Pattern Recognition 61 (2017) 92–103

Contents lists available at ScienceDirect

Pattern Recognition

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

Matrix exponential based semi-supervised discriminant embedding for image classification

F. Dornaika [a,b,*], Y. El Traboulsi [a]

[a] University of the Basque Country UPV/EHU, San Sebastian, Spain
[b] IKERBASQUE, Basque Foundation for Science, Bilbao, Spain

ARTICLE INFO

Article history:
Received 15 April 2016
Received in revised form
15 July 2016
Accepted 18 July 2016
Available online 20 July 2016

Keywords:
Graph-based semi-supervised learning
Small-sample-size (SSS) problem
Matrix exponential
Semi-supervised discriminant embedding (SDE)
Distance diffusion mapping
Feature extraction
Image classification

ABSTRACT

Semi-supervised Discriminant Embedding (SDE) is the semi-supervised extension of Local Discriminant Embedding (LDE). Since this type of methods is in general dealing with high dimensional data, the small-sample-size (SSS) problem very often occurs. This problem occurs when the number of available samples is less than the sample dimension. The classic solution to this problem is to reduce the dimension of the original data so that the reduced number of features is less than the number of samples. This can be achieved by using Principle Component Analysis for example. Thus, SDE needs either a dimensionality reduction or an explicit matrix regularization, with the shortcomings both techniques may suffer from. In this paper, we propose an exponential version of SDE (ESDE). In addition to overcoming the SSS problem, the latter emphasizes the discrimination property by enlarging distances between samples that belong to different classes. The experiments made on seven benchmark datasets show the superiority of our method over SDE and many state-of-the-art semi-supervised embedding methods.

1. Introduction

High-dimensional data is increasingly used in several domains (especially in computer vision and pattern recognition fields). For this reason, researchers have been interested in dimension reduction methods. These latter aim at transforming data from their original space to a meaningfully low dimensional one where tasks can be achieved in an efficient and accurate way. These methods can be categorized into several groups according to their properties: are they linear or non-linear? Do they use labeled samples or unlabeled ones (supervised, unsupervised and semi-supervised)? Principal Component Analysis (PCA) [1] is the most known dimensionality reduction method. It consists of an unsupervised features extraction method whose purpose is to project data in the direction of the maximal variance of the original data and so to make data reconstruction process as efficient as possible. Local Discriminant Analysis (LDA) [2] is a supervised method that, in addition to dimensionality reduction, helps in data discrimination. Kan et al. and Pang et al. propose the Adaptive Discriminant Learning (ADL) method [3] and the Regularized Locality Preserving Discriminant Embedding (RLPDE) method [4] respectively, which are two variants of LDA. PCA and LDA are intrinsically linear but, using the kernel trick, nonlinear versions have been proposed: Kernel PCA [5] and Kernel Fisher Discriminant analysis (KFD) [6] respectively. Local Discriminant Embedding (LDE) [7] is a supervised method that was proposed to overcome some limitations of LDA. It uses the neighborhood relation between samples when constructing the embedding and thus, is not affected by the distribution of data [8]. Due to their contribution to the dimensionality reduction methods, graph-based methods recently received a lot of attention by researchers [9–14]. Many graph-based nonlinear methods that preserve the intrinsic structure of data have been recently proposed: Locally Linear Embedding (LLE) [15], Laplacian Eigenmaps (LE) [16] and isometric mapping (ISOMAP) [17]. Despite the superior ability of nonlinear methods in improving the classification process when dealing with unlabeled samples, most of them suffer from the out-of-sample problem (i.e., there is no straightforward tool to predict the class of unseen data samples). For this reason, many significant efforts have been made in order to linearize some nonlinear methods by forcing the mapping model to be linear. In this context, Niyogi linearized Laplacian Eigenmaps (LEs) by proposing the Locality Preserving Projections (LPPs) method [18], and He et al. linearized Locally Linear Embedding (LLE) by introducing Neighborhood Preserving Embedding (NPE) [19]. Despite the fact that generally supervised methods outperform unsupervised ones, collecting labeled samples is too expensive or unfeasible in many real applications. In contrast, unlabeled samples are abundant and easy to obtain. For this

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

http://dx.doi.org/10.1016/j.patcog.2016.07.029
0031-3203/

reason, researchers are recently more and more interested in methods that use both labeled and unlabeled samples in the learning phase [20–23]. These methods are called semi-supervised methods. In this kind of approaches, labeled data are used in the aim of bringing samples belonging to the same class and widen distances between samples belonging to different classes. Gaussian Fields and Harmonic Functions (GFHFs) [24] and Local and Global Consistency (LGC) [13] are two well-known label propagation semi-supervised methods. These latter can only work on transductive setting in which labeled samples and unlabeled (test) ones are required during the learning phase. Cai et al. extended LDA into its semi-supervised version SDA [25] by adding a regularizer that preserves data smoothness. In the same way, LDE was also extended to its semi-supervised version Semi-supervised Discriminant Embedding (SDE) [7] by adding a similar regularizer. SDA and SDE are not limited to any transductive setting (i.e., they can project unseen data samples also to the new subspace).

Despite its high discriminative ability, in the absence of regularization, SDE suffers from the small-sample-size (SSS) problem when the number of samples is not sufficient. This problem occurs particularly when the number of observations is smaller than the dimension of a sample.

Using the matrix exponential, in this paper we propose an extended version of SDE called Exponential SDE (ESDE). The proposed method has two motivations. Firstly, it solves the SSS problem without any mandatory dimensionality reduction. Secondly, it preserves the discriminant nature of SDE and improves its efficiency by increasing the distances between samples belonging to different classes. In the literature, it should be noticed that matrix exponential was used for unsupervised and supervised learning schemes. To the best of our knowledge, this paper is the first work using the exponential framework for graph-based semi-supervised learning.

The most important properties that characterize our framework can be summarized as follows:

- ESDE overcomes the SSS problem without any information loss regardless of the dimension of the input samples.
- Unlike many dimensionality reduction methods, ESDE does not suffer from any transductive setting: it is straightforward to predict the embedding of any new (unseen) data sample since ESDE provides a linear transform.
- Unlike label propagation methods whose prediction is limited to the class of samples, our approach includes their embedding into the low dimensional subspace. Therefore, we have the freedom to choose the most convenient classifier.
- Since ESDE is an extension of SDE, it inherits its discriminant property based on the locality preserving scatter matrices, and also improves its performances.

This paper is organized as follows: in Section 2, we will review the most interesting graph-based learning methods, including the classic SDE. Our proposed method is introduced in Section 3.1. Section 4 presents the theoretical analysis of the proposed framework. Experimental work carried out on seven benchmark databases is presented in Section 5. Finally, in Section 6 we present our conclusion.

2. Related work

This section is dedicated to the description of some related state-of-the-art methods, namely: Local Discriminant Embedding (LDE) [7], Exponential Local Discriminant Embedding (ELDE) [26] and Semi-supervised Discriminant Embedding (SDE) [7]. In this paper capital bold letters denote matrices and small bold letters denote vectors.

2.1. Notations and preliminaries

Given N training samples, $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N \in \mathbb{R}^D$ where D is the dimension of each sample, we suppose that $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_l, \mathbf{x}_{l+1}, \ldots, \mathbf{x}_{l+u}] \in \mathbb{R}^{D\times N}$, where l is the number of labeled samples and u is the number of unlabeled ones ($N = l + u$), is the train matrix which includes all training samples. Let $\mathbf{X}_{\mathcal{L}} = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_l] \in \mathbb{R}^{D\times l}$ denote the matrix of labeled samples, and let $y_i \in \{1, 2, \ldots, C\}$ denote the label of $\mathbf{x}_i$ ($i \in \{1, 2, \ldots, l\}$) where C is the number of classes, and n_c ($c \in \{1, 2, \ldots, C\}$) is the number of labeled samples belonging to the class c. We assume that $\mathbf{S}$ is the similarity matrix defined by $\mathbf{S}(i, j) = sim(\mathbf{x}_i, \mathbf{x}_j)$ where $sim(\mathbf{x}_i, \mathbf{x}_j)$ represents a score related to the similarity between samples $\mathbf{x}_i$ and $\mathbf{x}_j$. The Laplacian matrix $\mathbf{L}$ corresponding to $\mathbf{S}$ is defined by $\mathbf{L} = \mathbf{D} - \mathbf{S}$ where $\mathbf{D}$ is a diagonal matrix whose elements are the row (or column since the similarity matrix is symmetric) sums of $\mathbf{S}$. We will define two graphs: the within-class graph $\mathbf{S}_w \in \mathbb{R}^{l\times l}$ and the between-class graph $\mathbf{S}_b \in \mathbb{R}^{l\times l}$. These two graphs are dedicated to labeled samples. The first one is defined by $\mathbf{S}_w(i, j) = sim(\mathbf{x}_i, \mathbf{x}_j)$ if $\mathbf{x}_i$ and $\mathbf{x}_j$ belong to the same class and $\mathbf{S}_w(i, j) = 0$ otherwise. The second one is defined by $\mathbf{S}_b(i, j) = sim(\mathbf{x}_i, \mathbf{x}_j)$ if $\mathbf{x}_i$ and $\mathbf{x}_j$ belong to different classes and $\mathbf{S}_b(i, j) = 0$ otherwise. We assume that $\mathbf{L}_{\rm w}$ and $\mathbf{L}_{\rm b}$ are the Laplacian matrices (defined similarly to $\mathbf{L}$) associated to $\mathbf{S}_{\rm w}$ and $\mathbf{S}_{\rm b}$ respectively.

2.2. Local discriminant embedding (LDE)

LDE is a supervised dimensionality reduction method that aims to maximize the similarity between samples belonging to the same class and the divergence between samples sharing dissimilar classes [27]. To this end, the intrinsic graph $\mathbf{G}_{\rm w}$ and the penalty graph $\mathbf{G}_{\rm b}$ are defined. The subset $N_w(\mathbf{x}_i)$ refers to the neighborhood of $\mathbf{x}_i$ sharing its label (e.g. K_1 nearest neighbors), and $N_b(\mathbf{x}_i)$ refers to the neighborhood of $\mathbf{x}_i$ sharing a label different of its own (e.g. K_2 nearest neighbors). We stress the fact that K_1 and K_2 are two different parameters. The similarity matrices of these two graphs are defined by:

$$\mathbf{S}_w(i, j) = \begin{cases} sim(\mathbf{x}_i, \mathbf{x}_j) & \text{if } \mathbf{x}_j \in N_w(\mathbf{x}_i) \text{ or } \mathbf{x}_i \in N_w(\mathbf{x}_j) \\ 0, & \text{otherwise} \end{cases} \tag{1}$$

$$\mathbf{S}_b(i, j) = \begin{cases} 1 & \text{if } \mathbf{x}_j \in N_b(\mathbf{x}_i) \text{ or } \mathbf{x}_i \in N_b(\mathbf{x}_j) \\ 0, & \text{otherwise} \end{cases} \tag{2}$$

where $sim(\mathbf{x}_i, \mathbf{x}_j)$ is a score that represents the similarity between $\mathbf{x}_i$ and $\mathbf{x}_j$.

The linear transformation matrix that projects samples from their original space into the subspace targeted by LDE is obtained by optimizing the following criteria:

$$\min_{\mathbf{W}} \frac{1}{2} \sum_i^l \sum_j^l \| \mathbf{W}^T(\mathbf{x}_i - \mathbf{x}_j)\|^2 \, \mathbf{S}_w(i, j) = \min_{\mathbf{W}} \mathrm{tr}(\mathbf{W}^T \mathbf{X}_{\mathcal{L}} \mathbf{L}_w \mathbf{X}_{\mathcal{L}}^T \mathbf{W}) \tag{3}$$

$$\max_{\mathbf{W}} \frac{1}{2} \sum_i^l \sum_j^l \| \mathbf{W}^T(\mathbf{x}_i - \mathbf{x}_j)\|^2 \, \mathbf{S}_b(i, j) = \max_{\mathbf{W}} \mathrm{tr}(\mathbf{W}^T \mathbf{X}_{\mathcal{L}} \mathbf{L}_b \mathbf{X}_{\mathcal{L}}^T \mathbf{W}) \tag{4}$$

where tr (.) denotes the trace of a matrix, $\mathbf{X}_{\mathcal{L}}$ is the matrix of labeled samples, $\mathbf{L}_{\rm w}$ and $\mathbf{L}_{\rm b}$ are the Laplacian matrices associated to $\mathbf{S}_{\rm w}$ and $\mathbf{S}_{\rm b}$ respectively and $\mathbf{W}$ is the projection matrix. Optimizing the problem represented by Eqs. (3) and (4) is equivalent to