

COMPUTATIONAL INTELLIGENCE APPROACH FOR GENE EXPRESSION DATA MINING AND CLASSIFICATION

Zuyi Wang^a Sun-Yuan Kung^b Junying Zhang^a Javed Khan^c Jianhua Xuan^a Yue Wang^a

^a Department of Electrical Engineering, The Catholic University of America, Washington, DC 20064

^b Department of Electrical Engineering, Princeton University, Princeton, NJ 08544

^c Advanced Technology Center, National Institutes of Health, Gaithersburg, MD 20877

ABSTRACT

The exploration of high dimensional gene expression microarray data demands powerful analytical tools. Our data mining software, VISual Data Analyzer (VISDA) for cluster discovery, reveals many distinguishing patterns among gene expression profiles. The model-supported hierarchical data exploration tool has two complementary schemes: discriminatory dimensionality reduction for structure-focused data visualization, and cluster decomposition by probabilistic clustering. Reducing dimensionality generates the visualization of the complete data set at the top level. This data set is then partitioned into subclusters that can consequently be visualized at lower levels and if necessary partitioned again. These approaches produce different visualizations that are compared against known phenotypes from the microarray experiments. For class prediction on cancers using microarray data, Multilayer Perceptrons (MLPs) are trained and optimized, whose architecture and parameters are regularized and initialized by weighted Fisher Criterion (wFC)-based Discriminatory Component Analysis (DCA). The prediction performance is compared and evaluated via multifold cross-validation.

1. INTRODUCTION

Microarrays have been used to find specific gene expression patterns in cancer cells to investigate the cause of various types of cancers. In a phenotype microarray data set, the representation of each sample's profile is described as a point in a d -dimensional gene expression space in which each dimension represents the expression level of one gene, the dimensionality can reach tens of thousands. The representations of phenotype-related samples form clusters. Discovering such cluster structure demands powerful data analytic tools that can handle large scale and high dimensional data, and the tools are not well developed [2], [4]. Another application of gene expression microarray is to examine expression changes of different genes over time due to the specific drug treatment, which is recorded in a temporal data set.

We propose a framework of hierarchical data exploration that aims at finding existing cluster structure, *i.e.*, cluster discovery, in high dimensional data. Our approaches for cluster discovery mainly include [2]: (1) statistical modeling of the gene microarray data using a Standard Finite Normal Mixture (SFNM) distribution; (2) dimensionality reduction and data visualization using multi-schematic unsupervised learning processes, including Principal Component Analysis (PCA), combined PCA and Projection Pursuit Method (PPM) – PCA-PPM, and Discriminatory Component Analysis; (3) user intervened model selection and probabilistic Expectation-Maximization (EM) clustering guided and validated by Minimum Description Length (MDL); (4) hierarchical data exploration scheme.

The study on cluster structure in phenotype microarray data greatly benefits molecular cancer diagnosis that is based on identified cancer genetic profiles. In this paper, we study class prediction that refers to the assignment of particular tumor samples to already-defined classes. Based on phenotype-known training samples, a classifier is designed and trained, mostly through supervised learning, and later used to classify future samples. Both conventional biostatistical predictors and neural network classifiers [1] have been proposed for class prediction with relative success. The ability of Multilayer Perceptrons (MLPs) to learn complex (non-linear) and multidimensional mapping from a collection of examples makes them ideal classifiers for class prediction. In this paper, we propose an optimized MLP, whose hidden nodes/weights correspond initially to the top eigenvectors and cluster-centers from the cluster structure-focused dimensionality reduction method DCA, and clustering, thus it offers an integrated procedure for gene extraction and classification.

2. METHODS AND RESULTS

2.1. Hierarchical Data Exploration Tool – VISDA

Assume that, in gene expression space t -space, a collection of N d -dimensional sample points $\{t_i\}$, forming K_0

clusters. The joint probability density function (pdf) of a d -dimensional multivariate SFNM model is a sum of individual normal distributions with their own density parameters while each gene is considered as a random variable,

$$p(\mathbf{t}_i) = \sum_{k=1}^{K_0} \pi_{\mathbf{t}k} g(\mathbf{t}_i | \boldsymbol{\mu}_{\mathbf{t}k}, \mathbf{C}_{\mathbf{t}k}) \quad (1)$$

where $\pi_{\mathbf{t}k}$ is the corresponding mixing proportion, with $0 \leq \pi_{\mathbf{t}k} \leq 1$ and $\sum \pi_{\mathbf{t}k} = 1$, and g is the Gaussian kernel, and $\boldsymbol{\mu}_{\mathbf{t}k}$ and $\mathbf{C}_{\mathbf{t}k}$ are the mean vector and covariance matrix of cluster k respectively. One natural criterion used for the modeling is the Maximum Likelihood (ML) estimation using the Expectation-Maximization (EM) algorithm [2], [4].

To address the problem of high dimensionality with microarray data, dimensionality reduction is intensively studied and well implemented. Principal component analysis is an effective unsupervised method for achieving dimensionality reduction. A set of top q principal axes \mathbf{w}_m , $m = 1 \dots q$, are orthogonal axes onto which the retained variances under projection are maximal. However, its global linearity without considering data structure does not support expected pattern separability-focused dimensionality reduction.

In order to find low dimensional projections that can provide the most revealing view of the data structure, we try to re-select the principal components from PCA based on a different criterion, *i.e.*, a non-Gaussianity measure, *kurtosis*, given by $kurt(m) = \frac{E((Y_m - \mu_{Y_m})^4)}{(E((Y_m - \mu_{Y_m})^2))^2} - 3$, where Y_m is a random variable consisting of the projections of all data points onto eigenvector \mathbf{w}_m derived from PCA, and $\mu_{Y_m} = E(Y_m)$. Kurtosis is one way to measure a type of departure from normality, and it ranges from -1 to 1 . If the distribution is normal, the value of kurtosis is equal to zero. A positive kurtosis indicates a more super-Gaussian (more peaked than the normal distribution); and a negative one indicates a sub-Gaussian (flatter than the normal). The principal components where the distribution of the projected data set is least Gaussian may contain plentiful structural information, and they will be chosen. This procedure is termed as PCA/PPM [2].

When class information is prior known or previously estimated, the search of projections for discovering cluster structure is under better guidance. An effective way is to emphasize the inter-cluster separation by using weighted Fisher criterion that is optimal for multi-class cases *i.e.*, to find the eigenvectors of

$$\mathbf{w} \mathbf{F} \mathbf{C} = \mathbf{S}_w^{-1} \sum_{k=1}^{K_0-1} \sum_{l=k+1}^{K_0} \pi_{\mathbf{t}k} \pi_{\mathbf{t}l} \varpi(\Delta_{kl}) \mathbf{S}_{kl} \quad (2)$$

where $\pi_{\mathbf{t}k}$ is the mixing proportion, the pairwise between-cluster scatter matrix $\mathbf{S}_{kl} = (\boldsymbol{\mu}_{\mathbf{t}k} - \boldsymbol{\mu}_{\mathbf{t}l})(\boldsymbol{\mu}_{\mathbf{t}k} - \boldsymbol{\mu}_{\mathbf{t}l})^T$, and within-cluster scatter matrix $\mathbf{S}_w = \sum_{k=1}^{K_0} \pi_{\mathbf{t}k} \mathbf{C}_{\mathbf{t}k}$, weighting function $\varpi(\Delta_{kl}) = \frac{1}{2\Delta_{kl}^2} \operatorname{erf}\left(\frac{\Delta_{kl}}{2\sqrt{2}}\right)$, where the distance

between centers $\Delta_{kl} = \|(\boldsymbol{\mu}_{\mathbf{t}k} - \boldsymbol{\mu}_{\mathbf{t}l})\|$, such that the separability of patterns is maximized, that is

$$\mathbf{W} = \arg \max_{\mathbf{W}} \{tr(\mathbf{W}^T \mathbf{w} \mathbf{F} \mathbf{C} \mathbf{W})\} \quad (3)$$

where \mathbf{W} is the set of eigenvectors. The motivation of the modification of Fisher criterion is to approximate Bayes error of pairs of classes into the separability measure for searching discriminative projections [2].

The application of DCA requires class information that is only available from the EM clustering. With the pre-estimated cluster means and covariance matrices, we can find a new data representation that emphasizes the separability of the already identified clusters. The data set is re-projected to the eigenvectors \mathbf{W} resulting from DCA to show and confirm the cluster structure. As a step for further cluster discovery, we believe that the importance of the separability analysis cannot be overstressed [2].

Using the dimensionality reduction tools, we can project the data set into a 2-D space, \mathbf{x} -space, through a linear transformation, $\mathbf{x}_i = \mathbf{W}^T \mathbf{t}_i$, where $\mathbf{W} = (\mathbf{w}_1, \mathbf{w}_2)$ is a set of top two most significant principal components resulting from PCA or PCA/PPM or DCA based on their specific principles. With such a perceptible data representation, the user is enabled to initialize cluster centers, which is a step included in the model selection procedure that refers to determination of structural information, *i.e.*, number of clusters, *e.g.*, K_1 for top level. In the 2-D \mathbf{x} -space, the user can generate several candidate models. With each model estimated by 2-D EM clustering, the optimal model with K components (clusters) can be determined by minimizing [2], [4],

$$\text{MDL}(K) = -\log(\mathcal{L}_{ML}) + 0.5(6K - 1) \log N \quad (4)$$

where \mathcal{L}_{ML} is the corresponding likelihood function.

As a general approach to maximum likelihood, EM algorithm is an iterative method for estimating mixture models. By the nature of EM, it ‘‘softly’’ partitions the data set to allow each data point to contribute simultaneously to all identified clusters through estimating z_{ik} , the conditional probability of sample i belonging to component (cluster) k . The EM Clustering procedure first takes place in the 2-D \mathbf{x} -space for two reasons [2], [4]: (1) estimate the 2-D mixture model for model selection; (2) remedy the problem of heavy loaded EM in high dimensional space by estimating model parameters first in 2-D \mathbf{x} -space and then fine tuning in \mathbf{t} -space. After the model selection, the chosen number of clusters along with the corresponding cluster parameters are used to initialize the EM in \mathbf{t} -space.

The strategy of the hierarchical data exploration and mining tool is that top level model and projection should discover as much information about the global picture of the data set, while lower level models explain the local and

internal structure between and within clusters, which may not be obvious in the high level models. With complementary mixture models and visualization projections, each level will be relatively simple while the complete hierarchy maintains overall flexibility yet still conveys considerable cluster information [2], [4]. The sub-level dimensionality reduction, mixture modeling and clustering focus on the each identified cluster/sub-cluster in a probabilistic and discriminative way.

At a second level, for model k found at the top level, we can further form a mixture of $K_{2,k}$ normal distributions, *i.e.*, an SFNM sub-model, a new complete mixture model is now formulated as,

$$p(\mathbf{t}_i) = \sum_{k=1}^{K_1} \pi_{\mathbf{t}k} \sum_{j=1}^{K_{2,k}} \pi_{\mathbf{t}(j|2,k)} g(\mathbf{t}_i | \boldsymbol{\mu}_{\mathbf{t}(2,k,j)}, \mathbf{C}_{\mathbf{t}(2,k,j)}) \quad (5)$$

where $\pi_{\mathbf{t}(j|2,k)}$ is the mixing proportion of the j th sum-model in the k th model at the top level with $0 \leq \pi_{\mathbf{t}(j|2,k)} \leq 1$, $\sum_{j=1}^{K_{2,k}} \pi_{\mathbf{t}(j|2,k)} = 1$, and $\sum_{k=1}^{K_1} K_{2,k} = K_0$, and the second level model $g(\mathbf{t} | \boldsymbol{\mu}_{\mathbf{t}(2,k,j)}, \mathbf{C}_{\mathbf{t}(2,k,j)})$ is the j th normal distribution within the cluster k , suppose all K_0 clusters are found at the second level. The second level data membership $z_{i(2,k,j)}$ given by,

$$z_{i(2,k,j)} = z_{ik} \frac{\pi_{\mathbf{t}(j|2,k)} g(\mathbf{t}_i | \boldsymbol{\mu}_{\mathbf{t}(2,k,j)}, \mathbf{C}_{\mathbf{t}(2,k,j)})}{g(\mathbf{t}_i | \boldsymbol{\mu}_{\mathbf{t}k}, \mathbf{C}_{\mathbf{t}k})} \quad (6)$$

with property $\sum_{k=1}^{K_1} \sum_{j=1}^{K_{2,k}} z_{i(2,k,j)} = 1$, and the density parameters $\boldsymbol{\mu}_{\mathbf{t}(2,k,j)}$, and $\mathbf{C}_{\mathbf{t}(2,k,j)}$ are estimated by sub-level EM [2], [4].

User interaction is also an important issue. We have developed a user-friendly graphical interface to facilitate the data visualization, which allows the user to select initial centers of perceivable clusters in \mathbf{x} -space, shown in Fig. 1. Our experience shows a great reduction of both computational complexity and local optimum likelihood [2], [4]. The completion of a hierarchy is determined by MDL and user visual inspection.

We test VISDA on a data set from National Cancer Institute (NCI) for small round blue cell tumors (SRBCTs) [1] including four subtypes EWS, RMS, BL, NB in Fig. 1, and a three-level hierarchy is constructed to fully explore the four-cluster data structure. The two-cluster structure within the cluster #2 at the top level is found at the second level which is not obvious at the top level. Note that the known class information is only used to represent the data points from various classes in different symbols and colors, but not in the clustering procedure.

2.2. Class Prediction

In our previous work [1], we have obtained promising results by using a linear single-layer perceptron to perform

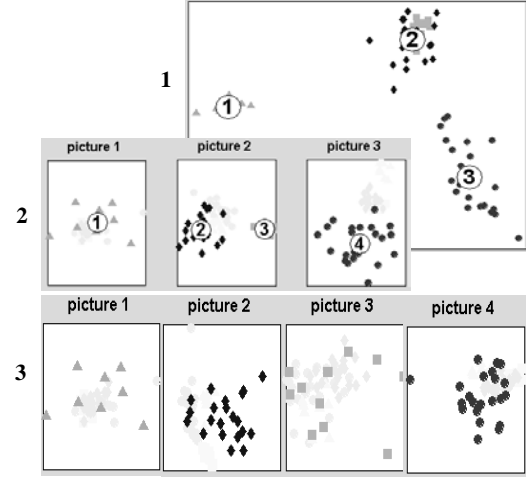


Fig. 1. The hierarchical data exploration on the NCI SRBCTs data set.

multi-class prediction on NCI SRBCTs data set. Since PCA is non-discriminatory, the resulting pseudo-genes do not guarantee to retain sufficient discriminatory power. While the MLPs can offer an integrated procedure for gene extraction and classification. We have developed MLP-based neural network classifier for the same task in [1].

In this experiment, we used $m_0 = 3 \sim 8$, $m_1 = 3 \sim 5$, and $m_2 = 4$ (*i.e.*, four outputs corresponding to the four subtypes). The values of m_0 and m_1 were determined by the number of the significant eigenvalues of the weighted Fisher criterion derived from Eq. 3. First, through a gene selection scheme that is also principled by weighted Fisher criterion, we selected top 60 and 150 genes as original inputs, and then the dimensionality was further reduced by wFC-DCA to $3 \sim 8$ gene extractions. The extracted $3 \sim 8$ dominant components contained more than 80% of the class separability for the four class in the gene expression space. Next, the 63 training samples were randomly partitioned into 3 groups (21 samples per group), among them one for validation and two for training. MLPs were then trained using standard back-propagation (BP) algorithm, in which the $3 \sim 8$ top wFC-DCA values were used as input and the cancer category as output for each sample. This process was repeated by rotating each of the 3 groups for validation. The samples were again randomly partitioned, and the entire process was repeated for 1250 times. For each shift of a validation group, one model was trained, resulting in a total of 3750 trained models. Our experiments with various MLP architectures have shown very satisfied classification performance. All cross-validations have achieved 0% misclassification rate and our experience has indicated that a ($m_0 = 3$, $m_1 = 3$, $m_2 = 4$) MLP is the most sufficient among tested

Table 1. Misclassification rates for the class prediction using MLPs on MIT ALL/AML data set.

Input	# of hidden nodes		
	1	2	3
Top 3	11.43%	14.29%	10%
Top 8	4.29%	2.86%	2.86%
Top 18	2.86%	1.43%	1.43%

ones. We have also tested leave-one-out cross-validation, and a perfect class prediction result was obtained.

We have also tested the class prediction on MIT's acute leukemia data set containing gene expressions of total 72 samples, among which 47 Acute Lymphoblastic Leukemia (ALL) and 25 Acute Myeloid Leukemia (AML) samples. We considered first-order linear correlation between the genes, and performed FC-DCA, finally reduced the dimension to 18 top pseudo-genes. Similar to the 3-fold cross-validation, a 5-fold cross-validation plan with four subsets for training and one for testing is applied. Note that two outputs are needed for two subtypes of leukemias in this case. Table 1 summarizes the misclassification rates for several different MLP configurations. The results show that the average probability of detection increased when we increased the number of dimension and the number of hidden nodes.

2.3. Temporal Microarray Data

Different from the phenotype microarray data, *e.g.*, NCI SRBCTs data set, an n -dimensional sample point in a temporal data set represents the expression change profile of one gene with each dimension is the expression level at a specific assaying time. Grouping genes with similar change pattern through clustering may help effectively identify distinct change profiles. To form an *in vivo* model of staged muscle regeneration, a 27-time point temporal microarray data set generated by Children's National Medical Center (CNMC) for determining potential downstream targets of MyoD that initiates the transcription of its downstream targets as a transcriptional factor [3]. Through gene clustering, eighteen clusters and their corresponding change profiles are identified by VISDA and presented in Fig. 2, and the results are used to identify the downstream targets Of MyoD [3].

3. CONCLUSION

With the techniques including, unsupervised discriminatory dimensionality reduction, probabilistic clustering, user interaction and hierarchical visualization, the developed hierarchical data exploration tool is proven to be able to effectively discover data structure in high dimensional space. The extraction of pseudo-genes through wFC-DCA enables

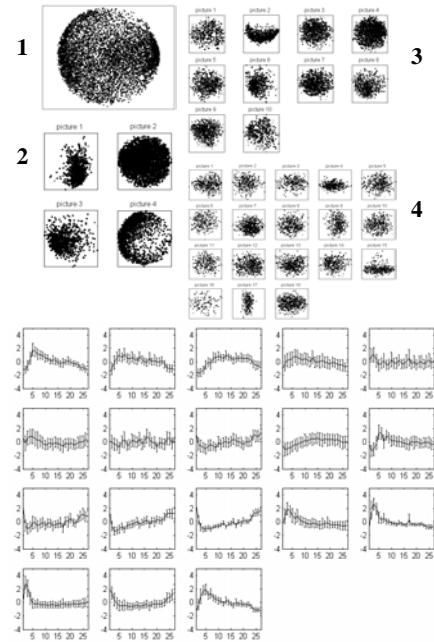


Fig. 2. The gene clustering on CNMC temporal microarray data set and the resulting gene expression change profiles.

the maximization of the class separability so that the MLP classifier can be optimized by taking the pseudo-genes as inputs.

4. REFERENCES

- [1] J. Khan, et. al, "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.
- [2] Z. Wang, et. al, "Discriminatory Mining of Gene Expression Microarray Data," *Journal of VLSI Signal Processing System*, in press, 2003.
- [3] P. Zhao, et. al, "In vivo filtering of in vitro MyoD target data: An approach for identification of biologically relevant novel downstream targets of transcription factors," *Comptes Rendus Biologies*, in press.
- [4] Y. Wang, et. al, "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.