arXiv:1411.5737v5 [cs.NE] 6 Oct 2015

Fuzzy Adaptive Resonance Theory, Diffusion Maps and their applications to Clustering and Biclustering

S. B. Damelin[1], Y. Gu[2], D. C. Wunsch II[3], R. Xu[4] *

[1] Mathematical Reviews
The American Mathematical Society
Ann Arbor, MI 48103 USA

[2] Department of Mathematics
University of Michigan – Ann Arbor
Ann Arbor, MI 48109 USA

[3] Applied Computational Intelligence Laboratory
Department of Electrical and Computer Engineering
University of Missouri – Rolla
Rolla, MO 65409-0249 USA

[4] GE Global Research
Niskayuna, NY 12309 USA

Dedicated to our friend and colleague Prof. Alexander Gorban.

Abstract. In this paper, we describe an algorithm FARDiff (Fuzzy Adaptive Resonance Diffusion) which combines Diffusion Maps and Fuzzy Adaptive Resonance Theory to do clustering on high dimensional data. We describe some applications of this method and some problems for future research.

1. Introduction

In this paper, we describe an algorithm FARDiff (Fuzzy Adaptive Resonance Diffusion) which combines Diffusion Maps and Fuzzy Adaptive Resonance Theory to do clustering on high dimensional data. This algorithm was introduced in the papers [9, 10, 18, 19, 21]. We describe some applications of this method and some problems for future research.

The dimensionality reduction part of FARDiff is achieved via a nonlinear diffusion map, which interprets the eigenfunctions of Markov matrices as a system of coordinates on the dataset in order to obtain

---
*Corresponding author. E-mail: `damelin@umich.edu`.

an efficient representation of certain geometric descriptions of the data. Our algorithm is sensitive to the connectivity of the data points. Clustering was achieved using Fuzzy Adaptive Resonance Theory.

The structure of this paper is as follows. Section 2 will describe the dimension reduction, section 3 will describe the clustering method, section 4 will discuss some applications, and section 5 will discuss work in progress.

2. Diffusion Maps

In this section, we describe the nonlinear dimension reduction spectral method part of FARDiff. We use nonlinear diffusion maps. See [6, 15] and references sited there in.

Let m be a large enough fixed non-negative integer, and let $\mathbf{X} = \{\mathbf{x}_i, i = 1, \ldots, N\} \subset \mathbb{R}^m$ be a data set consisting of $N \geq 1$ distinct points. 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 Gaussian kernel, $w : \mathbb{R}^m \times \mathbb{R}^m \to \mathbb{R}$, defined by

$$w(\mathbf{z}, \mathbf{z}') = \exp\left(-\frac{\|\mathbf{z} - \mathbf{z}'\|^2}{\sigma^2}\right), \qquad \mathbf{z}, \mathbf{z}' \in \mathbb{R}^m,$$

where $\|\cdot\|$ is the Euclidean norm in $\mathbb{R}^m$ and σ is a fixed nonzero real number to be chosen in a moment. w here satisfies symmetric property $w(\mathbf{z}, \mathbf{z}') = w(\mathbf{z}', \mathbf{z})$ and positivity preserving property $w(\mathbf{z}, \mathbf{z}') \geq 0$. The first property is useful for performing spectral analysis, and the second property allows it to be interpreted as a scaled probability.

We choose σ to be a parameter depending on the data set $\mathbf{X}$, and we shall write for each $1 \leq i, j \leq N$,

$$w(\mathbf{x}_i, \mathbf{x}_j) = \exp\left(-\frac{\|\mathbf{x}_i - \mathbf{x}_j\|^2}{\sigma^2}\right), \qquad \mathbf{x}_i, \mathbf{x}_j \in \mathbf{X}. \tag{2.1}$$

w reflects the degree of similarity between $\mathbf{x}_i$ and $\mathbf{x}_j$. The resulting symmetric semi-positive definite matrix $\mathbf{W} = \{w(\mathbf{x}_i, \mathbf{x}_j)\}_{N \times N}$ is called the affinity matrix.

Let $d : \mathbf{X} \to \mathbb{R}$ be defined by

$$d(\mathbf{x}) = \sum_{\mathbf{x}' \in \mathbf{X}} w(\mathbf{x}, \mathbf{x}'), \qquad \mathbf{x} \in \mathbf{X}, \tag{2.2}$$

be the degree of $\mathbf{x}$. For each $1 \leq i, j \leq N$, a Markov or affinity matrix $\mathbf{P}_{N \times N}$ is then constructed by calculating each entry of $\mathbf{P}$ as

$$p(\mathbf{x}_i, \mathbf{x}_j) = \frac{w(\mathbf{x}_i, \mathbf{x}_j)}{d(\mathbf{x}_i)}. \tag{2.3}$$

From the definition of w, $p(\mathbf{x}_i, \mathbf{x}_j)$ can be interpreted as the transition probability from $\mathbf{x}_i$ to $\mathbf{x}_j$ in one time step. From the definition of the Gaussian kernel it can be seen that the transition probability will be high for similar elements. This idea can be further extended by considering $p^t(\mathbf{x}_i, \mathbf{x}_j)$ in the tth power $\mathbf{P}^t$ of $\mathbf{P}$ as the probability of transition from $\mathbf{x}_i$ to $\mathbf{x}_j$ in t time steps [6], where t is a fixed non-negative integer. Here, $p^t(\mathbf{x}_i, \mathbf{x}_j)$ sums all paths of length t from point $\mathbf{x}_i$ to point $\mathbf{x}_j$. Hence, 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 size of t makes it possible to control the generation of more specific or broader clusters.

The diffusion distance is related to $\mathbf{P}$ and is given by

$$D_t(\mathbf{x}_i, \mathbf{x}_j)^2 = \sum_{\mathbf{x} \in \mathbf{X}} \left| p^t(\mathbf{x}_i, \mathbf{x}) - p^t(\mathbf{x}_j, \mathbf{x}) \right|^2. \tag{2.4}$$

It can be seen that the more paths that connect two points in the graph, the higher probability of paths of length t between two points is, and the smaller the diffusion distance is. Therefore, the diffusion distance can be used as a measurement of a specific geometric structure.

Next, we aim to find a mapping that reassembles the data in a lower dimensional space but preserves the diffusion distances between data points optimally. Consider a mapping

$$\boldsymbol{\Psi}'_t(\mathbf{x}_i) = \left(p^t(\mathbf{x}_i, \mathbf{x}_1), \ldots, p^t(\mathbf{x}_i, \mathbf{x}_N) \right)^T, \qquad 1 \le i \le N. \tag{2.5}$$

The Euclidean distance between two points $\mathbf{x}_i$ and $\mathbf{x}_j$ is

$$\left\| \boldsymbol{\Psi}'_t(\mathbf{x}_i) - \boldsymbol{\Psi}'_t(\mathbf{x}_j) \right\|^2 = \sum_{\mathbf{x} \in \mathbf{X}} \left| p^t(\mathbf{x}_i, \mathbf{x}) - p^t(\mathbf{x}_j, \mathbf{x}) \right|^2,$$

which is also the diffusion distance. The mapping thus preserves the original data structure in the sense of diffusion distance.

Next, we need to perform dimensionality reduction. Because of the symmetry property of w, for each t, we may obtain a sequence of N eigenvalues of $\mathbf{P}$, $1 = \lambda_1 > \lambda_2 > \cdots > \lambda_N \ge 0$, with corresponding eigenvectors $\{\boldsymbol{\Phi}_j, j = 1, \ldots, N\}$, satisfying,

$$\mathbf{P}^t \boldsymbol{\Phi}_j = \lambda_j^t \boldsymbol{\Phi}_j. \tag{2.6}$$

Suppose we wish to reduce our data from $\mathbb{R}^m$ to $\mathbb{R}^L$ where $1 \le L < m$[1]. To achieve it, we choose the first L eigenvalues and their corresponding eigenvectors of $\mathbf{P}$, and rewrite the mapping (2.5) as

$$\boldsymbol{\Psi}_t(\mathbf{x}_i) = \left(\lambda_1^t \boldsymbol{\Phi}_1(\mathbf{x}_i), \ldots, \lambda_L^t \boldsymbol{\Phi}_L(\mathbf{x}_i) \right)^T, \tag{2.7}$$

where $\boldsymbol{\Phi}_j(\mathbf{x}_i)$ is the ith element of the eigenvector $\boldsymbol{\Phi}_j$. We observe that the new data point $\mathbf{x}_i$'s dimension has been reduced to L. The new mapping reduces the dimensionality of the original data set by retaining the important geometric structure associated with dominant eigenvalues and eigenvectors. Correspondingly, the diffusion distance between a pair of points $\mathbf{x}_i$ and $\mathbf{x}_j$ is approximated with the Euclidean distance in $\mathbb{R}^L$, written as

$$D_t(\mathbf{x}_i, \mathbf{x}_j)^2 = \left\| \boldsymbol{\Psi}_t(\mathbf{x}_i) - \boldsymbol{\Psi}_t(\mathbf{x}_j) \right\|^2. \tag{2.8}$$

The kernel width parameter σ represents the rate at which the similarity between two points decays[2]. 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.

We have used a Gaussian Kernel scaled with parameter σ. This kernel works well for the two applications we have studied given both trade sparseness and affinity of the point set. FARDIFF however can be defined by way of a wide class of positive definite kernels, see [6,7] where the choice of kernel is typically application dependent. In addition to the choice of kernel, the trade off in sparsity can be handled by using a restricted isometry, see for example [1].

3. Fuzzy Adaptive Resonance Theory

In this section we describe the second part of FARDiff which uses Fuzzy Adaptive Resonance Theory (FA) [3,17,20] for clustering data points whose dimension has been reduced using the method of section 2. See [9,10,18,19,21].

---

[1]Note that here we require that $N \ge L$. This is not a strong requirement given the reduced dimension L is typically small and the number of points are always larger.

[2]Our choice of σ is determined by a trade off between sparseness of the kernel matrix (small σ) with adequate characterization of true affinity of two points.

FA allows stable recognition of clusters in response to both binary and real-valued input patterns with either fast or slow learning. 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 Figure 1. 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

$$T_j = T_j(\mathbf{x}, \mathbf{w}_j, \alpha) = \frac{\left|\mathbf{x} \wedge \mathbf{w}_j\right|}{\alpha + \left|\mathbf{w}_j\right|}, \tag{3.1}$$

where $\wedge$ is the fuzzy AND operator defined as follows. Given $\mathbf{y} = \{\mathbf{y}_i, \, i = 1, \ldots, N\} \subset \mathbb{R}^m$ and $\mathbf{y}' = \{\mathbf{y}'_i, \, i = 1, \ldots, N\} \subset \mathbb{R}^m$, define

$$\left(\mathbf{y} \wedge \mathbf{y}'\right)_i = \min_i \left(\mathbf{y}_i, \mathbf{y}'_i\right), \tag{3.2}$$

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,

$$T_J = \max_j \{T_j\}. \tag{3.3}$$

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 closed matched. If the match meets the vigilance criterion,

$$\rho \leq \frac{\left|\mathbf{x} \wedge \mathbf{w}_j\right|}{\left|\mathbf{x}\right|}, \tag{3.4}$$

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

$$\mathbf{w}_j(\text{new}) = \beta\left(\mathbf{x} \wedge \mathbf{w}_j(\text{old})\right) + \left(1 - \beta\right)\mathbf{w}_j(\text{old}), \tag{3.5}$$

where $\beta \in [0, 1]$ is the learning rate parameter. 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.

Some of ART's (Adaptive Resonance Theory) advantages are stability, biological plausibility, and responsiveness to the stability-plasticity dilemma. Further advantages are scalability, speed, configurability, and potential for parallelization. ART has been found to have good ability to interpret well, results on neural net learning and decision based networks with low complexity and good robustness [16].

4. Applications in Cancer Detection and Hyperspectral Clustering

The algorithm which we discuss in this paper which we call FARDiff (Fuzzy Adaptive Resonance Diffusion) combines 1. Diffusion Maps and 2. Fuzzy Adaptive Resonance Theory to do clustering on high dimensional data. This algorithm was introduced in the papers [9, 10, 18, 19, 21].

In [18, 21], we applied FARDiff to investigate cancer detection. Early detection of the site of the origin of a tumor is particularly important for cancer diagnosis and treatment. The employment of
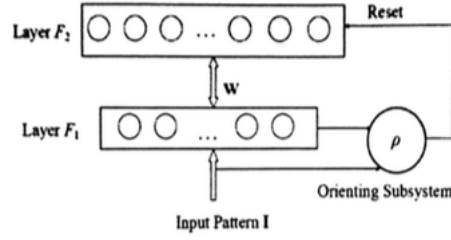
Figure 1: Topological structure of Fuzzy ART. Layers F_1 and F_2 are connected via adaptive weights $\mathbf{W}$. The orienting subsystem is controlled by the vigilance parameter ρ.

gene expression profiles for different cancer types or subtypes has already shown significant advantages over traditional cancer classification methods. We applied FARDiff 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, and compared results with other widely-used clustering algorithms. The experimental results demonstrate the effectiveness of our method in extracting useful information from the high-dimensional data sets.

In [9,10,19], we applied FARDiff to study hyperspectral imaging. The images are sampled at hundreds of frequencies, and the bands are regularly spaced, thus a continuous spectrum can be drawn for every pixel in the image. We reassembled our data into a "hypercube," called the hyperspectral data. Due to the fact that each pixel of the image contains more than a hundred bands, the occurrence of large amounts of hyperspectral data brings important challenges to storage and processing. We used FARDiff to investigate clustering of high dimensional hyperspectral image data from core samples provided by AngloCold Ashanti, and compared to results obtained by AngloGold Ashanti's proprietary method. The experimental results demonstrate the potential of our method in addressing the complicated hyperspectral data and identifying the minerals in core samples.

5. Open Questions For Future Research

In this section, we discuss some on going work. We are interested in extending the FARDiff algorithm to a biclustering framework and high dimensional data reduction. Biclustering is a technique which performs simultaneous clustering in many dimensions automatically integrating feature selection to clustering without any prior information. Two examples of good biclustering algorithms are BARTMAP and HBiFAM (Hierarchical Biclustering FA algorithm) [16, 22].

It is well known that clustering has been used extensively in the analysis of high-throughput messenger RNA (mRNA) expression profiling with microarrays. This technique is restrictive, in part, due to the existence of many uncorrelated genes with respect to sample or condition clustering, or many unrelated samples or conditions with respect to gene clustering. Biclustering offers a solution to such problems by performing simultaneous clustering on both dimensions, or automatically integrating feature selection to clustering without any prior information, so that the relations of clusters of genes (generally, features) and clusters of samples or conditions (data objects) are established. Challenges which need to be addressed using this method include computational complexity and high dimensional data reduction. Current work focuses on developing a natural framework for an analog of FARDiff for biclustering and related computational complexity challenges, for example, traveling salesman problems. A second question relates to applications of FARDiff to more generalized clustering performance. A third research project investigates unified learning schemes and hardware implementation with ART and FARDiff in clustering and biclustering. A valuable new area of innovation will be the application of FARDiff to more generalized data structures such as trees and grammars. Continued progress on distributed representations is valu-

able because of increased data representation capability, both in terms of system capacity and template complexity. Several aspects of these questions are being investigated in ongoing work.

References

[1] E. J. Candés and T. Tao. Near-optimal signal recovery from random projections: universal encoding strategies. *IEEE Trans. Inform. Theory*, (52)(2006), pp 5406-5425.

[2] G. Carpenter, S. Grossberg, N. Markuzon, J. Reynolds, and D. Rosen, Fuzzy ARTMAP: A neural network architecture for incremental supervised learning of analog multidimensional maps, *IEEE Transactions on Neural Networks*, (3)(1992), pp 698-713.

[3] G. Carpenter, S. Grossberg, and D. Rosen, Fuzzy ART: Fast Stable Learning and Categorization of Analog Patterns by an Adaptive Resonance, *Neural Networks*, (4)(1991), pp 759-771.

[4] K. Cawse, S. B. Damelin, R. McIntyre, M. Mitchley, L. du Plessis and M. Sears, An Investigation of data compression for Hyperspectral core image data, *Proceedings of the Mathematics in Industry Study Group, South Africa*, 2008, pp 1-25.

[5] K. Cawse, M. Sears, A. Robin, S. B. Damelin, K. Wessels, F. van den Bergh, R. Mathieu, Using random matrix theory to determine the number of endmembers in a hyperspectral image, *Proceedings of WHISPERS 2010, Reykjavik, Iceland*, pp 45-52.

[6] R. Coifman and S. Lafon, Diffusion maps. *Journal of Applied and Computational Harmonic Analysis*, April 2006, pp 5-30.

[7] S. B. Damelin, J. Levesley, D. L. Ragozin and X. Sun, Energies, Group Invariant Kernels and Numerical Integration on Compact Manifolds, *Journal of Complexity*, (25)(2009), pp 152-162.

[8] S. B. Damelin and W. Miller, Mathematics and Signal Processing, *Cambridge Texts in Applied Mathematics*, no. 48, February 2012.

[9] L. du Plessis, R. Xu, S. B. Damelin, M. Sears and D. C. Wunsch, Reducing dimensionality of hyperspectral data with diffusion maps and clustering with K-means and fuzzy art, *Int. J. Systems Control and Communications*, (3)(2011), pp 232-251.

[10] L. du Plessis, R. Xu, S. B. Damelin, M. Sears and D. C. Wunsch, Reducing dimensionality of hyperspectral data with diffusion maps and clustering with K-means and fuzzy art, *Proceedings of IJCNN 2009*, pp 32-36.

[11] C. Fefferman, S. B. Damelin and W. Glover, *BMO Theorems for ε distorted diffeomorphisms on $\mathbb{R}^D$ and an application to comparing manifolds of speech and sound*, Involve 5-2(2012), pp 159-172.

[12] S. Madeira and A. Oliveira, Biclustering algorithms for biological data analysis: A survey, *IEEE Transactions on Computational Biology and Bioinformatics*, 1-1(2004), pp 24-45.

[13] M. Mitchley, M. Sears and S. B. Damelin, Target detection I Hyperpectral mineral data using wavelet analysis, *Proceedings of the 2009 IEEE Geosciences and Remote Sensing Symposium, Cape Town*, pp 23-45.

[14] S. Mulder and D. Wunsch II, Million city traveling salesman problem solution by divide and conquer clustering with adaptive resonance neural networks, *Neural Networks*, (16)(2003), pp 827-832.

[15] J. de la Porte, B. M. Herbst, W. Hereman, S. J. van der Walt, An Introduction to Diffusion Maps, *The 19th Symposium of the Pattern Recognition Association of South Africa*. 2008.

[16] D. Wunsch II, ART Properties of Interest in Engineering Applications, *Proceedings of International Conference of Neural Networks, Atlanta*, 2009.

[17] R. Xu and D. Wunsch II. Clustering. *IEEE/Wiley*, 2008.

[18] R. Xu, S. B. Damelin, and D. C. 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*, 2007, pp 4613-4616.

[19] R. Xu, L. du Plessis, S. B. Damelin, M. Sears, and D. C. Wunsch II, Analysis of Hyperspectral Data with Diffusion Maps and Fuzzy ART. In *Proceedings of the 2009 international joint conference on Neural Networks*, pp 2302-2309.

[20] R. Xu and D. Wunsch II, Survey of Clustering Algorithms, *IEEE Transactions on Neural Networks*, 16-3(2005), pp 645-678.

[21] R. Xu, S. B. Damelin, B. Nadler, and D. C. Wunsch II, Clustering of High-Dimensional Gene Expression Data with Feature Filtering Methods and Diffusion Maps, *Bio-Medical Engineering and Informatics*, (1)(2008), pp 245-249.

[22] R. Xu and D. C. Wunsch II, BARTMAP: A Viable Structure for Biclustering, *Neural Networks*, (24)(7)(2011), pp 709-716.