Appl. Comput. Harmon. Anal. 40 (2016) 207–228

Contents lists available at ScienceDirect

Applied and Computational Harmonic Analysis

www.elsevier.com/locate/acha

Measure-based diffusion grid construction and high-dimensional data discretization

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

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

ARTICLE INFO

Article history:
Received 8 June 2014
Accepted 4 February 2015
Available online 11 February 2015
Communicated by Charles K. Chui

Keywords:
Dimensionality reduction
Kernel PCA
Diffusion-based kernel
Measure-based information
Grid construction
Data discretization

ABSTRACT

The diffusion maps framework is a kernel-based method for manifold learning and data analysis that models a Markovian process over data. Analysis of this process provides meaningful information concerning inner geometric structures in the data. Recently, it was suggested to replace the standard kernel by a measure-based kernel, which incorporates information about the density of the data. Thus, the manifold assumption is replaced by a more general measure assumption.

The measure-based diffusion kernel utilizes two separate independent datasets. The first is the set by which the measure is determined. This measure correlates with a density that represents normal behaviors and patterns in the data. The second set consists of the analyzed data points that are embedded by the metastable states of the underlying diffusion process. This set can either be contiguous or discrete.

In this paper, we present a data discretization methodology for analyzing a contiguous domain. The obtained discretization is achieved by constructing a uniform grid over this domain. This discretization is designed to approximate the continuous measure-based diffusion process by a discrete random walk process. This paper provides a proved criterion to determine the grid resolution that ensures a controllable approximation error for the continuous steady states by the discrete ones. Finally, the presented methodology is demonstrated on analytically generated data.

1. Introduction

Kernel methods constitute a wide class of algorithms for nonparametric data analysis of massive high dimensional datasets. Typically, a limited set of underlying factors generates the high dimensional observable parameters via nonlinear mappings. The nonparametric nature of these methods enables to uncover hidden structures in the data. These methods extend the well known MDS [5,14] 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

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

http://dx.doi.org/10.1016/j.acha.2015.02.001
1063-5203/

data that simplifies its analysis. Methods such as Isomap [21], LLE [19], Laplacian eigenmaps [1], Hessian eigenmaps [8] and local tangent space alignment [26,27], extend the MDS paradigm by considering 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 embedding spaces in these methods preserve the geometry of the manifold, which incorporates the underlying factors of the data.

The diffusion maps (DM) method [4] is a kernel method that models and analyzes a Markovian process over the data. It defines a transition probability operator based on local affinities between the multidimensional data points. By spectral decomposition of this operator, the data is embedded into a low dimensional Euclidean space, where distances 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.

The measure-based Gaussian correlation (MGC) framework [2,3] enhances the DM method by incorporating information about the distribution of the data 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 generalize and replace the manifold assumption. Thus, the diffusion process is accelerated in high density areas of the data, rather than depending solely on a manifold geometry. As shown in [2], the compactness of the associated integral operator enables dimensionality reduction by utilization of the DM framework.

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

In this paper, we present a data discretization methodology for analyzing a contiguous domain also called the analyzed domain. The presented discretization is obtained by constructing a uniform grid that will serve as a discrete version of the analyzed domain. This discretization is designed to approximate the continuous MGC diffusion process by a discrete random walk process. More precisely, the resolution of the uniform grid is related (via an upper bound) to the approximation quality of the steady-state stationary distribution of the MGC diffusion by the discrete stationary distribution of the random walk. Therefore, the size of the constructed grid, which is based on this resolution, is not determined by the size of the analyzed data,[1] but rather by the properties of the underlying MGC diffusion process and its stationary distribution.

The utilization of two separated sets of multidimensional points introduces a different approach for the analysis of very large problems. Big Data is evolving and the size of data becomes bigger every day. Twitter has more than 250 million tweets per day, Google has more than 1 billion queries per day, Facebook has more than 800 million updates per day, and YouTube has more than 4 billion views per day [23]. The data produced nowadays is estimated in the order of zettabytes (10^{10}). Current big data analytics methods focus on distributed and parallel methods such as MapReduce [6] or Hadoop [25]. However, these methods are not always the best analytics tool [16]. Hence, there is a need to find better and more efficient analysis techniques. Distribution based analysis reduces the task of data analytics to the determination of the relation between each multidimensional data point and the distribution. Any data analysis task that can be reduced to this relation (such as clustering, anomaly detection and classification) can be processed in a computational

[1] The grid size is determined by neither the size of the analyzed domain nor by the size of the measure domain, but merely by the analytic properties of the kernel and the underlying measure.