Pattern Recognition 60 (2016) 813–823

Contents lists available at ScienceDirect

Pattern Recognition

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

# Flexible constrained sparsity preserving embedding

L. Weng [a,b], F. Dornaika [b,c,*], Z. Jin [a]

[a] School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing, China
[b] University of the Basque Country UPV/EHU, San Sebastian, Spain
[c] IKERBASQUE, Basque Foundation for Science, Bilbao, Spain

ARTICLE INFO

*Article history:*
Received 29 February 2016
Received in revised form
15 June 2016
Accepted 28 June 2016
Available online 30 June 2016

*Keywords:*
Constrained embedding
Sparsity preserving projections
Flexible manifold embedding
Semi-supervised learning
Out-of-sample problem

ABSTRACT

In this paper, two semi-supervised embedding methods are proposed, namely Constrained Sparsity Preserving Embedding (CSPE) and Flexible Constrained Sparsity Preserving Embedding (FCSPE). CSPE is a semi-supervised embedding method which can be considered as a semi-supervised extension of Sparsity Preserving Projections (SPP) integrated with the idea of in-class constraints. Both the labeled and unlabeled data can be utilized within the CSPE framework. However, CSPE does not have an out-of-sample extension since the projection of the unseen samples cannot be obtained directly. In order to have an inductive semi-supervised learning, i.e. being able to handle unseen samples, we propose FCSPE which can simultaneously provide a non-linear embedding and an approximate linear projection in one regression function. FCSPE simultaneously achieves the following: (i) the local sparse structures is preserved, (ii) the data samples with a same label are mapped onto one point in the projection space, and (iii) a linear projection that is the closest one to the non-linear embedding is estimated. Experimental results on eight public image data sets demonstrate the effectiveness of the proposed methods as well as their superiority to many competitive semi-supervised embedding techniques.



## 1. Introduction

In many real world applications, such as face recognition and text categorization, the data are usually provided in a high dimension space. In many real-world problems, collecting a large number of labeled samples is practically impossible. The reasons are twofold. Firstly, these labeled samples can be very few. Secondly, acquiring labels requires expensive human labor. To deal with this problem, semi-supervised embedding methods can be used to project the data in the high-dimensional space into a space with fewer dimensions.

A lot of methods for dimension reduction have proposed. Principal Component Analysis [1] (PCA) and Multidimensional Scaling [2] (MDS) are two classic linear unsupervised embedding methods. Linear Discriminant Analysis [1] (LDA) is a supervised method. In 2000, Locally Linear Embedding [3] (LLE) and Isometric Feature Mapping (ISOMAP) [4] were separately proposed in *science* which laid a foundation of manifold learning. Soon afterward, Belkin et al. proposed Laplacian Eigenmaps [5] (LE). He et al. proposed both Locality Preserving Projection [6] (LPP), essentially a linearized version of LE, and Neighborhood Preserving Embedding [7] (NPE), a linearized version of LLE. LPP and NPE can be interpreted in a general graph embedding framework with different choices of graph structure. Most of these methods are unsupervised methods. Afterwards, sparse representation [8–10] based methods have attracted extensive attention. Lai et al. proposed a 2-D feature extraction method called sparse 2-D projections for image feature extraction [11]. In [12], a robust tensor learning method called sparse tensor alignment (STA) is then proposed for unsupervised tensor feature extraction based on the alignment framework. In [13], multilinear sparse principal component analysis (MSPCA) inherits the sparsity from the sparse PCA and iteratively learns a series of sparse projections that capture most of the variation of the tensor data.

Sparsity Preserving Projection (SPP) is an unsupervised learning method [10]. It can be considered as an extension to NPE since the latter has a similar objective function. However, SPP utilizes sparse representation over the whole data to obtain the affinity matrix.

In the last decade, semi-supervised learning algorithms have been developed to effectively utilize a large amount of unlabeled samples as well as the limited number of labeled samples for real world applications [14–22]. In the past years, many graph-based methods for semi-supervised learning have been developed [23–35].

Constrained Laplacian Eigenmaps [36] (CLE) is a semi-supervised embedding method. CLE constrains the solution space of Laplacian Eigenmaps only to contain embedding results that are

* Corresponding author at: University of the Basque Country UPV/EHU Manuel Lardizabal, 1 20018 San Sebastian, Spain.

http://dx.doi.org/10.1016/j.patcog.2016.06.027
0031-3203/

consistent with the labels. Labeled points belonging to the same class are merged together, labeled points belonging to different classes are separated, and similar points are close to one another. Similarly, Constrained Graph Embedding [37] (CGE) tries to project the data points from a same class onto one single point in the projection space with a constraint matrix.

Flexible Manifold Embedding [38] (FME) is a label propagation method. FME simultaneously estimates the non-linear embedding of unlabeled samples and the linear regression over these non-linear representations. In [39], the authors propose a whole learning process that can provide the data graph and a linear regression within a same framework.

SPP is a successful unsupervised learning method. To extend SPP to a semi-supervised embedding method, we introduce the idea of in-class constraints in CGE into SPP and propose a new semi-supervised method for data embedding named Constrained Sparsity Preserving Embedding (CSPE). The weakness of CSPE is that it can not handle the new coming samples which means a cascade regression should be performed after the non-linear mapping is obtained by CSPE over the whole training samples. Inspired by FME, we add a regression term in the objective function to obtain an approximate linear projection simultaneously when non-linear embedding is estimated and proposed Flexible Constrained Sparsity Preserving Embedding (FCSPE). So in this paper, two semi-supervised embedding methods namely CSPE and FCSPE are proposed. Compared to the existing works, the proposed CSPE retains the advantages of both CGE and SPP. On the other hand, the proposed FCSPE simultaneously estimates the non-linear mapping over the training samples and the linear projection for solving the out-of-sample problem, which is usually not provided by existing graph-based semi-supervised non-linear mapping methods.

This paper is organized as follows. Section 2 reviews the related methods including LPP, SPP, CGE and FME. Section 3 introduces the two proposed semi-supervised methods. Section 4 presents performance evaluations on six face image databases: Yale, ORL, FERET, PIE, Extended Yale B and LFW (the original version and the aligned version), one handwriting image database USPS and an object image database COIL-20. Section 5 presents some concluding remarks.

## 2. Related work

Some mathematical notations are listed and will be used in the next several sections. Let $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n] \in \Re^{m\times n}$ be the data matrix, where $n$ is the number of training samples and $m$ is the dimension of each sample. Let $\mathbf{y} = [y_1, y_2, \ldots, y_n]^T$ be a one-dimensional map of $\mathbf{X}$. Under a linear projection $\mathbf{y}^T = \mathbf{p}^T\mathbf{X}$, each data point $\mathbf{x}_i$ in the input space $\Re^m$ is mapped into $y_i = \mathbf{p}^T\mathbf{x}_i$ in the real line. Here $\mathbf{p} \in \Re^m$ is a projection axis. Let $\mathbf{Y} \in \Re^{d\times n}$ be the data projections in a $d$ dimensional space.

### 2.1. Locality preserving projection

Locality Preserving Projection [6] (LPP) is a classic unsupervised embedding method which aims to preserve the local structure of the data by keeping two sample points close in the projection space when they are similar in the original space. The reasonable criterion of LPP is to optimize the following objective function under some constraints:

$$\min \sum_{i,j} (y_i - y_j)^2 W_{ij}, \tag{1}$$

where $\mathbf{W}$ is the affinity matrix associated with the data and $W_{ij}$ represents the similarity between sample $\mathbf{x}_i$ and sample $\mathbf{x}_j$. Estimating the graph affinity $\mathbf{W}$ from data can be carried out by many graph construction methods [40]. The simplest method is based on the use of KNN graph.

The definition of KNN graph is as follows:

$$W_{ij} = \begin{cases} 1, & \mathbf{x}_i \in \delta_k(\mathbf{x}_j) \quad \text{or} \quad \mathbf{x}_j \in \delta_k(\mathbf{x}_i), \\ 0 & \text{otherwise,} \end{cases} \tag{2}$$

where $\delta_k(\mathbf{x}_i)$ means a set of the $k$ neighbors of $\mathbf{x}_i$.

After some simple algebraic formulations, we obtain:

$$\sum_{i,j} (y_i - y_j)^2 W_{ij} = 2\mathbf{p}^T\mathbf{X}\mathbf{L}\,\mathbf{X}^T\,\mathbf{p}, \tag{3}$$

where $\mathbf{L} = \mathbf{D} - \mathbf{W}$ is the Laplacian matrix and $\mathbf{D}$ is a diagonal matrix with $D_{ii} = \sum_j W_{ij}$.

With the constraint $\mathbf{p}^T\mathbf{X}\,\mathbf{D}\,\mathbf{X}^T\mathbf{p} = 1$, the problem becomes:

$$\min_{\mathbf{p}} \frac{\mathbf{p}^T\mathbf{X}\,\mathbf{L}\,\mathbf{X}^T\mathbf{p}}{\mathbf{p}^T\mathbf{X}\,\mathbf{D}\,\mathbf{X}^T\mathbf{p}}. \tag{4}$$

The optimal $\mathbf{p}$ is given by solving the minimum eigenvalue problem:

$$\mathbf{X}\,\mathbf{L}\,\mathbf{X}^T\mathbf{p} = \lambda\,\mathbf{X}\,\mathbf{D}\,\mathbf{X}^T\mathbf{p}. \tag{5}$$

The eigenvectors $\mathbf{p}_1, \ldots, \mathbf{p}_d$ corresponding to the $d$ smallest eigenvalues are then used as the columns of the projection matrix $\mathbf{P}$, i.e. $\mathbf{P} = [\mathbf{p}_1, \ldots, \mathbf{p}_d]$. The projected samples are obtained by $\mathbf{Y} = \mathbf{P}^T\mathbf{X}$.

### 2.2. Sparsity preserving projection

As LPP tries to preserve the neighborhood structure, Sparsity Preserving Projection [9,10] (SPP) aims to keep the structure over the whole data set by using sparse representation instead of the linear representation of $k$ nearest neighbors to get the weight matrix. For $\mathbf{x}_i$, the representative coefficients of the rest samples are obtained by solving an $\ell_1$ problem:

$$\min_{\mathbf{s}_i} \quad \| \mathbf{s}_i \|_1,$$
$$\text{s. t.} \quad \mathbf{x}_i = \mathbf{X}\,\mathbf{s}_i, \tag{6}$$

where $\mathbf{s}_i = [s_{i1}, \ldots, s_{i(i-1)}, 0, s_{i(i+1)}, \ldots, s_{in}]^T$. The problem of SPP is:

$$\min_{\mathbf{p}} \sum_i \left( y_i - \sum_j s_{ij} y_j \right)^2 = \min_{\mathbf{p}} \sum_i \left( \mathbf{p}^T\mathbf{x}_i - \mathbf{p}^T\mathbf{X}\,\mathbf{s}_i \right)^2. \tag{7}$$

With the constraint $\mathbf{p}^T\mathbf{X}\,\mathbf{X}^T\mathbf{p} = 1$, Eq. (7) becomes:

$$\min_{\mathbf{p}} \frac{\mathbf{p}^T\mathbf{X}(\mathbf{I} - \mathbf{S} - \mathbf{S}^T + \mathbf{S}^T\mathbf{S})\mathbf{X}^T\mathbf{p}}{\mathbf{p}^T\mathbf{X}\,\mathbf{X}^T\mathbf{p}} = \max_{\mathbf{p}} \frac{\mathbf{p}^T\mathbf{X}\,\tilde{\mathbf{S}}\mathbf{X}^T\mathbf{p}}{\mathbf{p}^T\mathbf{X}\,\mathbf{X}^T\mathbf{p}}, \tag{8}$$

where $\tilde{\mathbf{S}} = \mathbf{S} + \mathbf{S}^T - \mathbf{S}^T\mathbf{S}$, and $\mathbf{S} = [\mathbf{s}_1, \ldots, \mathbf{s}_n]^T$. The corresponding eigenvalue problem is:

$$\mathbf{X}\,\tilde{\mathbf{S}}\,\mathbf{X}^T\mathbf{p} = \lambda\,\mathbf{X}\,\mathbf{X}^T\mathbf{p}. \tag{9}$$

The eigenvectors $\mathbf{p}_1, \ldots, \mathbf{p}_d$ corresponding to the $d$ largest eigenvalues are the columns of the sought linear transform, i.e., $\mathbf{P} = [\mathbf{p}_1, \ldots, \mathbf{p}_d]$, and $\mathbf{Y} = \mathbf{P}^T\mathbf{X}$.

### 2.3. Constrained Graph Embedding

Constrained Graph Embedding [37] (CGE) is a semi-supervised non-linear embedding method which uses the label information as additional constraints mapping the samples with a same label to one point in the projection space. We assume that the first $l$ samples are with labels from $c$ classes. In the projection space, a constraint matrix $\mathbf{U}$ is used to keep the samples with a same label