Appl. Comput. Harmon. Anal. 41 (2016) 190–213

Contents lists available at ScienceDirect

Applied and Computational Harmonic Analysis

www.elsevier.com/locate/acha

Diffusion-based kernel methods on Euclidean metric measure spaces

Amit Bermanis [a], Guy Wolf [b], Amir Averbuch [b,*]

[a] *Department of Applied Mathematics, School of Mathematical Sciences, Tel Aviv University, Tel Aviv 69978, Israel*
[b] *School of Computer Science, Tel Aviv University, Tel Aviv 69978, Israel*

ARTICLE INFO

Article history:
Received 25 September 2014
Received in revised form 23 July 2015
Accepted 25 July 2015
Available online 30 July 2015
Communicated by Spec. Issue Guest Editor

Keywords:
Diffusion maps
Kernel methods
Diffusion-based kernel
Spectral analysis
Measure-based information
Manifold learning

ABSTRACT

Diffusion-based kernel methods are commonly used for analyzing massive high dimensional datasets. These methods utilize a non-parametric approach to represent the data by using an affinity kernel that represents similarities, distances or correlations between data points. The kernel is based on a Markovian diffusion process, whose transition probabilities are determined by local distances between data points. Spectral analysis of this kernel provides a representation of the data, where Euclidean distances correspond to diffusion distances between data points. When the data lies on a low dimensional manifold, these diffusion distances encompass the geometry of the manifold. In this paper, we present a generalized approach for defining diffusion-based kernels by incorporating measure-based information, which represents the density or distribution of the data, together with its local distances. The generalized construction does not require an underlying manifold to provide a meaningful kernel interpretation but assumes a more relaxed assumption that the measure and its support are related to a locally low dimensional nature of the analyzed phenomena. This kernel is shown to satisfy the necessary spectral properties that are required in order to provide a low dimensional embedding of the data. The associated diffusion process is analyzed via its infinitesimal generator and the provided embedding is demonstrated in two geometric scenarios.

1. Introduction

The utilization of kernel methods is a common practice in a non-parametric data analysis of massive high dimensional datasets. Usually, a limited set of underlying factors generates the high dimensional observable parameters via non-linear mappings. The non-parametric nature of this analysis overcomes the redundancies of the observable parameters and uncovers their underlying factors. These methods extend the well known MDS [9,18] method. They are based on a construction of an affinity kernel that encapsulates the relations

[*] Corresponding author. Fax: +972 153 54 5694455.
E-mail address: amir@math.tau.ac.il (A. Averbuch).

http://dx.doi.org/10.1016/j.acha.2015.07.005

(distances, similarities or correlations) between data points. Spectral analysis of this kernel provides an efficient representation of the data that simplifies its analysis.

The MDS method uses the eigenvectors of a Gram matrix, which contains the inner products between the data points in the analyzed dataset, to define a mapping of data points into an embedded space that preserves most of these inner products. This method is equivalent to PCA [17,16], which projects the data onto the span of the principal directions of the variance of the data. Both of these methods capture linear structures on the data. They separate between meaningful directions, which represent the distribution of the data, and noisy uncorrelated directions. The former ones are associated with significant eigenvalues (and eigenvectors) of the Gram matrix, while the latter ones are associated with small eigenvalues.

Kernel methods, such as Isomap [28], LLE [25] and Laplacian eigenmaps [1], Hessian eigenmaps [12] and local tangent space alignment [29,31], extend the MDS paradigm by considering locally linear structures in the data. These structures are assumed to form a low dimensional manifold that captures the dependencies between the observable parameters of the data. This is called the manifold assumption, and the data is assumed to be sampled from this manifold. The spectral embedding space in these methods preserves the geometry of the manifold, which incorporates the underlying factors of the data.

The diffusion maps (DM) method [6] is a popular kernel method that utilizes a stochastic diffusion process to analyze the data. It defines diffusion affinities via symmetric conjugation of a transition probability operator. These probabilities are based on local distances between the data points. The Euclidean distances in the embedded space represent the diffusion distances in the original space. When the data is sampled from a low dimensional manifold, the diffusion paths follow the manifold and the diffusion distances capture its geometry.

In this paper, we enhance the DM method by incorporating information about the distribution of the data, in addition to local distances on which DM is based. This distribution is expressed in term of a measure over the observable space. The measure (and its support) replace the manifold assumption. We assume that the measure quantifies the likelihood for the presence of data over the geometry of the space. This assumption is significantly less restrictive than the need to have a manifold present. In practice this measure can either be provided as an input (e.g., by a-priori knowledge or a statistical model), or deduced from a given training set (e.g., by a density estimator). The manifold assumption can be expressed in terms of the measure assumption by setting the measure to be concentrated around an underlying manifold or (in the extremely restrictive case), to be supported by the manifold. Therefore, the measure assumption is not only less restrictive than the manifold assumption but it also generalizes it.

Data sampling densities were considered (and modeled by density measures) in previous variations of the DM framework, such as [6,10,11]. However, such sampling densities are typically an artifact resulting from nonuniform sampling of the underlying geometry, and the analysis does not use them to directly model the geometry of the data. Indeed, the anisotropic kernel [6], for example, is specifically aimed to separate the sampling density from the manifold geometry by either fully or partly canceling its effects on the diffusion process via appropriate kernel normalization. An alternative approach, presented in [10,11], is to use the sampling densities to locally adjust the diffusion scales, which determine the sizes of local data patches over the underlying geometry. In both these cases, the used densities are estimated directly from the sampled data that is used to construct the kernel, using unnormalized version of the kernel itself. In order for this density estimation to be accurate, large amounts of data are required both for measuring and for representing the densities. Such amounts are indeed commonly available in many Big Data applications. However, as in most kernel method, the size of the DM kernel is quadratically related to the size of dataset. Thus, computational requirements limit the sampled dataset size that can be effectively used for its construction in applicative settings. Therefore, tying the density estimation process directly to the kernel construction may be impractical.

In the suggested construction, the used measure is separated from the distances and from the analyzed dataset. As mentioned before, this measure can either represent densities or some other distribution