Appl. Comput. Harmon. Anal. 36 (2014) 79–107

Contents lists available at ScienceDirect

Applied and Computational Harmonic Analysis

www.elsevier.com/locate/acha

Diffusion maps for changing data

Ronald R. Coifman, Matthew J. Hirn *

Yale University, Department of Mathematics, P.O. Box 208283, New Haven, CT 06520-8283, USA

ARTICLE INFO

Article history:
Received 25 August 2012
Received in revised form 5 March 2013
Accepted 5 March 2013
Available online 14 March 2013
Communicated by Charles K. Chui

Keywords:
Diffusion distance
Graph Laplacian
Manifold learning
Dynamic graphs
Dimensionality reduction
Kernel method
Spectral graph theory

ABSTRACT

Graph Laplacians and related nonlinear mappings into low dimensional spaces have been shown to be powerful tools for organizing high dimensional data. Here we consider a data set X in which the graph associated with it changes depending on some set of parameters. We analyze this type of data in terms of the diffusion distance and the corresponding diffusion map. As the data changes over the parameter space, the low dimensional embedding changes as well. We give a way to go between these embeddings, and furthermore, map them all into a common space, allowing one to track the evolution of X in its intrinsic geometry. A global diffusion distance is also defined, which gives a measure of the global behavior of the data over the parameter space. Approximation theorems in terms of randomly sampled data are presented, as are potential applications.

1. Introduction

In this paper we consider a changing graph depending on certain parameters, such as time, over a fixed set of data points. Given a set of parameters of interest, our goal is to organize the data in such a way that we can perform meaningful comparisons between data points derived from different parameters. In some scenarios, a direct comparison may be possible; on the other hand, the methods we develop are more general and can handle situations in which the changes to the data prevent direct comparisons across the parameter space. For example, one may consider situations in which the mechanism or sensor measuring the data changes, perhaps changing the observed dimension of the data. In order to make meaningful comparisons between different realizations of the data, we look for invariants in the data as it changes. We model the data set as a normalized, weighted graph, and measure the similarity between two points based on how the local subgraph around each point changes over the parameter space. The framework we develop will allow for the comparison of any two points derived from any two parameters within the graph, thus allowing one to organize not only along the data points but the parameter space as well.

An example of this type of data comes from hyperspectral image analysis. A hyperspectral image is in fact a set of images of the same scene that are taken at different wavelengths. Put together, these images form a data cube in which the length and width of the cube correspond to spatial dimensions, and the height of the cube corresponds to the different wavelengths. Thus each pixel is in fact a vector corresponding to the spectral signature of the materials contained in that pixel. Consider the situation in which we are given two hyperspectral images of the same scene, and we wish to highlight the anomalous (e.g., man made) changes between the two. Assume though, that for each data set, different cameras were

* Corresponding author.
E-mail addresses: coifman@math.yale.edu (R.R. Coifman), matthew.hirn@yale.edu (M.J. Hirn).
URL: http://www.math.yale.edu/~mh644 (M.J. Hirn).

1063-5203/$ – see front matter
http://dx.doi.org/10.1016/j.acha.2013.03.001

used which measured different wavelengths, perhaps also at different times of day under different weather conditions. In such a scenario a direct comparison of the spectral signatures between different days becomes much more difficult. Current work in the field often times goes under the heading change detection, as the goal is to often find small changes in a large scene; see [1] for more details.

Other possible areas for applications come from the modeling of social networks as graphs. The relationships between people change over time and determining how groups of people interact and evolve is a new and interesting problem that has usefulness in marketing and other areas. Financial markets are yet another area that lends itself to analysis conducted over time, as are certain evolutionary biological questions and even medical problems in which patient tests are updated over the course of their lives.

The tools developed in this paper are inspired by high dimensional data analysis, in which one assumes that the data has a hidden, low dimensional structure (for example, the data lies on a low dimensional manifold). The goal is to construct a mapping that parameterizes this low dimensional structure, revealing the intrinsic geometry of the data. We are interested in high dimensional data the evolves over some set of parameters, for example time. We are particularly interested in the case in which one does not have a given metric by which to compare the data across time, but can only compare data points from the same time instance. The hyperspectral data situation described above is one such example of this scenario; due to the differing sensor measurements at different times, a direct comparison of images is impossible.

Let $\mathcal{I}$ denote our parameter space, and let X_α, with $\alpha \in \mathcal{I}$, be the data in question. The elements of our data set are fixed, but the graph changes depending on the parameter α. In other words, there is a known bijection between X_α and X_β for $\alpha, \beta \in \mathcal{I}$, but the corresponding graph weights of X have changed between the two parameters. For a fixed α, the diffusion maps framework developed in [2] gives a multiscale way of organizing X_α. If X_α has a low dimensional structure, then the diffusion map will take X_α into a low dimensional Euclidean space that characterizes its geometry. More specifically, the diffusion mapping maps X_α into a particular ℓ^2 space in which the usual ℓ^2 distance corresponds to the diffusion distance on X_α; in the case of a low dimensional data set, the ℓ^2 space can be "truncated" to $\mathbb{R}^d$, with the standard Euclidean distance. However, for different parameters α and β, the diffusion map may take X_α and X_β into different ℓ^2 spaces, thus meaning that one cannot take the standard ℓ^2 distance between the elements of these two spaces. Our contribution here is to generalize the diffusion maps framework so that it works independently of the parameter α. In particular, we derive formulas for the distance between points in different embeddings that are in terms of the individual diffusion maps of each space. It is even possible to define a mapping from one embedding to the other, so that after applying this mapping the standard ℓ^2 distance can once again be used to compute diffusion distances. In particular, this additional mapping gives a common parameterization of the data across all of $\mathcal{I}$ that characterizes the evolving intrinsic geometry of the data. Once this generalized framework has been established, we are able to define a global distance between all of X_α and X_β based on the behavior of the diffusions within each data set. This distance in turn allows one to model the global behavior of X_α as it changes over $\mathcal{I}$.

Earlier results that use diffusion maps to compare two data sets can be found in [3]. Furthermore, there is recent work contained in [4] that also involves combining diffusion geometry principles via tree structures with evolving graphs. In [5], the author considers the case of an evolving Riemannian manifold on which a diffusion process is spreading as the manifold evolves. In our work, we separate out the two processes, effectively using the diffusion process to organize the evolution of the data. Also tangentially related to this work are the results contained in [6] on shape analysis, in which shapes are compared via their heat kernels. More generally, this paper fits into the larger class of research that utilizes nonlinear mappings into low dimensional spaces in order to organize potentially high dimensional data; examples include locally linear embedding (LLE) [7], ISOMAP [8], Hessian LLE [9], Laplacian eigenmaps [10], and the aforementioned diffusion maps [2].

An outline of this paper goes as follows: in the next section, we take care of some notation and review the diffusion mapping first presented in [2]. In Section 3 we generalize the diffusion distance for a data set that changes over some parameter space, and show that it can be computed in terms the spectral embeddings of the corresponding diffusion operators. We also show how to map each of the embeddings into one common embedding in which the ℓ^2 distance is equal to the diffusion distance. The global diffusion distance between graphs is defined in Section 4; it is also seen to be able to be computed in terms of the eigenvalues and eigenfunctions of the relevant diffusion operators. In Section 5 we set up and state two random sampling theorems, one for the diffusion distance and one for the global diffusion distance. The proofs of these theorems are given in Appendix B. Section 6 contains some applications, and we conclude with some remarks and possible future directions in Section 7.

2. Notation and preliminaries

In this section we introduce some basic notation and review certain preliminary results that will motivate our work.

2.1. Notation

Let $\mathbb{R}$ denote the real numbers and let $\mathbb{N} \triangleq \{1, 2, 3, \ldots\}$ be the natural numbers. Often we will use constants that depend on certain variables or parameters. We let $C(\cdot)$, $C_1(\cdot)$, $C_2(\cdot)$, etc., denote these constants; note that they can change from line to line.