Appl. Comput. Harmon. Anal. • • • (• • • •) • • •–• • •

Contents lists available at ScienceDirect

Applied and Computational Harmonic Analysis

www.elsevier.com/locate/acha

Diffusion representations

Moshe Salhov, Amit Bermanis, Guy Wolf, Amir Averbuch *

School of Computer Science, Tel Aviv University, Tel Aviv 69978, Israel

ARTICLE INFO

Article history:
Received 19 November 2015
Received in revised form 2 October 2016
Accepted 7 October 2016
Available online xxxx
Communicated by Gregory Beylkin

Keywords:
Manifold learning
Kernel PCA
Diffusion Maps
Diffusion distance
Distance preservation

ABSTRACT

Diffusion Maps framework is a kernel based method for manifold learning and data analysis that defines diffusion similarities by imposing a Markovian process on the given dataset. Analysis by this process uncovers the intrinsic geometric structures in the data. Recently, it was suggested to replace the standard kernel by a measure-based kernel that incorporates information about the density of the data. Thus, the manifold assumption is replaced by a more general measure-based assumption.

The measure-based diffusion kernel incorporates two separate independent representations. The first determines a measure that correlates with a density that represents normal behaviors and patterns in the data. The second consists of the analyzed multidimensional data points.

In this paper, we present a representation framework for data analysis of datasets that is based on a closed-form decomposition of the measure-based kernel. The proposed representation preserves pairwise diffusion distances that does not depend on the data size while being invariant to scale. For a stationary data, no out-of-sample extension is needed for embedding newly arrived data points in the representation space. Several aspects of the presented methodology are demonstrated on analytically generated data.

1. Introduction

Kernel methods constitute of a wide class of algorithms for non-parametric data analysis of massive high dimensional datasets. Typically, a limited set of underlying factors generates the high dimensional observable parameters via non-linear mappings. The non-parametric nature of these methods enables to uncover hidden structures in the data. These methods extend the well known Multi-Dimensional Scaling (MDS) [6,14] method. They are based on an affinity kernel construction that encapsulates the relations (distances, similarities or correlations) among multidimensional data points. Spectral analysis of this kernel provides an efficient representation of the data that simplifies its analysis. Methods such as Isomap [29], LLE [22], Laplacian eigenmaps [1], Hessian eigenmaps [8] and local tangent space alignment [31,33], extend

* Corresponding author. Fax: +972 3 6422020.
E-mail address: amir@math.tau.ac.il (A. Averbuch).

http://dx.doi.org/10.1016/j.acha.2016.10.003
1063-5203/

the MDS paradigm by assuming to satisfy the manifold assumption. Under this assumption, the data is assumed to be sampled from a low intrinsic dimensional manifold that captures the dependencies between the observable parameters. The corresponding spectral-based embedding performed by these methods preserves the geometry of the manifold that incorporates the underlying factors in the data.

The diffusion maps (DM) method [5] is a kernel-based method that defines diffusion similarities for data analyzes by imposing a Markovian process over the dataset. It defines a transition probability operator based on local affinities between multidimensional data points. By spectral decomposition of this operator, the data is embedded into a low dimensional Euclidean space, where Euclidean distances represent the diffusion distances in the ambient space. When the data is sampled from a low dimensional manifold, the diffusion paths follow the manifold and the diffusion distances capture its geometry.

DM embedding was utilized for a wide variety of data and pattern analysis techniques. For example it was used to improve audio quality by suppressing transient interference [28]. It was utilized in [25] for detecting and classifying moving vehicles. Additionally, DM was applied to scene classification [12], gene expression analysis [24] and source localization [27]. Furthermore, the DM method can be utilized for fusing different sources of data [16,13].

DM embeddings in both the original version [5,15] and in the measure-based Gaussian correlation (MGC) version [4,3], are obtained by the principal eigenvectors of the corresponding diffusion operator. These eigenvectors represent the long-term behavior of the diffusion process that captures its metastable states [11] as it converges to a unique stationary distribution.

The MGC framework [4,3] enhances the DM method by incorporating information about data distribution in addition to the local distances on which DM is based. This distribution is modeled by a probability measure, which is assumed to quantify the likelihood of data presence over the geometry of the space. The measure and its support in this method replace the manifold assumption. Thus, the diffusion process is accelerated in high density areas in the data rather than being depended solely on the manifold geometry. As shown in [4], the compactness of the associated integral operator enables to achieve dimensionality reduction by utilizing the DM framework.

This MGC construction consists of two independent data points representations. The first represents the domain on which the measure is defined or, equivalently, the support of the measure. The second represents the domain on which the MGC kernel function and the resulting diffusion process are defined. These *measure domain* and the *analyzed domain* may, in some cases, be identical, but separate sets can also be considered by the MGC-based construction. The latter case utilizes a training dataset, which is used as the measure domain to analyze any similar data that is used as the analyzed domain. Furthermore, instead of using the collected data as an analyzed domain, it can be designed as a dictionary or as a grid of representative data points that capture the essential structure of the MGC-based diffusion.

In general, kernel methods can find geometrical meaning in a given data via the application of spectral decomposition. However, this representation changes as additional data points are added to the given dataset. Furthermore, the required computational complexity, which is dictated by spectral decomposition, is $O(n^3)$ that is not feasible for a very large dataset. For example, a segmentation of a medium size image of 512×512 pixels requires a kernel matrix of size $2^{18} \times 2^{18}$. The size of such matrix necessitated about 270 GByte of memory assuming double precision. Spectral decomposition procedure applied to such a matrix is a formidable slow task. Hence, there is a growing need to have more computationally efficient methods that are practical for processing large datasets.

Recently, a method to produce random Fourier features from a given data and a positive kernel was proposed in [21]. The suggested method provides a random mapping of the data such that the inner product of any two mapped points approximates the kernel function with high probability. This scheme utilizes Bochner's theorem [23] that says that any such kernel is a Fourier transform of a uniquely defined probability measure. Later, this work was extended in [17,30] to find explicit features for image classification.

Please cite this article in press as: M. Salhov et al., Diffusion representations, Appl. Comput. Harmon. Anal. (2016), http://dx.doi.org/10.1016/j.acha.2016.10.003