

Clustering of High-Dimensional Gene Expression Data with Feature Filtering Methods and Diffusion Maps

Rui Xu¹, Steven Damelin², Boaz Nadler³, and Donald C. Wunsch II¹

¹Applied Computational Intelligence Laboratory

Dept. of Electrical & Computer Engineering, Missouri University of Science & Technology
Rolla, MO 65409-0249 USA

²School of Applied and Computational Mathematics, University of the Witwatersrand
Private Bag 3, Wits, 2050, South Africa

³Dept. of Computer Science and Applied Mathematics, Weizmann Institute of Science,
Rehovot, Israel

rxu@mst.edu, damelin@georgiasouthern.edu, boaz.nadler@weizmann.ac.il,
dwunsch@mst.edu

Abstract

The importance of gene expression data in cancer diagnosis and treatment by now has been widely recognized by cancer researchers in recent years. However, one of the major challenges in the computational analysis of such data is the curse of dimensionality, due to the overwhelming number of measures of gene expression levels versus the small number of samples. Here, we use a two-step method to reduce the dimension of gene expression data. At first, we extract a subset of genes based on the statistical characteristics of their corresponding gene expression measurements. For further dimensionality reduction, we then apply diffusion maps, which interpret the eigenfunctions of Markov matrices as a system of coordinates on the original data set in order to obtain efficient representation of data geometric descriptions, to the reduced data. A neural network clustering theory, Fuzzy ART, is applied to the resulting data to generate clusters of cancer samples. Experimental results on the small round blue-cell tumor (SRBCT) data set, compared with other widely-used clustering algorithms, demonstrate the effectiveness of our proposed method in addressing multidimensional gene expression data.

1. Introduction

In recent years, the importance of gene expression data from DNA microarrays in cancer diagnosis, together with its advantage over the traditional, morphological appearance-based cancer classification methods, by now has been widely recognized by cancer researchers [1-3]. In this context, different cancer types or subtypes are discriminated through their corresponding gene expression profiles.

One of the major challenges of microarray data analysis is the overwhelming number of measures of

gene expression levels compared with the small number of cancer samples. Specifically, the samples display different behaviors in only a few of the features (genes). On the other hand, on most of the features, the behavior of the different classes is roughly the same; thus, these features can be regarded as noise. From the computational point of view, the existence of numerous irrelevant and redundant features or non-informative genes not only increases the computational complexity, but impairs the effective discovery of the cancer clusters. In this sense, feature selection or extraction is critically important for dimensionality reduction and further analysis. Many methods have been proposed to address this problem [4-5]. However, most of these methods work in a supervised way, and the lack of training data intensifies the problem. We remark that in recent years various unsupervised methods to detect bi-clusters have also been developed, see [6-7] and references therein.

In our previous research [8-9], we used diffusion maps [10-11] to address the high-dimensional problem. Here, we show that the performance of diffusion maps can be further improved by removing those non-informative genes based on the statistical characteristics of their corresponding gene expression measurement, such as high correlation coefficient to other genes, large variance, and a bimodal probability density distribution. Unlike other supervised methods, no prior or label information for the samples is required. This assumption is reasonable with the requirement for discovering unknown and novel cancer types or subtypes. The reduced data are then clustered with a neural network cluster theory, Fuzzy ART (FA) [12], to generate a partition of the cancer samples. We compared the performance of the proposed methods with those of diffusion maps-FA, hierarchical clustering algorithms, and K -means - on the small round blue-cell tumor (SRBCT) data set [3]. The experimental results demonstrate the effectiveness of

our proposed method in addressing multidimensional gene expression data and ultimately identifying corresponding cancer types.

The remainder of this paper is organized as follows. Section II and III discuss diffusion maps and the feature selection methods, respectively. Section IV presents an introduction to FA. The experimental results are presented and discussed in section V, and section VI concludes the paper.

2. Diffusion maps

Given a data set $\mathbf{X}=\{\mathbf{x}_i, i=1, \dots, N\}$, where N is large enough and fixed, on an m -dimensional data space, where m is also large enough and fixed, a finite graph with N nodes corresponding to N data points can be constructed on \mathbf{X} as follows. Every two nodes in the graph are connected by an edge weighted through a non-negative, symmetric, and positive definite kernel $w: \mathbf{X} \times \mathbf{X} \rightarrow (0, \infty)$. An archetypal example is the Gaussian kernel, given by,

$$w(\mathbf{x}_i, \mathbf{x}_j) = \exp\left(-\frac{\|\mathbf{x}_i - \mathbf{x}_j\|^2}{2\sigma^2}\right), \quad (1)$$

where σ is the kernel width parameter. The kernel reflects the degree of similarity between \mathbf{x}_i and \mathbf{x}_j , and $\|\cdot\|$ is the Euclidean norm in \mathfrak{R}^m .

Let

$$d(\mathbf{x}_i) = \sum_{\mathbf{x}_j \in \mathbf{X}} w(\mathbf{x}_i, \mathbf{x}_j) \quad (2)$$

be the degree of \mathbf{x}_i ; the Markov or affinity matrix \mathbf{P} is then constructed by calculating each entry as

$$p(\mathbf{x}_i, \mathbf{x}_j) = \frac{w(\mathbf{x}_i, \mathbf{x}_j)}{d(\mathbf{x}_i)}. \quad (3)$$

From the definition of the weight function, $p(\mathbf{x}_i, \mathbf{x}_j)$ can be interpreted as the transition probability from \mathbf{x}_i to \mathbf{x}_j in one time step. This idea can be further extended by considering $p^t(\mathbf{x}_i, \mathbf{x}_j)$ in the t^{th} power \mathbf{P}^t of \mathbf{P} as the probability of transition from \mathbf{x}_i to \mathbf{x}_j in t time steps [10]. Therefore, the parameter t defines the granularity of the analysis. With the increase of the value of t , local geometric information of data is also integrated. The change in direction of t makes it possible to control the generation of more specific or broader clusters.

Since the matrix \mathbf{P} is adjoint to a symmetric matrix, its spectrum is composed of real eigenvalues and the corresponding right and left eigenvectors form a basis of \mathfrak{R}^N . Assuming the Markov matrix \mathbf{P} is irreducible (e.g., the graph is connected), its largest eigenvalue is 1 with multiplicity one. We denote the eigenvalues of \mathbf{P} by $1=\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_N$, and the corresponding right eigenvectors by $\{\boldsymbol{\varphi}_j, j=1, \dots, N\}$,

$$\mathbf{P}'\boldsymbol{\varphi}_j = \lambda_j' \boldsymbol{\varphi}_j. \quad (4)$$

As described in [13], given the definition of a random walk on the graph of the data, one can define two key concepts. The first is a *diffusion distance* between the nodes on the graph that captures their dynamical proximity as follows,

$$D_t(\mathbf{x}_i, \mathbf{x}_j) = \left\| p^t(\mathbf{x}_i, \cdot) - p^t(\mathbf{x}_j, \cdot) \right\|_{\phi_0}, \quad (5)$$

where ϕ_0 is the unique stationary distribution

$$\phi_0(\mathbf{x}) = \frac{d(\mathbf{x})}{\sum_{\mathbf{x}_i \in \mathbf{X}} d(\mathbf{x}_i)}, \mathbf{x} \in \mathfrak{R}^d, \quad (6)$$

The second is a *diffusion map*, which is a mapping of the nodes of the graph from the original data space into an L -dimensional Euclidean space \mathfrak{R}^L . This is done via the eigenvectors of \mathbf{P} , viewed as a new set of coordinates on the data set, as follows

$$\boldsymbol{\Psi}_t : \mathbf{x}_i \rightarrow (\lambda_1' \boldsymbol{\varphi}_1(\mathbf{x}_i), \dots, \lambda_L' \boldsymbol{\varphi}_L(\mathbf{x}_i))^T. \quad (7)$$

The relationship between the two concepts is that the Euclidean distance between all N eigenvector coordinates is equal to the diffusion distance,

$$D_t(\mathbf{x}_i, \mathbf{x}_j) = \left\| \boldsymbol{\Psi}_t(\mathbf{x}_i) - \boldsymbol{\Psi}_t(\mathbf{x}_j) \right\|, \quad (8)$$

where $\|\cdot\|$ is the Euclidean norm in \mathfrak{R}^L . The diffusion distance captures the dynamic proximity of nodes on the graph, since the more paths connect two nodes, the smaller their diffusion distance is. Further, since the eigenvalues of \mathbf{P} decay to zero, one can approximate the diffusion distance by relatively few eigenvector coordinates ($L \ll N$).

3. Gene filtering

As aforementioned, many features are non-informative for discrimination of cancer types. Removing these genes before applying diffusions is important. The main reason is that the distance between samples in the representation that takes all features into account contains a lot of noise and makes distances not very informative. For standard statistical methods of classification and regression, noise and high dimensional data have a detrimental effect leading to error terms of the form (variance)* d/N . If d/N is large, the error can be quite large [14]. In the rest of the section, we will discuss three methods for selecting informative genes, using the SRBCT data set as an example, which is introduced in Section V.

The correlation coefficient reflects the statistical measure of the strength of a linear relationship between variables. For gene expression data sets, there are usually quite a few genes that are highly correlated. Fig. 1 depicts four of the most correlated pairs of features of the SRBCT data set. It can be seen that all these

gene pairs can be beneficial for separation of one group of samples from the others. This also makes sense because we expect one class of samples to behave differently on more than a single feature. Therefore, it should not be surprising that such genes are correlated.

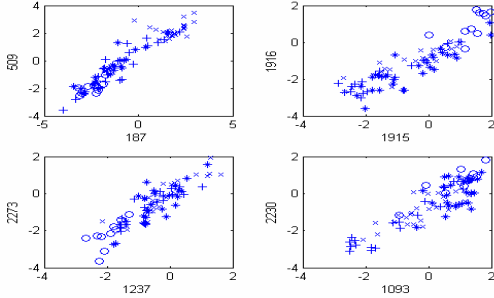


Fig. 1. Expression levels of four pairs of most correlated genes. The different point types correspond to the four categories.

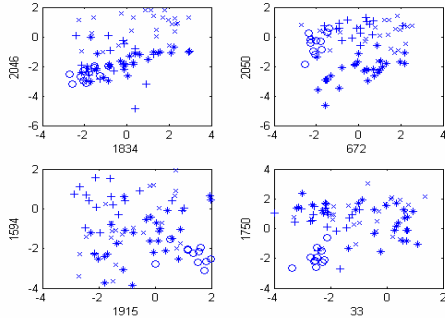


Fig. 2. Expression levels of four pairs of genes with the highest variance. The different point types correspond to the four categories.

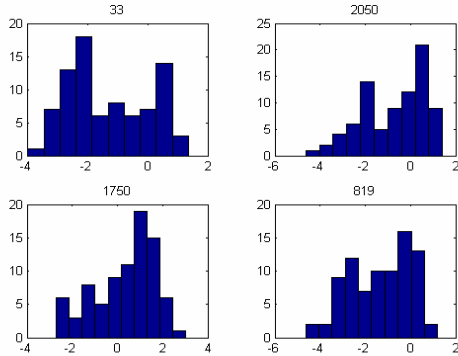


Fig. 3. Histogram of four genes that show double hump density.

Large variance is another important statistical method that indicates the usefulness of the corresponding feature in discriminating between different categories. Discriminating genes exhibit quite different behaviors in different cancer categories, leading to large variance. Fig. 2 shows a plot of 8 of

the features with the highest variance, grouped in pairs for visualization purposes. It is clear that these features are very informative in disclosing the category structure.

Fig. 3 illustrates the histogram of four features that display double hump densities, which look like the sum of two well-separated Gaussians. Such features are also useful in clustering.

4. Fuzzy ART

The basic FA architecture consists of two-layer nodes or neurons, the feature representation field F_1 , and the category representation field F_2 . The neurons in layer F_1 are activated by the input pattern, while the prototypes of the formed clusters are stored in layer F_2 . The neurons in layer F_2 that are already being used as representations of input patterns are said to be committed. Correspondingly, the uncommitted neuron encodes no input patterns. The two layers are connected via adaptive weights w_j , emanating from node j in layer F_2 . After an input pattern is presented, the neurons (including a certain number of committed neurons and one uncommitted neuron) in layer F_2 compete by calculating the category choice function

$$T_j = \frac{|\mathbf{x} \wedge \mathbf{w}_j|}{\alpha + |\mathbf{w}_j|}, \quad (9)$$

where \wedge is the fuzzy AND operator defined by

$$(\mathbf{x} \wedge \mathbf{y})_i = \min(x_i, y_i), \quad (10)$$

and $\alpha > 0$ is the choice parameter to break the tie when more than one prototype vector is a fuzzy subset of the input pattern, based on the winner-take-all rule,

$$T_j = \max_j \{T_j\}. \quad (11)$$

The winning neuron J then becomes activated, and an expectation is reflected in layer F_1 and compared with the input pattern. The orienting subsystem with the pre-specified vigilance parameter ρ ($0 \leq \rho \leq 1$) determines whether the expectation and the input pattern are closely matched. If the match meets the vigilance criterion,

$$\rho \leq \frac{|\mathbf{x} \wedge \mathbf{w}_j|}{|\mathbf{x}|}, \quad (12)$$

weight adaptation occurs, where learning starts and the weights are updated using the following learning rule,

$$\mathbf{w}_j(\text{new}) = \beta(\mathbf{x} \wedge \mathbf{w}_j(\text{old})) + (1 - \beta)\mathbf{w}_j(\text{old}), \quad (13)$$

where $\beta \in [0, 1]$ is the learning rate parameter. This procedure is called resonance, which suggests the name of ART. On the other hand, if the vigilance criterion is not met, a reset signal is sent back to layer F_2 to shut off the current winning neuron, which will remain disabled for the entire duration of the

Table 1. Performance results of diffusion maps and Fuzzy ART on the entire SRBCT data set. The corresponding ρ is indicated in the parentheses.

	$RI(\rho)$					
	$\sigma=22$	$\sigma=24$	$\sigma=26$	$\sigma=28$	$\sigma=30$	$\sigma=32$
$L=5$	0.7417 (0.5)	0.7661 (0.5)	0.7802 (0.5)	0.7761 (0.5)	0.7743 (0.45)	0.7708 (0.6)
$L=10$	0.8569 (0.3)	0.8260 (0.35)	0.8187 (0.45)	0.9019 (0.2)	0.8601 (0.3)	0.8760 (0.2)
$L=15$	0.8795 (0.35)	0.8290 (0.35)	0.8560 (0.35)	0.8431 (0.2)	0.8619 (0.3)	0.8322 (0.25)
$L=20$	0.8707 (0.25)	0.8346 (0.4)	0.8795 (0.35)	0.8284 (0.25)	0.8578 (0.4)	0.8160 (0.45)
$L=50$	0.8437(0.3)	0.8149 (0.35)	0.8175 (0.5)	0.8137 (0.55)	0.8354 (0.35)	0.8196 (0.6)

Table 2. Performance results of diffusion maps and Fuzzy ART on the reduced SRBCT data set with 30 genes. The corresponding ρ is indicated in the parentheses

	$RI(\rho)$					
	$\sigma=1$	$\sigma=2$	$\sigma=4$	$\sigma=6$	$\sigma=8$	$\sigma=10$
$L=5$	0.5498 (0.6)	0.9227 (0.5)	0.8490 (0.55)	0.8554 (0.45)	0.8110 (0.6)	0.8237 (0.35)
$L=8$	0.6100 (0.2)	0.8751 (0.45)	0.8598 (0.5)	0.8463 (0.2)	0.8516 (0.2)	0.8422 (0.45)
$L=10$	0.8149 (0.45)	0.9042 (0.2)	0.8554 (0.5)	0.8845 (0.3)	0.8422 (0.45)	0.8287 (0.35)
$L=15$	0.9124 (0.5)	0.9427 (0.3)	0.8842 (0.4)	0.9042 (0.35)	0.8516 (0.5)	0.7749 (0.4)

presentation of this input pattern, and a new competition is performed among the remaining neurons. This new expectation is then projected into layer F_1 , and this process repeats until the vigilance criterion is met. In the case that an uncommitted neuron is selected for coding, a new uncommitted neuron is created to represent a potential new cluster.

5. Experimental results

We applied the proposed method to the data set on the diagnostic research of small round blue-cell tumors (SRBCTs) of childhood. The SRBCT data set consists of 83 samples from four categories, known as Burkitt lymphomas (BL), the Ewing family of tumors (EWS), neuroblastoma (NB) and rhabdomyosarcoma (RMS) [3]. Gene expression levels of 2,308 genes were measured using cDNA microarray. The relative red intensity (RRI) of a gene is defined as the ratio between the mean intensity of that particular spot and the mean intensity of all filtered genes, and the ultimate expression level measure is the natural logarithm of RRI. In our further analysis, an additional logarithm was taken to linearize the relations between different genes and to lessen very high expression levels.

Because we already have a pre-specified partition \mathbf{H} of the data set, which is also independent from the clustering structure \mathbf{C} resulting from the use of a clustering algorithm, the performance can be evaluated by comparing \mathbf{C} to \mathbf{H} in terms of external criteria, such as the Rand index [15]. Considering a pair of tissue samples \mathbf{x}_i and \mathbf{x}_j , there are four different cases based on how \mathbf{x}_i and \mathbf{x}_j are placed in \mathbf{C} and \mathbf{H} .

- Case 1: \mathbf{x}_i and \mathbf{x}_j belong to the same clusters of \mathbf{C} and the same category of \mathbf{H} .
- Case 2: \mathbf{x}_i and \mathbf{x}_j belong to the same clusters of \mathbf{C} but different categories of \mathbf{H} .

- Case 3: \mathbf{x}_i and \mathbf{x}_j belong to different clusters of \mathbf{C} but the same category of \mathbf{H} .
- Case 4: \mathbf{x}_i and \mathbf{x}_j belong to different clusters of \mathbf{C} and a different category of \mathbf{H} .

Correspondingly, the number of pairs of samples for the four cases are denoted as a , b , c , and d , respectively. The Rand index used in our analysis can then be defined as follows:

$$R = (a + d) / (a + b + c + d); \quad (14)$$

As can be seen from the definition, the larger the values of R , the more similar are \mathbf{C} and \mathbf{H} .

Fig. 4 shows the best Rand index scores for diffusion maps and FA on the entire data set and the reduced data set with features selected by the methods discussed in Section III. Thirty genes were chosen with indexes listed as follows: $\text{idx}=[4, 33, 58, 107, 129, 187, 509, 672, 735, 819, 989, 1237, 1263, 1594, 1750, 1769, 1781, 1803, 1834, 1890, 1915, 1916, 2046, 2050, 2060, 2086, 2211, 2214, 2273, 2290]$. For the purpose of comparison, we also illustrate the best results with hierarchical clustering (HC) algorithm (single-linkage) and the K -means (KM) algorithm on both the entire data set and the data after diffusion maps are used. Here, the dendrogram of HC are cut at different levels to generate 2-10 clusters, respectively, and the value of K varies from 2 to 10. From the figure, it can be seen that filtering the features in advance can consistently improve performance. Also, diffusion maps are important in exposing the data structure; the performance of all three clustering algorithms without diffusion maps deteriorates dramatically, especially for the hierarchical clustering algorithm. Another observation from Fig. 4 is that FA can achieve better partitions of the given samples than the other two methods.

Tables 1 and 2 further summarize the results of diffusion maps and FA on the original and reduced data set. The dimensions of the transformed space are

chosen at 5, 10, 15, 20, and 50 when the entire data set is used and 5, 8, 10, and 15 when the 30-gene subset is used. For each designated dimension, we adjusted the kernel width parameter σ and vigilance parameter ρ . As shown in the tables, both parameters play an important role in the sample partition. However, there is still no effective criterion to decide these parameters, and their selection is based on cross validation. Again, the effectiveness of feature selection before applying diffusion maps is demonstrated.

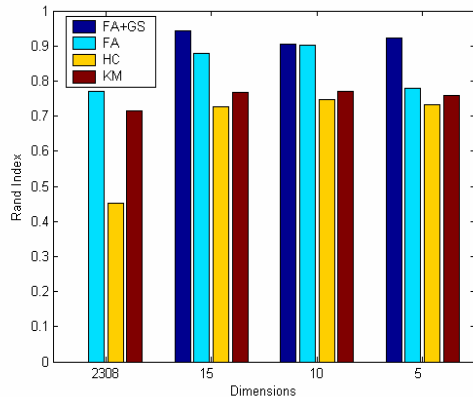


Fig. 4. The best clustering scores of Rand index for the SRBCT data set. The dimension is set as 2308, 15, 10, and 5, respectively. The order for the bars is FA with gene selection (GS), FA, HC, KM, from left to right.

6. Conclusions

Cancer classification based on gene expression profiles provides a promising method for cancer diagnosis and treatment. Here, we propose to use feature selection methods and diffusion maps to address the problem of high dimensions, a major challenge in gene expression data analysis. Fuzzy ART is then used to form the clusters of cancer samples. The experimental results on the SRBCT data set demonstrate the potential of the proposed methods in achieving useful information from the high-dimensional gene expression data.

Acknowledgment

Partial support for this research from the National Science Foundation, and from the M.K. Finley Missouri endowment, is gratefully acknowledged.

References

[1] T. Golub, D. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. Mesirov, H. Coller, M. Loh, J. Downing, M.

- Caligiuri, C. Bloomfield, and E. Lander, "Molecular classification of cancer: class discovery and class prediction by gene expression monitoring," *Science*, vol. 286, pp. 531-537, 1999.
- [2] R. Shyamsundar, Y. Kim, J. Higgins, K. Montgomery, M. Jorden, A. Sethuraman, M. van de Rijn, D. Botstein, P. Brown and J. Pollack, "A DNA microarray survey of gene expression in normal human tissues," *Genome Biology*, vol. 6, R22, 2005.
- [3] J. Khan, J. Wei, M. Ringnér, L. Saal, M. Ladanyi, F. Westermann, F. Berthold, M. Schwab, C. Antonescu, C. Peterson, and P. Meltzer, "Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks," *Nature Medicine*, vol. 7, pp. 673-679, 2001.
- [4] A. Ben-Dor, L. Bruhn, N. Friedman, I. Nachman, M. Schummer, and Z. Yakhini, "Tissue classification with gene expression profiles," *Proceedings of the Fourth Annual International Conference on Computational Molecular Biology*, pp. 583-598, 2000.
- [5] G. Cawley and N. Talbot, "Gene selection in cancer classification using sparse logistic regression with Bayesian regularization," *Bioinformatics*, vol. 22, no. 19, pp. 2348-2355, 2006.
- [6] A. Tanay, R. Sharan, and R. Shamir, "Biclustering algorithms, a survey", *Handbook of Computational Molecular Biology*, Edited by S. Aluru, Chapman, 2004.
- [7] S. Madeira and A. Oliveira, "Biclustering algorithms for biological data analysis: A survey". *IEEE Transactions on Computational Biology and Bioinformatics*, vol. 1, no. 1, pp. 24-45, 2004.
- [8] R. Xu, S. Damelin, and D. Wunsch II, "Applications of diffusion maps in gene expression data-based cancer diagnosis analysis," In *proceedings of the 29th Annual International Conference of IEEE Engineering in Medicine and Biology Society*, Lyon, France, August, 2007.
- [9] R. Xu, S. Damelin, and D. Wunsch II, "Clustering of cancer tissues using diffusion maps and fuzzy ART with gene expression data," in preparation.
- [10] R. Coifman and S. Lafon, "Diffusion maps," *Applied and Computational Harmonic Analysis*, vol. 21, pp. 5-30, 2006.
- [11] S. Lafon, Y. Keller, and R. Coifman, "Data fusion and multicue data matching by diffusion maps," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 28, no. 11, pp. 1784-1797, 2006.
- [12] G. Carpenter, S. Grossberg, and D. Rosen, "Fuzzy ART: fast stable learning and categorization of analog patterns by an adaptive resonance system," *Neural Networks*, vol. 4, pp. 759-771, 1991.
- [13] B. Nadler, S. Lafon, R. Coifman, and I. Kevrekidis, "Diffusion Maps, Spectral Clustering and Eigenfunctions of Fokker-Planck Operators," *Neural Information Processing Systems (NIPS)*, vol 18, 2005.
- [14] B. Nadler and R. Coifman, "The prediction error in CLS and PLS: The importance of feature selection prior to multivariate calibration," *Journal of Chemometrics*, vol. 19, no. 2, pp.107-118, 2005.
- [15] R. Xu and D. Wunsch II, "Survey of Clustering Algorithms," *IEEE Transactions on Neural Networks*, vol.16, no.3, pp.645-678, 2005.