radmacher, R. Simon, Z. Yakhinl, A. Ben-Dor, N. Sampas, E. Dougherty, E. Wang, F. Marincola, C. Gooden, J. Lueders, A. Glatfelter, P. Pollock, J. Carpten, E. Gillanders, D. Leja, K. Dietrich, C. Beaudry, M. Berens, D. Alberts, V. Sondak, N. Hayward, and J. Trent, "Molecular classification of cutaneous malignant melanoma by gene expression profiling," *Nature*, vol. 406, no. 3, pp. 536-540, August 2000.

[4] H. Zhang, C-Y. Yu, B. Singer, and M. Xiong, "Recursive partitioning for tumor classification with gene expression microarray data," *Proc. Natl. Acad. Sci.*, vol. 98, no. 12, pp. 6730-6735, June 2001.

[5] 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.

[6] J. Khan, J. S. Wei, M. Ringner, L. H. Saal, M. Lananyi, F. Westermann, F. Berthold, M. Schwab, C. R. Antonescu, C. Peterson, and P. S. Meltzer, "Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks," *Nature Medicine*, vol. 7, no. 6, pp. 673-679, June 2001.

[7] P. Tamayo, D. Slonim, J. Msirov, et.al, "Interpreting pattern of gene expression with self-organizing maps: methods and application to hematopoietic differentiation", Proc. Natl. Acad. Sci., Vol. 96, pp. 2907-2912, March 1999.

[8] E. Hartuv, A. O. Schmitt, L. Lange, S. Meier-Ewert, H. Lehrach, and R. Schamir, "An algorithm for clustering cDNA fingerprints", Genomics, vol. 66, pp. 249-256, 2000.

[9] A. Ben-Hur, D. Horn, H. T. Siegelmann, and V. Vapnik, "Support vector clustering", J. Machine Learning Research, vol. 2, pp. 125-137, 2001.

[10] A. K. Jain, R. P. W. Duin, and J. Mao, "Statistical pattern recognition: a review," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 22, no. 1, pp. 4-37, Jan. 2000.

[11] E. Mjolsness and D. DeCoste, "Machine learning for science: state of the art and future prospects," *Science*, vol. 293, pp. 2051-2055, Sept. 2001.

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

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

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

[15] B. Ripley, *Pattern Recognition and Neural Networks*, Cambridge Univ. Press, 1996.

[16] 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. Neural Nets*, vol. 11, no. 3, pp. 625-636, May 2000.

[17] Y. Wang, J. Lu, and Z. Wang, et al., "Discriminative mining of gene microarray data", *Proc. of IEEE Neural Network for Signal Processing Workshop*, pp. 23-32, Sept. 2001.

[18] S. T. Roweis and L. K. Saul, "Nonlinear dimensionality reduction by locally linear embedding," *Science*, vol. 290, pp. 2323-2326, Dec. 2000.

[19] R. N. Bracewell, *Two-Dimensional Imaging*, Prentice-Hall, Inc., 1995.

[20] J. H. Friedman, "Exploratory projection pursuit," *J. Ame. Stat. Asso.*, vol. 82, no. 397, pp. 249-266, Mar. 1987.