Appl. Comput. Harmon. Anal. 38 (2015) 399–419

Contents lists available at ScienceDirect

Applied and Computational Harmonic Analysis

www.elsevier.com/locate/acha

Approximately-isometric diffusion maps

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

[a] School of Computer Science, Tel Aviv University, Tel Aviv 69978, Israel
[b] Department of Mathematical Information Technology, University of Jyväskylä, Finland

ARTICLE INFO

Article history:
Received 8 July 2013
Received in revised form 25 May 2014
Accepted 30 May 2014
Available online 4 June 2014
Communicated by Amit Singer

Keywords:
Dimensionality reduction
Manifold learning
Kernel PCA
Diffusion maps
Diffusion distance
Distance preservation

ABSTRACT

Diffusion Maps (DM), and other kernel methods, are utilized for the analysis of high dimensional datasets. The DM method uses a Markovian diffusion process to model and analyze data. A spectral analysis of the DM kernel yields a map of the data into a low dimensional space, where Euclidean distances between the mapped data points represent the diffusion distances between the corresponding high dimensional data points. Many machine learning methods, which are based on the Euclidean metric, can be applied to the mapped data points in order to take advantage of the diffusion relations between them. However, a significant drawback of the DM is the need to apply spectral decomposition to a kernel matrix, which becomes infeasible for large datasets.

In this paper, we present an efficient approximation of the DM embedding. The presented approximation algorithm produces a dictionary of data points by identifying a small set of informative representatives. Then, based on this dictionary, the entire dataset is efficiently embedded into a low dimensional space. The Euclidean distances in the resulting embedded space approximate the diffusion distances. The properties of the presented embedding and its relation to DM method are analyzed and demonstrated.

1. Introduction

Recent methods process massive amounts of high dimensional data by utilizing a manifold structure on which data points are assumed to lie. This manifold is immersed in the ambient space that is defined by observable/measurable parameters. Kernel methods are designed to support data analysis tasks by utilizing the intrinsic manifold geometry. These methods are based on a kernel matrix that is designed to quantify the similarity between data points on the manifold. Spectral analysis of the kernel in these methods reveals the internal geometric structure of the data [10]. This analysis decomposes the designed kernel and generates eigenvectors that map the data from the ambient space into an embedded space that is usually

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

http://dx.doi.org/10.1016/j.acha.2014.05.002
1063-5203/

low dimensional. Spectral kernel methods have an impact on a wide range of optimization problems from graph coloring [4,3,2] to image segmentation [27] and web search [8].

Kernel methods extend the classic Multi Dimensional Scaling (MDS) method [21] by replacing its Gram matrix with a kernel matrix whose spectral decomposition preserves similarities between data points instead of preserving the inner products that MDS preserves. Some examples for kernel methods are: LLE [24], Isomaps [32], Laplacian eigenmaps [5], Hessian eigenmaps [15], local tangent space alignment [34,35] and Diffusion Maps [11].

For a sufficiently small dataset, kernel methods can be implemented and executed on relatively standard computing devices. However, even for moderate size datasets, the necessary computational requirements to process them are unreasonable and, in many cases, impractical. For example, a segmentation of a medium size image with 512×512 pixels requires a $2^{18} \times 2^{18}$ kernel matrix. The size of such a matrix necessitated about 270 GB of memory assuming double precision. Furthermore, the spectral decomposition procedure applied to such a matrix will be a formidable slow task. Hence, there is a growing need to have more computationally efficient methods that are practical for processing large datasets.

The main computational load associated with kernel methods is generated by the application of a spectral decomposition to a kernel matrix. Sparsification by a sparse eigensolver such as Lanczos, which computes the relevant eigenvectors [12] of the kernel matrix, is widely used to reduce the computational load involved in processing a kernel matrix. Another sparsification approach is to transform the dense kernel matrix into a sparse matrix by selectively truncating elements outside a given neighborhood radius of each dataset member. Other approaches to achieve matrix sparsification are described in [33]. Given a dataset with n data points, common methods including the one described in this paper for processing kernel methods require at least $O(n^2)$ operations to determine which entries to either calculate or to threshold. While there are methods to alleviate these computational complexities [1], kernel sparsification might result in a significant loss of intrinsic geometric information such as distances and similarities.

A prominent approach to reduces the discussed computational load is based on the Nyström extension method [17], which estimates the eigenvectors needed for an embedding. This approach is based on three phases:

1. The dataset is subsampled uniformly over the set of indices that are randomly chosen without repetition.
2. The subsamples define a smaller (than the dataset size) kernel. SVD is applied to the small kernel.
3. Spectral decomposition of a small kernel is extended by the application of the Nyström extension method to the entire dataset.

This three-phase approach reduces the computational load, but the approximated spectral decomposition output suffers from several major problems. Subsampling affects the quality of the spectral approximation. In addition, the Nyström extension method exhibits ill-conditioned behavior that also affects the spectral approximation [6]. Uniform subsampling of a sufficient number of data points captures most of the data probability distribution. However, rare events, compared to the subsampled size, might get lost. The results from this loss of information degrades the quality of the estimated embedded distances.

The Nyström extension method is based on inverting a kernel matrix that was derived from a uniform sampling. This kernel does not necessarily has a full rank. Therefore, a direct kernel matrix inversion is ill-conditioned. The Moor–Penrose pseudo-inverse operator can overcome the ill-conditioned effect in Nyström extension. However, this solution may generate an inaccurate extension. Therefore, combining Nyström extension with random sampling can result in inaccurate approximations of spectral decomposition.

Recently, a multiscale scheme, which is called multiscale extension (MSE), was suggested in [6]. The scheme, which samples scattered data and extends functions defined on sampled data points, overcomes some of the limitations of the Nyström method. The MSE method is based on mutual distances between