01 Jun 2008

Clustering of Cancer Tissues Using Diffusion Maps and Fuzzy ART with Gene Expression Data

Rui Xu
Missouri University of Science and Technology

Steven Damelin

Donald C. Wunsch
Missouri University of Science and Technology, dwunsch@mst.edu

Recommended Citation

R. Xu et al., "Clustering of Cancer Tissues Using Diffusion Maps and Fuzzy ART with Gene Expression Data," *Proceedings of the 2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence)*, Institute of Electrical and Electronics Engineers (IEEE), Jun 2008.
The definitive version is available at https://doi.org/10.1109/IJCNN.2008.4633787

Clustering of Cancer Tissues Using Diffusion Maps and Fuzzy ART with Gene Expression Data

Rui Xu, Steven Damelin, and Donald C. Wunsch II

Abstract—**Early detection of a tumor's site of origin is particularly important for cancer diagnosis and treatment. The employment of gene expression profiles for different cancer types or subtypes has already shown significant advantages over traditional cancer classification methods. Here, we apply a neural network clustering theory, Fuzzy ART, to generate the division of cancer samples, which is useful in investigating unknown cancer types or subtypes. On the other hand, we use 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, for dimensionality reduction. The curse of dimensionality is a major problem in cancer type recognition-oriented gene expression data analysis due to the overwhelming number of measures of gene expression levels versus the small number of 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.**

I. Introduction

CANCERS of various types, for many decades, have been a leading cause of death in the world. For example, according to the report released by the National Center for Health Statistics in 2004, cancer accounts for 22.9% (550,270) deaths in the United States, only less than the number caused by heart diseases [1]. Given the tremendous complexity of various types of cancers, it is believed that the single most important indicator for surviving cancers is and will be early detection and, subsequently, early treatments. Early cancer diagnoses require accurately identifying the site of origin of a tumor. However, the traditional cancer classification methods that are largely dependent on the morphological appearance of tumors and parameters derived from clinical observations cannot meet such an expectation [2]. Their applications are limited by the existing uncertainties, and their prediction accuracy is very low. Tumors with similar appearances may have quite different origins and may therefore respond differently to the same treatment therapy. For example, diffuse large B-cell lymphoma (DLBCL), the most common type of lymphoma in adults, can only be cured by chemotherapy in 35-40 percent of patients due to the existence of unknown subtypes that cannot be discriminated based only on their morphologic parameters [3].

Fortunately, the recently-developed DNA microarray technologies [4-5], which can measure the expression levels of tens of thousands of genes simultaneously, offer cancer researchers a novel method to investigate the pathologies of cancers from a molecular angle. Under such a systematic framework, cancer types or subtypes can be identified through the corresponding gene expression profiles. Research on gene expression profile-based cancer type recognition has already attracted numerous efforts from a wide variety of research communities [6-7]. Investigations on leukemia [2], lymphoma [3], colon cancer [8], cutaneous melanoma [9], bladder cancer [10], breast cancer [11], lung cancer [12], and so on show very promising results. Supervised computational methods, such as multi-layer perceptron [13], naïve Bayes [14], support vector machines [14-15], semi-supervised Ellipsoid ARTMAP [16], and *k*-Top Scoring Paris [17], to name a few, have already been used in cancer diagnosis-oriented gene expression data analysis.

In this paper, we consider the situation in which we do not have labels for the cancer samples. This assumption is reasonable with the requirement for discovering unknown and novel cancer types or subtypes. In this case, unsupervised learning or cluster analysis [6] is required in order to explore the underlying data structure of the obtained data and provide cancer researchers with meaningful insights on the possible partition of the samples. Given N tumor samples measured over D genes, the corresponding microarray data matrix is represented as $\mathbf{X}=\{x_{ij}\}$, $1\leq i\leq N$, $1\leq j\leq D$, where x_{ij} represents the expression level of gene j in tissue sample i. The goal of our work is to generate a K-partition $\mathbf{C}=\{C_k\}$, $1\leq k\leq K$, such that $C_k\neq\phi$, $\bigcup_{k=1}^{K} C_k = \mathbf{C}$, and $C_k \cap C_l = \phi, k,l = 1,...,K$ and $k \neq l$.

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 samples, which is caused by factors such as sample collections and experiment cost. This problem is well known as the 'curse of dimensionality' in machine learning, which is introduced to indicate the exponential growth in computational complexity

Manuscript received November 19, 2007.

R. Xu is with the Applied Computational Intelligence Laboratory, Department of Electrical & Computer Engineering, University of Missouri - Rolla, Rolla, MO 65409 USA (phone: 573-341-6811; fax: 573-341-4521; e-mail: rxu@umr.edu).

S. Damelin is with the Department of Mathematical Sciences, Georgia Southern University, Statesboro, GA 30460 USA (e-mail: damelin@georgiasouthern.edu).

D. C. Wunsch is with the Department of Electrical & Computer Engineering, University of Missouri - Rolla, Rolla, MO 65409 USA (e-mail: dwunsch@umr.edu).

978-1-4244-1821-3/08/$25.00©2008 IEEE

and the demand for more samples as a result of high dimensionality in the feature space [18]. Not all of these genes (features) are relevant to the discrimination of tumors. From the computational point of view, the existence of numerous irrelevant features 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. Major explorations include principal component analysis [19] and ranking-based methods, such as signal-to-noise ratio [2], Fisher discriminant score [20], t-statistics score [21], and nonparametric test statistics like the TNoM score [22] and the Park score [23]. Most of these methods work in a supervised way, and the lack of such prior information makes the problem more difficult.

Here, we address the high-dimensional problem using diffusion maps, which consider 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 [24-26]. The new data obtained are then clustered with a neural network cluster theory, Fuzzy ART (FA) [27], to generate a partition of the cancer samples of interest. FA is based on Adaptive Resonance Theory (ART) [28-29], which was inspired by neural modeling research and was developed as a solution to the plasticity-stability dilemma: how adaptable (plastic) should a learning system be so that it does not suffer from catastrophic forgetting of previously-learned rules (stability)? ART can learn arbitrary input patterns in a stable, fast, and self-organizing way, thus overcoming the effect of learning instability that plagues many other competitive networks. Experimental results on a publicly accessible benchmark cancer data set, compared with other widely-used clustering algorithms, such as hierarchical clustering algorithms and K-means [6], demonstrate the effectiveness of our proposed method in addressing multidimensional gene expression data and ultimately identifying corresponding cancer types. A brief version of the work is published in [30].

The remainder of this paper is organized as follows. Section II and III present introductions to diffusion maps and FA, respectively. The experimental results are presented and discussed in section IV, and section V concludes the paper.

II. Diffusion Maps

Given a data set $\mathbf{X}=\{\mathbf{x}_i, i=1,\ldots,N\}$ on a m-dimensional data space, 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: $\mathbf{X} \times \mathbf{X} \rightarrow (0, \infty)$. Typically, a Gaussian kernel is defined as

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

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 $\Re^m$.

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 [24]. 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.

Because of the symmetry property of the kernel function, for each $t \geq 1$, we may obtain a sequence of N eigenvalues of $\mathbf{P}$ $1=\lambda_0 \geq \lambda_1 \geq \ldots \geq \lambda_N$, with the corresponding eigenvectors $\{\boldsymbol{\varphi}_j, j=1,\ldots,N\}$, satisfying,

$$\mathbf{P}^t \boldsymbol{\varphi}_j = \lambda^t \boldsymbol{\varphi}_j. \quad (4)$$

Using the eigenvectors as a new set of coordinates on the data set, the mapping from the original data space to an L-dimensional ($L< d$) Euclidean space $\Re^L$ can be defined as

$$\boldsymbol{\Psi}_t : \mathbf{x}_i \rightarrow \left(\lambda_1^t \boldsymbol{\varphi}_1(\mathbf{x}_i), \ldots, \lambda_L^t \boldsymbol{\varphi}_L(\mathbf{x}_i) \right)^T. \quad (5)$$

Correspondingly, the diffusion distance between a pair of points $\mathbf{x}_i$ and $\mathbf{x}_j$

$$D_t(\mathbf{x}_i, \mathbf{x}_j) = \left\| p^t(\mathbf{x}_i, \cdot) - p^t(\mathbf{x}_j, \cdot) \right\|_{1/\varphi_0}, \quad (6)$$

where φ_0 is the unique stationary distribution

$$\phi_0(\mathbf{x}) = \frac{d(\mathbf{x})}{\sum_{\mathbf{x}_i \in \mathbf{X}} d(\mathbf{x}_i)}, \mathbf{x} \in \Re^m, \quad (7)$$

is approximated with the Euclidean distance in $\Re^L$, written as

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

where $\|\cdot\|$ is the Euclidean norm in $\Re^L$. It can be seen that the more paths that connect two points in the graph, the smaller the diffusion distance is.

The kernel width parameter σ represents the rate at which the similarity between two points decays. There is no good theory to guide the choice of σ. Several heuristics have been proposed, and they boil down to trading off sparseness of the kernel matrix (small sigma) with adequate characterization of the true affinity of two points. One of the main reasons for using spectral clustering methods is that, with sparse kernel matrices, long range affinities are accommodated through the chaining of many local interactions as opposed to standard Euclidean distance methods - e.g. correlation - that impute global influence into each pair-wise affinity metric, making long range interactions dominate local interactions.

III. Fuzzy ART

Fuzzy ART (FA) incorporates fuzzy set theory into ART and extends the ART family by allowing stable recognition of clusters in response to both binary and real-valued input patterns with either fast or slow learning [27]. 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 $\mathbf{w}_j$, 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

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

where $\wedge$ is the fuzzy AND operator defined by

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

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,

$$T_J = \max_j \{T_j\} \ . \tag{11}$$

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,

$$\rho \leq \frac{|\mathbf{x} \wedge \mathbf{w}_J|}{|\mathbf{x}|} \ , \tag{12}$$

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

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

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 presentation of this input pattern, and a new competition is performed among the rest of the 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.

IV. 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) [13]. Gene expression levels of 6,567 genes were measured using cDNA microarrays for each sample, 2,308 of which passed the filter that requires the red intensity of a gene to be greater than 20 and were kept for further analyses. The relative red intensity (RRI) of a gene is defined as the ratio between the mean

TABLE I. Performance Results of Diffusion Maps And Fuzzy ART on the SRBCT Data Set.

	σ=22			σ=24			σ=26		
	RI (ρ)	*JC* (ρ)	*FM* (ρ)	*RI* (ρ)	*JC* (ρ)	*FM* (ρ)	*RI* (ρ)	*JC* (ρ)	*FM* (ρ)
L=5	0.7417 (0.5)	0.2337 (0.25)	0.3875 (0.25)	0.7661 (0.5)	0.3017 (0.25)	0.4646 (0.25)	0.7802 (0.5)	0.3082 (0.4)	0.4835 (0.4)
L=10	0.8569 (0.3)	0.5120 (0.3)	0.6929 (0.3)	0.8260 (0.35)	0.4136 (0.2)	0.6064 (0.35)	0.8187 (0.45)	0.3652 (0.45)	0.5778 (0.45)
L=15	0.8795 (0.35)	0.5648 (0.35)	0.7436 (0.35)	0.8290 (0.35)	0.4853 (0.2)	0.6539 (0.2)	0.8560 (0.35)	0.5015 (0.35)	0.6879 (0.35)
L=20	0.8707 (0.25)	0.6245 (0.2)	0.7704 (0.2)	0.8346 (0.4)	0.5047 (0.25)	0.6715 (0.25)	0.8795 (0.35)	0.5652 (0.35)	0.7437 (0.35)
L=50	0.8437 (0.3)	0.5503 (0.3)	0.7099 (0.3)	0.8149 (0.35)	0.4981 (0.25)	0.6672 (0.25)	0.8175 (0.5)	0.5040 (0.3)	0.6740 (0.3)

	σ=28			σ=30			σ=32		
	RI (ρ)	*JC* (ρ)	*FM* (ρ)	*RI* (ρ)	*JC* (ρ)	*FM* (ρ)	*RI* (ρ)	*JC* (ρ)	*FM* (ρ)
L=5	0.7761 (0.5)	0.2950 (0.2)	0.4599 (0.45)	0.7743 (0.45)	0.2898 (0.4)	0.4668 (0.45)	0.7708 (0.6)	0.3044 (0.25)	0.4670 (0.25)
L=10	0.9019 (0.2)	0.6673 (0.2)	0.8039 (0.2)	0.8601 (0.3)	0.5036 (0.3)	0.6953 (0.3)	0.8760 (0.2)	0.5977 (0.2)	0.7513 (0.2)
L=15	0.8431 (0.2)	0.5389 (0.2)	0.7006 (0.2)	0.8619 (0.3)	0.5553 (0.3)	0.7186 (0.3)	0.8322 (0.25)	0.4742 (0.25)	0.6499 (0.25)
L=20	0.8284 (0.25)	0.5485 (0.2)	0.7129 (0.2)	0.8578 (0.4)	0.5268 (0.25)	0.6907 (0.25)	0.8160 (0.45)	0.4316 (0.25)	0.6031 (0.25)
L=50	0.8137 (0.55)	0.4052 (0.25)	0.5882 (0.25)	0.8354 (0.35)	0.5495 (0.35)	0.7102 (0.35)	0.8196 (0.6)	0.4985 (0.35)	0.6665 (0.35)

RI: Rand index;
JC: Jaccard coefficient;
FM: Fowlkes and Mallows index

TABLE II. PERFORMANCE RESULTS OF DIFFUSION MAPS AND HIERARCHICAL CLUSTERING ON THE SRBCT DATA SET.

	σ=22			σ=24			σ=26		
	RI	*JC*	*FM*	*RI*	*JC*	*FM*	*RI*	*JC*	*FM*
L=5	0.6706	0.2544	0.4255	0.7114	0.2682	0.4327	0.7320	0.2525	0.4111
L=10	0.6920	0.2547	0.4738	0.7267	0.3416	0.5729	0.7473	0.2493	0.4540
L=15	0.6324	0.2642	0.4976	0.6567	0.2617	0.4972	0.6550	0.2549	0.4740
L=20	0.4126	0.2712	0.5044	0.4572	0.2617	0.4972	0.4790	0.2578	0.4913
L=50	0.5075	0.2703	0.5102	0.4364	0.2578	0.4913	0.5145	0.2578	0.4913

	σ=28			σ=30			σ=32		
	RI	*JC*	*FM*	*RI*	*JC*	*FM*	*RI*	*JC*	*FM*
L=5	0.7150	0.2504	0.4011	0.6820	0.2236	0.3744	0.7138	0.2077	0.3525
L=10	0.7217	0.2941	0.4550	0.7011	0.2380	0.4222	0.7085	0.3091	0.4991
L=15	0.7264	0.2514	0.4145	0.7044	0.2525	0.4586	0.7094	0.2612	0.4937
L=20	0.7220	0.2338	0.4161	0.6441	0.2594	0.4937	0.7032	0.2638	0.4937
L=50	0.5786	0.2912	0.4937	0.6682	0.2594	0.4937	0.7073	0.3186	0.5334

RI: Rand index;
JC: Jaccard coefficient;
FM: Fowlkes and Mallows index

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. The data are expressed as a matrix, $\mathbf{X}=\{x_{ij}\}_{83\times 2,308}$. In our further analysis, an additional logarithm was taken to linearize the relations between different genes and to make very high expression levels less high.

According to our experiments, we find that the clustering results are not sensitive to the category choice parameter α, which is then set as 0.1 for our further study. We adjusted the kernel width parameter σ and vigilance parameter ρ, and observed the performance of the proposed method. Because we already have a pre-specified partition **H** of the data set, which is also independent from the clustering structure **C** resulting from the use of FA, the performance can be evaluated by comparing **C** to **H** in terms of external criteria, such as the Rand index, the Jaccard coefficient, and the Fowlkes and Mallows index [31].

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 **C** and **H**.

- Case 1: $\mathbf{x}_i$ and $\mathbf{x}_j$ belong to the same clusters of **C** and the same category of **H**.
- Case 2: $\mathbf{x}_i$ and $\mathbf{x}_j$ belong to the same clusters of **C** but different categories of **H**.
- Case 3: $\mathbf{x}_i$ and $\mathbf{x}_j$ belong to different clusters of **C** but the same category of **H**.
- Case 4: $\mathbf{x}_i$ and $\mathbf{x}_j$ belong to different clusters of **C** and a different category of **H**.

Correspondingly, the number of pairs of samples for the four cases are denoted as *a*, *b*, *c*, and *d*, respectively. Because the total number of pairs of samples is *N*(*N*-1)/2, denoted as *M*, we have $a+b+c+d=M$. The external criteria that we used in our analysis can then be defined as follows:

1. Rand index

$$R=(a+d)/M\,; \tag{14}$$

2. Jaccard coefficient

$$J=a/(a+b+c)\,; \tag{15}$$

3. Fowlkes and Mallows index

$$FM=\sqrt{\frac{a}{a+b}\frac{a}{a+c}}\,; \tag{16}$$

As can be seen from the definition, the larger the values of these indices, the more similar are **C** and **H**. Specifically, the values of both the Rand index and the Jaccard coefficient are in the range of [0, 1]. The major difference between these two statistics is that the Rand index emphasizes the situation in which pairs of samples belong to the same group or different groups in both **C** and **H**, but the Jaccard coefficient excludes *d* in the similarity measure.

Table I summarizes the best clustering results for the SRBCT data set using the three external criteria, with σ varying from 22 to 32. The corresponding ρ is also indicated in the table. The dimensions of the transformed space are chosen at 5, 10, 15, 20, and 50, respectively. From the table, it can be seen that the effective dimensions for representing the data are 10 and 15, among those selected. The values of the indices decrease as the dimension becomes either smaller or larger.

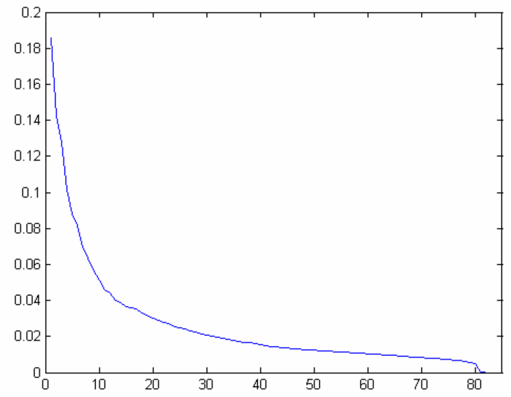

Fig. 1. The eigenvalues of the affinity matrix for the SRBCT data set. σ, ρ, and L are chosen at 30, 0.3, and 15, respectively. For clarification, the first eigenvalue that is equal to 1 is not shown here.

TABLE III. PERFORMANCE RESULTS OF DIFFUSION MAPS AND *K*-MEANS ON THE SRBCT DATA SET.

	σ=22			σ=24			σ=26		
	RI	*JC*	*FM*	*RI*	*JC*	*FM*	*RI*	*JC*	*FM*
L=5	0.7285	0.3283	0.5492	0.7493	0.2694	0.4285	0.7549	0.2019	0.3774
L=10	0.7567	0.2292	0.3986	0.7667	0.3392	0.5066	0.7555	0.2407	0.4067
L=15	0.7388	0.2276	0.3718	0.7670	0.2443	0.4340	0.7423	0.2675	0.4289
L=20	0.7344	0.2676	0.4294	0.7658	0.2951	0.4727	0.7561	0.3159	0.4885
L=50	0.7200	0.2498	0.4203	0.7579	0.2749	0.4316	0.7364	0.2487	0.4007

	σ=28			σ=30			σ=32		
	RI	*JC*	*FM*	*RI*	*JC*	*FM*	*RI*	*JC*	*FM*
L=5	0.7379	0.2706	0.4271	0.7576	0.3039	0.4726	0.7294	0.2222	0.3717
L=10	0.7482	0.2391	0.3969	0.7540	0.2882	0.4475	0.7699	0.2405	0.4294
L=15	0.7643	0.2319	0.4018	0.7402	0.2840	0.4579	0.7482	0.2279	0.3949
L=20	0.7505	0.2299	0.3826	0.7514	0.3310	0.4989	0.7443	0.2415	0.4052
L=50	0.7367	0.2265	0.3829	0.7332	0.2695	0.4300	0.7435	0.2307	0.3868

RI: Rand index;
JC: Jaccard coefficient;
FM: Fowlkes and Mallows index

We further examine the eigenvalues for the corresponding affinity matrix (see Fig. 1), 15 of which are listed below, in decreasing order:

1.0000 0.1856 0.1424 0.1277 0.1017 0.0879 0.0816
0.0698 0.0633 0.0569 0.0514 0.0459 0.0442
0.0401 0.0391 …

Obviously, the curve decays rapidly for the first 15 eigenvalues and then decreases gradually. This explains the deterioration of the clustering performance when we use only 5 corresponding eigenvectors to construct the mapping, which causes the loss of too much information.

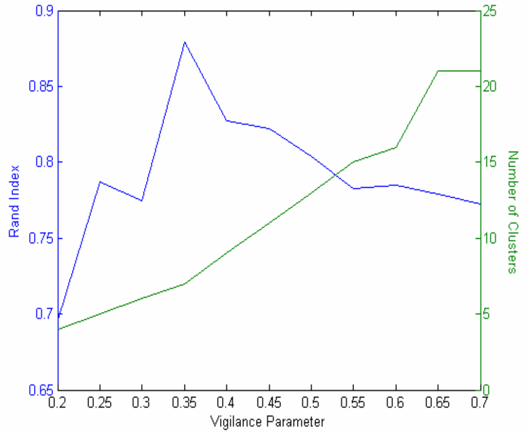

Fig. 2. The vigilance parameter vs. the number of clusters and the value of the Rand index for the SRBCT data set.

Fig. 2 shows the influence of ρ on the number of generated clusters and the value of the Rand index. σ and L are set as 22 and 15, respectively. As ρ increases, the number of clusters increases, too. However, the largest value of the Rand index is achieved at ρ=0.35. When performing the stricter vigilance tests, the samples that belong to the same category are divided into more small categories, which causes the decrease of the value of the Rand index.

TABLE IV. PERFORMANCE RESULT OF FA, HIERARCHICAL CLUSTERING, AND *K*-MEANS ON THE SRBCT DATA SET WITHOUT USING DIFFUSION MAPS.

	RI	*JC*	*FM*
FA	0.7705	0.3089	0.5183
Hierarchical Clustering	0.4505	0.2578	0.4919
K-means	0.7138	0.1970	0.3330

RI: Rand index;
JC: Jaccard coefficient;
FM: Fowlkes and Mallows index

The best clustering results with hierarchical clustering algorithms (single-linkage) and *K*-means on the SRBCT data set are summarized in Tables II and III, respectively. It can be seen that FA can consistently achieve better partitions of the given samples than the other two methods. We further examined the clustering results of FAs, hierarchical clustering algorithms, and *K*-means algorithms when the diffusion maps are not used. As shown in Table IV, the effectiveness of diffusion maps is obvious: the performance of all three algorithms without dimension reduction deteriorates dramatically, especially for the hierarchical clustering algorithm.

V. CONCLUSIONS

Cancer classification is important for subsequent diagnosis and treatment. DNA microarray technologies provide a promising way to address the problem, while bringing many challenges. Particularly, publicly accessible gene expression data sets usually include a small set of samples for each tumor type, in contrast to the rapidly and persistently increasing capability of gene chip technologies. Here, we propose to use the diffusion maps to reduce the high dimensions of gene expression data and Fuzzy ART to form the clusters of cancer samples. The experimental results demonstrate the potential of the proposed method in extracting useful information from these high-dimensional data sets.

ACKNOWLEDGMENT

We thank Boaz Nadler for valuable comments. Partial support for this research from the National Science Foundation, and from the M.K. Finley Missouri endowment, is gratefully acknowledged.

REFERENCES

[1] A. Miniño, M. Heron, and B. Smith, "Deaths: Preliminary data for 2004," *National Vital Statistics Reports*, vol. 54, no. 19, 2006.

[2] 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.

[3] A. Alizadeh, M. Eisen, R. Davis, C. Ma, I. Lossos, A. Rosenwald, J. Boldrick, H. Sabet, T. Tran, X. Yu, J. Powell, L. Yang, G. Marti, T. Moore, J. Hudson, Jr, L. Lu, D. Lewis, R. Tibshirani, G. Sherlock, W. Chan, T. Greiner, D. Weisenburger, J. Armitage, R. Warnke, R. Levy, W. Wilson, M. Grever, J. Byrd, D. Bostein, P. Brown, and L. Staudt, "Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling," *Nature*, vol. 403, pp. 503-511, 2000.

[4] M. Eisen, and P. Brown, "DNA arrays for analysis of gene expression," *Methods Enzymol*, vol. 303, pp. 179-205, 1999.

[5] R. Lipshutz, S. Fodor, T. Gingeras, and D. Lockhart, "High density synthetic oligonucleotide arrays," *Nature Genetics*, vol. 21, pp. 20-24, 1999.

[6] R. Xu and D. Wunsch II, "Survey of Clustering Algorithms," *IEEE Transactions on Neural Networks*, vol.16, no.3, pp.645-678, 2005.

[7] G. McLachlan, K. Do, and C. Ambroise, "Analyzing microarray gene expression data," John Wiley & Sons, Inc., Hoboken, NJ, 2004.

[8] U. Alon, N. Barkai, D. Notterman, K. Gish, S. Ybarra, D. Mack, and A. Levine, "Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays," *Proc. Natl. Acad. Sci. USA* 96, pp. 6745-6750, 1999.

[9] M. Bittner, P. Meltzer, Y. Chen, Y. Jiang, E. Seftor, M. Hendrix, M. Radmacher, R. Simon, Z. Yakhini, 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, pp. 536-540, 2000.

[10] L. Dyrskjøt, T. Thykjaer, M. Kruhøffer, J. Jensen, N. Marcussen, S. Hamilton-Dutoit, H. Wolf, and T. Ørntoft, "Identifying distinct classes of bladder carcinoma using microarrays," *Nature Genetics*, vol. 33, pp. 90-96, 2003.

[11] C. Perou, T. Sørlie, M. Eisen, M. Rijn, S. Jeffrey, C. Rees, J. Pollack, D. Ross, J. Johnsent, L. Akslen, Ø. Fluge, A. Pergamenschlkov, C. Williams, S. Zhu, P. Lønning, A. Børresen-Dale, P. Brown, and D. Botstein, "Molecular portraits of human breast tumors," *Nature*, vol. 406, pp. 747-752, 2000.

[12] M. Garber, O. Troyanskaya, K. Schluens, S. Petersen, Z. Thaesler, M. Pacyna-Gengelbach, M. Rijn, G. Rosen, C. Perou, R. Whyte, R. Altman, P. Brown, D. Botstein, and I. Petersen, "Diversity of gene expression in adenocarcinoma of the lung," *Proc. Natl. Acad. Sci. USA* 98, pp. 13784-13789, 2001.

[13] 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.

[14] T. Li, C. Zhang, and M. Ogihara, "A comparative study of feature selection and multiclass classification methods for tissue classification based on gene expression," *Bioinformatics*, vol. 20, no. 15, pp. 2429-2437, 2004.

[15] A. Statnikov, C. Aliferis, I. Tsamardinos, D. Hardin, and S. Levy, "A comprehensive evaluation of multicategory classification methods for microarray gene expression cancer diagnosis," *Bioinformatics*, vol. 21, no. 5, pp. 631-643, 2005.

[16] R. Xu, G. Anagnostopoulos, and D. Wunsch II, "Multi-class cancer classification using semi-supervised Ellipsoid ARTMAP and particle swarm optimization with gene expression data," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 4, no. 1, pp. 65-77, 2007.

[17] A. Tan, D. Naiman, L. Xu, R. Winslow, and D. Geman, "Simple decision rules for classifying human cancers from gene expression profiles," *Bioinformatics*, vol. 21, no. 20, pp. 3896-3904, 2005.

[18] R. Bellman, "Adaptive control processes: A guided tour," Princeton University Press, Princeton, NJ, 1961.

[19] K. Yeung, and W. Ruzzo, "Principal component analysis for clustering gene expression data," *Bioinformatics*, vol. 17, pp. 763-774, 2001.

[20] J. Jaeger, R. Sengupta, and W. Ruzzo, "Improved gene selection for classification of microarrays," *Pacific Symposium on Biocomputing* 8**,** pp. 53-64, 2003.

[21] D. Nguyen, and D. Rocke, "Multi-class cancer classification via partial least squares with gene expression profiles," *Bioinformatics*, vol. 18, pp. 1216-1226, 2002.

[22] 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.

[23] P. Park, M. Pagano, and M. Boneti, "A nonparametric scoring algorithm for identifying informative genes from microarray data," *Pacific Symposium on Biocomputing*, 6, pp. 52-63, 2001.

[24] R. Coifman and S. Lafon, "Diffusion maps," *Applied and Computational Harmonic Analysis*, vol. 21, pp. 5-30, 2006.

[25] S. Lafon and A. Lee, "Diffusion maps and coarse-graining: A unified framework for dimensionality reduction, graph partitioning, and data set parameterization," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 28, no. 9, pp. 1393-1403, 2006.

[26] 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.

[27] 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.

[28] G. Carpenter, and S. Grossberg, "A massively parallel architecture for a self-organizing neural pattern recognition machine," *Computer Vision, Graphics, and Image Processing*, vol. 37, pp. 54-115, 1987.

[29] S. Grossberg, "Adaptive pattern recognition and universal encoding II: feedback, expectation, olfaction, and illusions," *Biological Cybernetics*, vol. 23, pp. 187-202, 1976.

[30] 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.

[31] A. Jain and R. Dubes, "Algorithms for clustering data," Prentice Hall, Englewood Cliffs, NJ, 1988.