Signal Processing 111 (2015) 137–149

Contents lists available at ScienceDirect

Signal Processing

journal homepage: www.elsevier.com/locate/sigpro

Embedding and function extension on directed graph

Saman Mousazadeh *, Israel Cohen

Department of Electrical Engineering, Technion-Israel Institute of Technology, Technion City, Haifa 32000, Israel

ARTICLE INFO

Article history:
Received 14 July 2014
Received in revised form
13 November 2014
Accepted 16 December 2014
Available online 24 December 2014

Keywords:
Directed graph
Asymmetric kernel
Function extension

ABSTRACT

In this paper, we propose a novel technique for finding the graph embedding and function extension for directed graphs. We assume that the data points are sampled from a manifold and the similarity between the points is given by an asymmetric kernel. We provide a graph embedding algorithm which is motivated by Laplacian type operator on manifold. We also introduce a Nyström type eigenfunctions extension which is used both for extending the embedding to new data points and to extend an empirical function on new data set. For extending the eigenfunctions to new points, we assume that only the distances of the new points from the labelled data are given. Simulation results demonstrate the performance of the proposed method in recovering the geometry of data and extending a function on new data points.

1. Introduction

Recent advances in geometric based methods of data mining and machine learning lead to efficient algorithms for lots of applications such as dimensionality reduction, function extension, classification and clustering, just to name a few. Most of these methods are graph based techniques. Graphs offer an advantageous compromise between their simplicity, interpretability and their ability to express complex relationships between data points. The core idea in such algorithms is to construct a weighted graph on data points such that each vertex of the graph represents a data point, and a weighted edge, connecting two vertices to each other, represents the similarity between the two corresponding data points. In the context of networks (e.g., social networks), the data naturally lead themselves to graph modelling [1]. The graph based representation of data combined with Markov chain techniques exhibits extremely successful results. The main idea here is based on the fact that the eigenvectors of Markov matrices can be regarded as coordinates on the data set. Among vast techniques incorporating Markov chain methods in data processing, kernel eigenmap methods have attracted much research attention recently. The algorithmic consequences of these methods are local linear embedding (LLE) [2], Laplacian eigenmaps [3], Hessian eigenmaps [4], local tangent space alignment [5] and diffusion maps [6].

In most of these kernel eigenmaps based methods, the similarity between points is given by a symmetric positive semi-definite kernel. In some practical applications the similarity between points is not necessarily symmetric. Typical examples are web information retrieval based on hyperlink structure, document classification based on citation graphs [7], web information retrieval based on hyperlink structure, and protein clustering based on pairwise alignment scores [8]. Some works have been done to deal with the ranking problem on link structure of the Web. Although much progress in the field, it is still a hard task to do general data analysis on directed graphs such as classification and clustering. Chen et al. [9] proposed an algorithm for embedding vertices on directed graphs to vector spaces. This algorithm explores the inherent pairwise relation between vertices of the directed graph by using transition probability and the stationary distribution of Markov random walks, and embeds the vertices into vector spaces preserving such relation optimally. Recently,

* Corresponding author.
E-mail addresses: smzadeh@tx.technion.ac.il (S. Mousazadeh), icohen@ee.technion.ac.il (I. Cohen).

http://dx.doi.org/10.1016/j.sigpro.2014.12.019
0165-1684/

Perrault-Joncas and Meilă [10] proposed an algorithm based on the analysis of Laplacian type operators and their continuous limit as generators of diffusions on a manifold. They modelled the observed graph as a sample from a manifold endowed with a vector field, and designed an algorithm that separates and recovers the features of this process: the geometry of the manifold, the data density, and the vector field. The most important shortcoming of these methods is not providing a straightforward procedure to extend the embedding to new points in case only the distances of the new point to the original data points are known, which encountered in applications. In [11], Coifman and Hirn introduced a simple procedure for the construction of a bi-stochastic kernel for an arbitrary data set that is derived from an asymmetric affinity function. The affinity function measures the similarity between points in test set and some reference set.

These geometric based algorithms have been applied in various signal processing applications. In a pioneering work, Shi and Malik [12] used spectral methods for image segmentation. Later many researchers have used geometric based methods in applications such as image clustering [13,14], image completion [15], speech enhancement in the presence of transient noise [16], voice activity detection in the presence of transient noise [17], linear and nonlinear independent component analysis [18,19], parametrization of linear systems [20], and single channel source localization [21]. Most if not all of these applications can be regarded as out of sample function extension, in which an empirical function is extended to unlabelled data. In these applications, usually a large amount of data is involved and the only way to perform a task like clustering, regression, or classification is to subsample the data set $\overline{X}$ in order to reduce the size of the problem, process the new set X, and then extend the results to the original data $\overline{X}$. Coifman and Lafon [22] proposed a geometric harmonics procedure, inspired from the Nyström method, to perform this task. More specifically, they assumed that the similarity between the data points is given by a symmetric positive semi-definite kernel. Then it is shown that the eigenfunctions of the integration operator defined by this kernel form an orthogonal basis for the space of squared integrable functions defined on $\overline{X}$ (i.e. $L^2(\overline{X})$). In order to extend a function defined on the set X to the data set $\overline{X}$, first the eigenfunctions computed on X are extended to the data set $\overline{X}$ using the Nyströmmethod. The function is then approximated as the weighted sum of these extended eigenfunctions. Kushnir et al. [23] and Singer et al. [19] introduced a method for parameterizing high dimensional data into its independent physical parameters, which enables the identification of the parameters and a supervised extension of the re-parametrization to new observations. In their work, a novel diffusion processes was used, utilizing only the small observed set, that approximates the isotropic diffusion on the parametric manifold. They utilized Nyström-type extension of the embedding of that small observed data-set to the embedding into the independent components on a much larger data-set.

In this paper, we propose a novel technique for embedding a directed graph to Euclidean space. We model the observed data as samples from a manifold where the similarity between the points is given by an asymmetric kernel. This asymmetric kernel is modelled utilizing a vector field, and we design an algorithm that separates and recovers the geometry of the manifold, the data density, and the vector field. We further provide a simple Nyström extension procedure which allows us to extend both the embedding and the estimated vector field to new data points. More precisely, we adopt the method presented in [23] into the case when the kernel is asymmetric. The rest of this paper is organized as follows. In Section 2, we provide a model which can be used in directed graph modelling. We also introduce our results regarding the limit of Laplacian type operators and provide an algorithm for obtaining the embedding of a directed graph. We also propose a Nyström extension procedure for extending the embedding and the vector field to new data points. In Section 3 we provide some experimental results. We conclude the paper in Section 4.

2. Problem formulation, embedding and function extension

Let X be a set of n data points sampled according to a distribution $p=e^{-U}$ from an unknown smooth manifold $\mathcal{M}\subset\mathbb{R}^{\ell}$ with intrinsic dimension $d<\ell$. Let G be a directed graph with n nodes constructed from the data set X, where each nodes of the graph (e.g. the node i) corresponds to a point in the set X (e.g. $x_i\in X$). We assume that the edge weight $K_{i,j}$ between nodes i and j is given by a positive asymmetric similarity kernel $k_\epsilon(\cdot,\cdot)$ (i.e. $K_{i,j}=k_\epsilon(x_i,x_j)\geq 0$). We also assume that the directional component of $k_\epsilon(\cdot,\cdot)$ is derived by a vector field $\boldsymbol{r}$ on the manifold $\mathcal{M}$, which will be precisely defined shortly. This vector field $\boldsymbol{r}$ is sufficient to characterize any directionality associated with a drift component and as it turns out, the component of $\boldsymbol{r}$ normal to $\mathcal{M}\subset\mathbb{R}^{\ell}$ can also be used to characterize any source component, see [10] for further discussion. The problem is finding an embedding of G into $\mathbb{R}^m; m\leq d$ which approximates the generative process geometry $\mathcal{M}$, the sampling distribution $p=e^{-U}$, and the directionality $\boldsymbol{r}$. This embedding needs to be consistent as sample size increases and the bandwidth of the kernel vanishes.

2.1. Anisotropic diffusion operator

Any kernel $k_\epsilon(x,y)$ can be decomposed into symmetric and anti-symmetric parts as follows:

$$k_\epsilon(x,y)=h_\epsilon(x,y)+a_\epsilon(x,y), \tag{1}$$

where $h_\epsilon(x,y)=h_\epsilon(y,x)$ is the symmetric component and $a_\epsilon(x,y)=-a_\epsilon(y,x)$ is the antisymmetric component of the kernel. As in [10], we assume that the symmetric and anti-symmetric parts can be written as

$$h_\epsilon(x,y)=\frac{h(\|x-y\|^2/\epsilon)}{\epsilon^{d/2}} \tag{2}$$

$$a_\epsilon(x,y)=\frac{\boldsymbol{r}(x,y)}{2}\cdot(y-x)\frac{h(\|x-y\|^2/\epsilon)}{\epsilon^{d/2}}, \tag{3}$$

respectively, where $\boldsymbol{r}(x,y)=\boldsymbol{r}(y,x)$ and $h\geq 0$ is an arbitrary exponentially decreasing function when $\|x-y\|$ converges to infinity.