Embedding new observations via sparse-coding for non-linear manifold learning

Bogdan Raducanu [a], Fadi Dornaika [b,c,*]

[a] Computer Vision Center, Barcelona, Spain
[b] Department of Computer Science and Artificial Intelligence, University of the Basque Country UPV/EHU, San Sebastian, Spain
[c] IKERBASQUE, Basque Foundation for Science, Bilbao, Spain

ARTICLE INFO

ABSTRACT

Non-linear dimensionality reduction techniques are affected by two critical aspects: (i) the design of the adjacency graphs, and (ii) the embedding of new test data—the out-of-sample problem. For the first aspect, the proposed solutions, in general, were heuristically driven. For the second aspect, the difficulty resides in finding an accurate mapping that transfers unseen data samples into an existing manifold. Past works addressing these two aspects were heavily parametric in the sense that the optimal performance is only achieved for a suitable parameter choice that should be known in advance.

In this paper, we demonstrate that the sparse representation theory not only serves for automatic graph construction as shown in recent works, but also represents an accurate alternative for out-of-sample embedding. Considering for a case study the Laplacian Eigenmaps, we applied our method to the face recognition problem. To evaluate the effectiveness of the proposed out-of-sample embedding, experiments are conducted using the K-nearest neighbor (KNN) and Kernel Support Vector Machines (KSVM) classifiers on six public face datasets. The experimental results show that the proposed model is able to achieve high categorization effectiveness as well as high consistency with non-linear embeddings/manifolds obtained in batch modes.

1. Introduction

Manifold learning refers to the problem of recovering the structure of a manifold from a set of unordered sample data. Manifold learning is often equated with dimensionality reduction, where the goal is to find an embedding or 'unrolling' of the manifold into a lower dimensional space such as certain relationships between samples are preserved. Such embeddings are typically used for visualization. In recent years, a new family of non-linear dimensionality reduction techniques for manifold learning has emerged. The most known are Kernel Principal Component Analysis (KPCA) [1], Locally Linear Embedding (LLE) [2,3], Isomap [4], Supervised Isomap [5], Laplacian Eigenmaps (LE) [6,7]. This family of non-linear embedding techniques appeared as an alternative to their linear counterparts which suffer severe limitation when dealing with real-world data: (i) they assume that the data lie in an Euclidean space and (ii) they may fail to get a faithful representation of data distribution when the number of samples is too small. On the other hand, the non-linear dimensionality techniques are able to discover the intrinsic data structure by exploiting the local topology. In general, they attempt to optimally preserve the local geometry around each data sample while using the rest of the samples to preserve the global structure of the data.

The non-linear methods such as Locally Linear Embedding (LLE), Laplacian Eigenmaps, Isomap, Hessian LLE (hLLE) [8] focus on preserving the local structure of data. LLE formulates the manifold learning problem as a neighborhood-preserving embedding, which learns the global structure by exploiting the local linear reconstructions. It estimates the reconstruction coefficients by minimizing the reconstruction error of the set of all local neighborhoods in the dataset. Isomap extends the classical Multidimensional Scaling (MDS) [9] by computing the pairwise distances in the geodesic space of the manifold. Essentially, Isomap attempts to preserve geodesic distances when data are embedded in the new low dimensional space. Based on the spectral decomposition of the Laplacian of a graph, Laplacian Eigenmaps actually try to find Laplacian eigenfunction on the manifold. Maximum Variance Unfolding (MVU) [10] is a global algorithm for non-linear dimensionality reduction, in which all the data pairs, nearby and far, are considered. MVU attempts to 'unfold' a dataset by pulling the input patterns as far apart as possible subject to constraints that distances and angles between neighboring points are strictly preserved.

* Corresponding author at: Department of Computer Science and Artificial Intelligence, University of the Basque Country UPV/EHU, San Sebastian, Spain. Tel.: +34 943018034; fax: +34 943015590.
E-mail addresses: fadi.dornaika@ehu.es, fdornaika@hotmail.fr (F. Dornaika).

The main issues of the non-linear methods are (1) the quality of embedded space is very sensitive to the choice of free parameters used in the data graph construction [11,12], and (2) they do not provide an explicit mapping function between low and high dimensional spaces [13,14]. Such function is essential for ensuring the continuity of low dimensional representation and projecting data between spaces. Many existing manifold learning techniques do not naturally contain an out-of-sample extension, so research has been undertaken to find ways of extending manifold learning techniques to handle new samples. The out-of-sample extension problem has not received much attention by researchers since it was considered a pure non-linear regression problem [15,16]. Therefore, the out-of-sample problem has been addressed quite satisfactorily by applying Radial Basis Function networks in order to approximate the optimal mapping function [15]. However, the quality of Radial Basis Function networks relies on the careful selection of a few parameters which are chosen empirically [17,18]. In [19], the author presented an algorithm, Locally Smooth Manifold Learning, for learning the structure of a manifold in terms of tangent vectors. Rather than pose manifold learning as the problem of recovering an embedding, they posed the problem in terms of learning a warping function for traversing the manifold using the learned tangent vectors. Smoothness assumptions on this warp allowed the method generalize to unseen data.

In [20], the authors cast MDS, ISOMAP, LLE, and LE in a common framework, in which these methods are seen as learning eigenfunctions of a kernel. The authors try to generalize the dimensionality reduction results for the unseen data samples. In [21], the author proposes a method based on probabilistic mixtures of factor analyzers to (1) model the density of images sampled from such manifolds and (2) recover global parameterizations of the manifold. A globally non-linear probabilistic two-way mapping between coordinates on the manifold and images is estimated by combining several, locally valid, linear mappings. In [22], the authors propose a novel solution which involves approximating the kernel eigenfunction using Gaussian basis functions. They also show how the width of the Gaussian can be tuned to achieve extrapolation. Their method was applied to Maximum Variance Unfolding (MVU) method [10]. In [23], the proposed method works by learning the transformation that maps the neighborhood of the unlearnt sample from the high to the low-dimensional space. This transformation is then applied to the new sample to obtain an estimation of its low-dimensional embedding.

In this paper, we address the out-of-sample extension problem. We adopt the sparse representation approach as an optimal solution to the 'out-of-sample' problem. The sparse representation was recently used as an effective alternative to the parametric construction of the adjacency graph [12]. Without any loss of generality, we chose the Laplacian Eigenmaps as one of the non-linear dimensionality reduction techniques to test our method. We present a generalized out-of-sample extension solution using the recent findings in sparse coding theory. Unlike existing approaches we do not require information to be retained from the learning process, such as the pairwise distance matrix or the resultant eigenvectors, we simply learn the mapping from the original high-dimensional data and its low-dimensional counterpart. Although the proposed method integrates the locality preserving principle in its derivation, it is intended to be independent of any specific manifold learning algorithm.

The paper is structured as follows. In Section 2, we briefly review the Laplacian Eigenmaps as well as the L_1 graph construction. In Section 3, we introduce our proposed approach for the out-of-sample problem based on sparse representation. Section 4 contains the experimental results performed on six face datasets. We evaluate the performance of the proposed out-of-sample method for the face recognition problem. Finally, in Section 5 we present our conclusions.

2. Background

2.1. Review of Laplacian Eigenmaps

Laplacian Eigenmaps is a recent non-linear dimensionality reduction technique that aims to preserve the local structure of data [6]. Using the notion of the Laplacian of a graph, this non-supervised algorithm computes a low-dimensional representation of the dataset by optimally preserving local neighborhood information in a certain sense. We assume that we have a set of N samples $\{\mathbf{x}_i\}_{i=1}^{N} \subset \mathbb{R}^D$. The original LE starts with building a graph on the data samples. In this graph, the nodes represent the data samples and the edges quantify the similarity among pairs of samples. There are several ways for setting the edges of the graph. For instance, the most common strategy is to use a K-nearest-neighbor or ϵ−ball graph, or a full mesh (all pairs are connected). Once the edges are set, one can weigh each edge $\mathbf{x}_i \sim \mathbf{x}_j$ by a symmetric affinity function $W_{ij} = K(\mathbf{x}_i; \mathbf{x}_j)$, typically Gaussian:

$$W_{ij} = \exp\left(-\frac{\|\mathbf{x}_i - \mathbf{x}_j\|^2}{\beta}\right) \tag{1}$$

where β is a suitable positive scalar. It is usually set to the average of squared distances between all pairs.

LE seeks latent points $\{\mathbf{y}_i\}_{i=1}^{N} \subset \mathbb{R}^L$ that minimize $\frac{1}{2}\sum_{i,j}\|\mathbf{y}_i - \mathbf{y}_j\|^2 W_{ij}$, which discourages placing far apart latent points that correspond to similar observed points. If $\mathbf{W} \equiv W_{ij}$ denotes the symmetric affinity matrix and $\mathbf{D}$ is the diagonal weight matrix, whose entries are column (or row, since $\mathbf{W}$ is symmetric) sums of $\mathbf{W}$, then the Laplacian matrix is given $\mathbf{L} = \mathbf{D} - \mathbf{W}$. The objective function can also be written as

$$\frac{1}{2}\sum_{i,j}\|\mathbf{y}_i - \mathbf{y}_j\|^2 W_{ij} = \text{tr}(\mathbf{Z}^T \mathbf{L} \mathbf{Z}) \tag{2}$$

where $\mathbf{Z}^T = \mathbf{Y} = [\mathbf{y}_1, ..., \mathbf{y}_N]$ is the $L \times N$ matrix of embedded data and $tr(\cdot)$ denotes the trace of a matrix. The ith row of the matrix $\mathbf{Z}$ provides the vector $\mathbf{y}_i$—the embedding coordinates of the sample $\mathbf{x}_i$.

The matrix $\mathbf{Z}$ (or equivalently $\mathbf{Y}$) is the solution of the optimization problem:

$$\min_{\mathbf{Z}} \text{tr}(\mathbf{Z}^T \mathbf{L} \mathbf{Z}) \quad \text{s.t.} \quad \mathbf{Z}^T \mathbf{D} \mathbf{Z} = \mathbf{I}, \quad \mathbf{Z}^T \mathbf{L} \mathbf{1} = \mathbf{0} \tag{3}$$

where $\mathbf{I}$ is the identity matrix and $\mathbf{1} = (1, ..., 1)^T$. The first constraint eliminates the trivial solution $\mathbf{Z} = \mathbf{0}$ (by setting an arbitrary scale) and the second constraint eliminates the trivial solution $\mathbf{1}$ (all samples are mapped to the same point). Standard methods show that the embedding matrix is provided by the matrix of eigenvectors corresponding to the smallest eigenvalues of the generalized eigenvector problem:

$$\mathbf{L}\mathbf{z} = \lambda \mathbf{D}\mathbf{z} \tag{4}$$

Let the column vectors $\mathbf{z}_0, ..., \mathbf{z}_{N-1}$ be the solutions of (4), ordered according to their eigenvalues, $\lambda_0 = 0 \le \lambda_1 \le \cdots \le \lambda_{N-1}$. The eigenvector corresponding to eigenvalue 0 is left out and only the next eigenvectors for embedding are used. The embedding of the original samples is given by the row vectors of the matrix $\mathbf{Z}$, that is, $\mathbf{Y} = [\mathbf{y}_1, \mathbf{y}_2, ..., \mathbf{y}_N] = \mathbf{Z}^T$.

$$\mathbf{x}_i \longrightarrow \mathbf{y}_i = (z_1(i), ..., z_L(i))^T \tag{5}$$

where $L < N$ is the dimension of the new space.

From Eq. (4), we can observe that the dimensionality of the subspace obtained by LE is limited by the number of samples N.