8-1-2007

Applications of Diffusion Maps in Gene Expression Data-Based Cancer Diagnosis Analysis

Rui Xu
Missouri University of Science and Technology

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

Steven Damelin

Recommended Citation

R. Xu et al., "Applications of Diffusion Maps in Gene Expression Data-Based Cancer Diagnosis Analysis," *Proceedings of the 29th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, 2007*, Institute of Electrical and Electronics Engineers (IEEE), Aug 2007.
The definitive version is available at https://doi.org/10.1109/IEMBS.2007.4353367

Proceedings of the 29th Annual International
Conference of the IEEE EMBS
Cité Internationale, Lyon, France
August 23-26, 2007.

Applications of Diffusion Maps in Gene Expression Data-Based Cancer Diagnosis Analysis

Rui Xu, *Member, IEEE*, Steven Damelin., and Donald C. Wunsch II, *Fellow, IEEE*

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. One of the major problems in cancer type recognition-oriented gene expression data analysis is the overwhelming number of measures of gene expression levels versus the small number of samples, which causes the curse of dimension issue. Here, 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 derived data are then clustered with Fuzzy ART to form the division of the cancer samples. Experimental results on the small round blue-cell tumor data set demonstrate the effectiveness of our proposed method in addressing multidimensional gene expression data and identifying different types of tumors.**

Keywords—**Gene expression profiles, Diffusion maps, Dimensionality reduction, Fuzzy ART**.

I. Introduction

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, which are largely dependent on the morphological appearance of tumors and parameters derived from clinical observations [1-5]. Tumors with similar appearance may have quite different origins and may therefore respond differently to the same treatment therapy. In contrast, DNA microarray technologies [6-7], which can measure the expression levels of tens of thousands of genes simultaneously, offer cancer researchers a new 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.

Cancer type recognition-oriented gene expression data analysis raises many computational challenges [8-9]. Particularly, one of the major challenges is the 'curse of dimensionality' due to the overwhelming number of measures of gene expression levels compared to the small number of tissue samples. This issue is well known in the community of computational intelligence and machine learning, where 'curse of dimensionality' indicates the exponential growth in computational complexity and the demand for more samples as a result of high dimensionality in the feature space [10]. In the context of gene expression data-based cancer diagnosis, the existence of numerous genes or features that are irrelevant to the discrimination of specific tumors not only increases the computational complexity, but impairs the effective discovery of the real cancer clusters. In this sense, feature selection or extraction is critically important for dimensionality reduction and further cancer type discrimination [1, 11-19].

In this work, we address the high-dimensionality problem using diffusion maps that consider the eigenfunctions of Markov matrices as a system of coordinates on the original data set in order to obtain an efficient representation of data geometric descriptions [20-22]. The derived new data are then clustered with Fuzzy ART (FA) [23] to generate the division of the cancer samples. FA is based on Adaptive Resonance Theory (ART) [24-25], 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). FA has the desirable characteristics of fast and stable learning, transparent learning paradigm, and atypical pattern detection. Experimental results on the small round blue-cell tumor data set [26] demonstrate the effectiveness of our proposed method in addressing multidimensional gene expression data and ultimately identifying corresponding cancer types or subtypes.

The remainder of this paper is organized as follows. Section II introduces the proposed methods, including diffusion maps and FA. The experimental results are presented and discussed in section III, and section IV concludes the paper.

II. Methods

A. Diffusion Maps

Given a data set $\mathbf{X}=\{\mathbf{x}_i, i=1,\ldots,N\}$ on a d-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

R. Xu is with the Applied Computational Intelligence Laboratory, Department of Electrical And Computer Engineering, University of Missouri – Rolla, Rolla, MO 65409-0249 USA (e-mail: rxu@ umr.edu).

S. Damelin is with Department of Mathematical Sciences, Georgia Southern University, Statesboro, GA, 30460-8093 USA (e-mail: damelin@georgiasouthern.edu).

D. C. Wunsch is with the Applied Computational Intelligence Laboratory, Department of Electrical And Computer Engineering, University of Missouri – Rolla, Rolla, MO 65409-0249 USA (e-mail: dwunsch@umr.edu).

non-negative, symmetric, and positive definite kernel w: $\mathbf{X} \times \mathbf{X} \rightarrow \Re$. Typically, a Gaussian kernel, 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, satisfies such constraints and reflects the degree of similarity between $\mathbf{x}_i$ and $\mathbf{x}_j$, and $\|\cdot\|$ is the Euclidean norm in $\Re^d$.

Letting

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

be the degree of $\mathbf{x}_i$, the Markov or affinity matrix $\mathbf{P}=\{p(\mathbf{x}_i, \mathbf{x}_j)\}$ 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 [20]. Therefore, the parameter t defines the granularity of the analysis. When the value of t increases, local geometric data information 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 can 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

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

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 connecting 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; they boil down to trading off sparseness of the kernel matrix (small sigma) with adequate characterization of true affinity of two points. The reason that spectral clustering methods work 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 that impute global influence into each pair-wise affinity metric, making long range interactions wash out local interactions.

B. Fuzzy ART

FA incorporates fuzzy set theory into ART and extends the ART family by being capable of learning stable recognition clusters in response to both binary and real-valued input patterns with either fast or slow learning [23]. The basic FA architecture consists of two-layer nodes or neurons, the feature representation field F_1, and the category representation field F_2, as shown in Fig. 1.

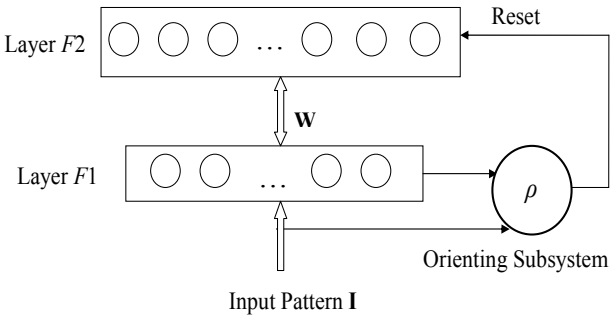

Fig. 1. Topological structure of Fuzzy ART. Layers $F1$ and $F2$ are connected via adaptive weights W. The orienting subsystem is controlled by the vigilance parameter ρ.

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

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

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

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

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,

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

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

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

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. If an uncommitted neuron is selected for coding, then a new uncommitted neuron is created to represent a potential new cluster.

III. Results

We applied the proposed method to the small round blue-cell tumor (SRBCT) data set, which is published from the diagnostic research of small round blue-cell tumors in children. The SRBCT data set consists of 83 samples belonging to four categories: Burkitt lymphomas (BL), the Ewing family of tumors (EWS), neuroblastoma (NB) and rhabdomyosarcoma (RMS) [26]. Gene expression levels of 6,567 genes were measured using cDNA microarray 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 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 $E=\{e_{ij}\}_{83\times 2,308}$, where e_{ij} represents the expression level of gene j in tissue sample i.

For the following analysis, we set the category choice parameter α to 0.1 and observed the effect of the kernel width parameter σ and vigilance parameter ρ on the performance of the proposed method. Because we already have a pre-specified partition $\boldsymbol{P}$ of the data set, which is also independent from the clustering structure $\boldsymbol{C}$ resulting from the use of FA, the performance can be evaluated by comparing $\boldsymbol{C}$ to $\boldsymbol{P}$ in terms of external criteria [27], such as Rand index.

Considering a pair of tissue samples $\mathbf{e}_i$ and $\mathbf{e}_j$, there are four different cases based on how $\mathbf{e}_i$ and $\mathbf{e}_j$ are placed in $\boldsymbol{C}$ and $\boldsymbol{P}$.

- Case 1: $\mathbf{e}_i$ and $\mathbf{e}_j$ belong to the same clusters of $\boldsymbol{C}$ and the same category of $\boldsymbol{P}$.
- Case 2: $\mathbf{e}_i$ and $\mathbf{e}_j$ belong to the same clusters of $\boldsymbol{C}$ but different categories of $\boldsymbol{P}$.
- Case 3: $\mathbf{e}_i$ and $\mathbf{e}_j$ belong to different clusters of $\boldsymbol{C}$ but the same category of $\boldsymbol{P}$.
- Case 4: $\mathbf{e}_i$ and $\mathbf{e}_j$ belong to different clusters of $\boldsymbol{C}$ and different category of $\boldsymbol{P}$.

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 $N(N\text{-}1)/2$, denoted as M, we have $a+b+c+d=M$. The Rand index is defined as follows:

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

Table I summarizes the best clustering results based on Rand index, with σ varying from 24 to 32. The corresponding ρ is also indicated in the table. The dimension of the transformed space is 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 or 15, among the selected possibilities. The number of incorrect assignments increases as the dimension becomes either too small or too large. For the former case, too much useful information on class discrimination is discarded and for the latter case, some irrelevant components are still kept, which disturbs the effective unveiling of the underlying data structure.

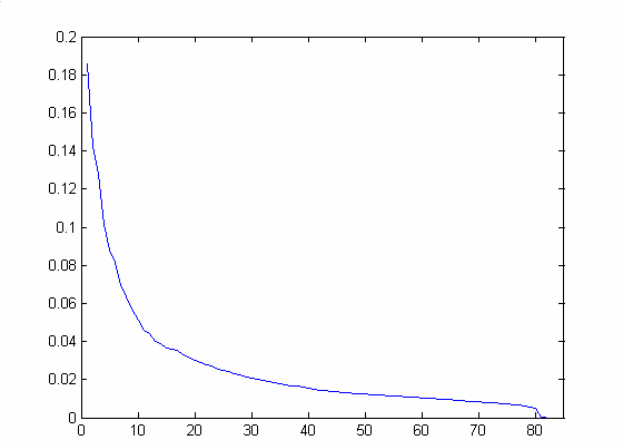

Fig. 2. The eigenvalues of the affinity matrix for the SRBCT data set. σ, ρ, and L are chosen at 30, 0.3, and 15, respectively. For clarification, the first eigenvalue that is equal to 1 is not shown here.

We further examine the eigenvalues for the corresponding affinity matrix (see Fig. 2), 15 of which are listed below, in a 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.

IV. Conclusion

Cancer classification is important for subsequent diagnosis and treatment. DNA microarray technologies provide a promising way to improve prediction accuracy of cancer types, but they inevitably bring 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

TABLE I. PERFORMANCE RESULTS OF DIFFUSION MAPS AND FUZZY ART ON THE SRBCT DATA SET.

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

RI: Rand index;

of gene chip technologies that provide cancer researchers with rich gene expression level measurement. Here, we propose to use the diffusion maps to reduce the high dimensionality of gene expression data, and furthermore, to adopt Fuzzy ART to form the clusters of cancer samples. The experimental results demonstrate the potential of the proposed method to extract useful information from high-dimensional data and provide meaningful insights in discriminating different types of cancer samples. Future research includes the construction of hierarchical clustering by considering the transition probability for more than one time step and the investigation of the performance of the method to more benchmark data sets.

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] 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.

[3] 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.

[4] 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.

[5] 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*, 6:R22, 2005.

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

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

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

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

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

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

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

[13] 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.

[14] 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.

[15] V. Tseng and C. Kao, "Efficiently mining gene expression data via a novel parameterless clustering method," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 2, pp. 355 – 365, 2005

[16] L. Wang, F. Chu, and W. Xie, "Accurate cancer classification using expressions of very few genes," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 4, no. 1, pp. 40-53, 2007.

[17] S. Mitra and Y. Hayashi, "Bioinformatics with soft computing," *IEEE Transactions on Systems, Man and Cybernetics, Part C*, vol. 36, pp. 616-635, 2006.

[18] B. Liu, C. Wan, and L. Wang, "An efficient semi-unsupervised gene selection method via spectral biclustering," *IEEE Transactions on Nano-Bioscience*, vol. 5, no. 2, pp. 110-114, 2006.

[19] S. Ray, S. Bandyopadhyay, P. Mitra, and S. Pal, "Bioinformatics in neurocomputing framework," *IEE Proceedings Circuits, Devices and Systems*, vol. 152, pp. 556-564, 2005.

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

[21] 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.

[22] 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.

[23] 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.

[24] 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.

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

[26] 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.

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