Pattern Recognition 60 (2016) 275–285

Contents lists available at ScienceDirect

Pattern Recognition

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

Inductive and flexible feature extraction for semi-supervised pattern categorization

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

[a] University of the Basque Country UPV/EHU, Manuel Lardizabal, 1, 20018 San Sebastian, Spain
[b] IKERBASQUE, Basque Foundation for Science, Maria Diaz de Haro 3, 48013 Bilbao, Spain
[c] Doctoral School of Sciences and Technology, Lebanese University, Mitein Street, Tripoli, Lebanon

ARTICLE INFO

Article history:
Received 17 June 2015
Received in revised form
10 December 2015
Accepted 28 April 2016
Available online 22 May 2016

Keywords:
Feature extraction
Semi-supervised discriminant analysis
Graph-based embedding
Out-of-sample extension
Pattern categorization

ABSTRACT

This paper proposes a novel discriminant semi-supervised feature extraction method for generic classification and recognition tasks. This method, called inductive flexible semi-supervised feature extraction, is a graph-based embedding method that seeks a linear subspace close to a non-linear one. It is based on a criterion that simultaneously exploits the discrimination information provided by the labeled samples, maintains the graph-based smoothness associated with all samples, regularizes the complexity of the linear transform, and minimizes the discrepancy between the unknown linear regression and the unknown non-linear projection. We extend the proposed method to the case of non-linear feature extraction through the use of kernel trick. This latter allows to obtain a nonlinear regression function with an output subspace closer to the learned manifold than that of the linear one. Extensive experiments are conducted on ten benchmark databases in order to study the performance of the proposed methods. Obtained results demonstrate a significant improvement over state-of-the-art algorithms that are based on label propagation or semi-supervised graph-based embedding.

1. Introduction

Feature extraction with dimensionality reduction is an important step and essential process in embedding data analysis. By computing an adequate representation of data that has a low dimension, more efficient learning and inference [1–4] can be achieved. There are two main reasons for estimating a low-dimensional representation of high-dimensional data: reducing measurement cost of further data analysis and beating the curse of dimensionality. The dimensionality reduction can be achieved either by feature extraction or feature selection. Feature extraction refers to methods that create a set of new features based on transformations and/or combinations of the original features, while feature selection methods select the most representative and relevant subset from the original feature set [5]. Feature extraction methods can be classified into two main classes: (1) linear methods, and (2) non-linear methods. Besides this categorization, these methods can also be classified into three categories: (i) supervised, (ii) semi-supervised, and (iii) unsupervised.

The linear techniques have been increasingly important in pattern recognition [6–8,3] since they permit a relatively simple mapping of data onto a lower-dimensional subspace, leading to simple and computationally efficient classification strategies. The classical linear embedding methods (e.g., PCA, Linear Discriminant Analysis (LDA), Maximum Margin Criterion (MMC) [9]) and Locally LDA [10] are demonstrated to be computationally efficient and suitable for practical applications, such as pattern classification and visual recognition. PCA projects the samples along the directions of maximal variances and aims to preserve the Euclidean distances between the samples. Unlike PCA which is unsupervised, Linear Discriminant Analysis (LDA) [11,8] is a supervised technique. One limitation of PCA and LDA is that they only see the linear global Euclidean structure.

The non-linear methods such as Locally Linear Embedding (LLE) [12] and Laplacian eigenmaps [13] focus on preserving the local structures. Isomap [14] is a non-linear projection method that globally preserve the data. It also attempts to preserve the geodesic distances between samples.

Although the supervised feature extraction methods had been successfully applied to many pattern recognition applications, they require a full labeling of data samples. It is well-known that it is much easier to collect unlabeled data than labeled samples. The labeling process is often expensive, time consuming, and requires intensive human involvement. As a result, partially labeled datasets are more frequently encountered in real-world problems.

[*] Corresponding author at: University of the Basque Country UPV/EHU, Manuel Lardizabal, 1, 20018 San Sebastian, Spain.
E-mail address: fadi.dornaika@ehu.es (F. Dornaika).

http://dx.doi.org/10.1016/j.patcog.2016.04.024
0031-3203/© 2016 Elsevier Ltd. All rights reserved.

In the last decade, semi-supervised learning algorithms have been developed to effectively use limited number of labeled samples and a large amount of unlabeled samples for real-world applications [15,16]. In the past years, many graph-based methods for semi-supervised learning have been developed. The main advantage of such methods is their ability to identify classes of arbitrary distributions. The use of data-driven graphs has led to many progresses in the field of semi-supervised learning [17–25]. Toward classification, an excellent subspace should be smooth as well as discriminative. Hence, a graph-theoretic learning framework is usually deployed to simultaneously meet the smoothness requirement among nearby points and the discriminative requirement among differently labeled points [26]. In addition to the use of partial labelling in semi-supervised learning, many researchers use pairwise constraints which can be seen as another form of side information [27]. These constraints simply indicate whether two instances are similar (must-link) or dissimilar (cannot-link). They are usually used for getting a linear or non-linear embedding by adding them to the criterion derived from unlabelled data samples [28–31]. The final application is to help spectral clustering recover from an undesirable partition.

From the point of view of manifold learning, semi-supervised extensions can generally improve the performance over their supervised counterparts. Nevertheless, despite the success of many graph-based algorithms in dealing with partially labeled problems [32], there are still some problems that are not properly addressed. Almost all semi-supervised feature extraction techniques can suffer from one of the following limitations:

1. The non-linear semi-supervised approaches do not have, in general, an implicit function that can map unseen data samples. In other words, the non-linear methods provide embedding for only the training data. This is the transductive setting, i.e., the test set coincides with the set of unlabeled samples in the training dataset. Indeed, solving the out-of-sample extension is still an open problem for those techniques adopting non-linear embedding.
2. Almost all proposed semi-supervised approaches target the estimation of a linear transform that maps original data into a low dimensional space. While this simplifies the learning processes and gets rid of the out-of-sample problem, there is no guarantee that such approaches will be optimal for all datasets. The main reason behind this is that the criterion used is already a rigid constraint that contains only the linear mapping. Thus, any coordinate in the low-dimensional space is supposed to be a linear combination of the original features. For that reason, these approaches have not the flexibility to adapt their linear model to a more generic non-linear model.

In addition to the above limitations, it is not clear what would be the performance of the semi-supervised approaches when minimal labeling is used. For instance, in the domain of face recognition the so-called one sample problem can be a challenging issue. In this paper, we propose an Inductive Flexible Semi Supervised Feature Extraction method as well as its kernelized version. The aim is to combine the merits of Flexible Manifold Embedding and the non-linear graph based embedding. The proposed linear method will be flexible since it estimates a non-linear manifold that is the closest to a linear embedding. The proposed kernelized method will be also flexible since it estimates the non-linear manifold that is the closest to a kernel-based embedding. In both proposed methods, the non-linear manifold as well as the mapping (linear transform for the linear regression and the kernel multipliers for the non-linear regression) are simultaneously estimated. This simultaneous estimation is the main reason that makes the proposed frameworks superior to many existing algorithms as it will be shown in the sequel. We can also notice that the dimension of the final embedding obtained by the proposed methods is not limited to the number of classes. This allows the application of any kind of classifiers once the data are embedded in new spaces. In contrast with non-linear dimensionality reduction approaches, our proposed methods have an obvious advantage that the learnt subspace has a direct out-of-sample extension to novel samples, and are thus easily generalized to the entire high-dimensional input space. The main differences between our proposed method and the other state-of-the-art ones are (1) our method is *not based on label propagation* and (2) it *simultaneously* exploits labeled discrimination information, uses graph-based smoothness, and estimates a regression function whose output is close at most to the non-linear model.

The paper is structured as follows. In Section 2, we briefly review the main methods for semi-supervised learning including the graph-based label propagation and the semi-supervised embedding methods. In Section 3, we introduce the IFSSFE method and its kernel variant. Section 4 depicts the experimental results obtained with ten real datasets and compares the performance of the proposed method with those of the competing ones. Finally, in Section 5 we present our conclusions. In the sequel, capital bold letters denote matrices and small bold letters denote vectors.

2. Related work

In order to make the paper self-contained, this section will briefly describe some state-of-the art semi-supervised methods.

2.1. Notation and preliminaries

We define the training data matrix as $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_l, \mathbf{x}_{l+1}, \ldots, \mathbf{x}_{l+u}] \in \mathbb{R}^{D\times(l+u)}$, where $\mathbf{x}_i|_{i=1}^{l}$ and $\mathbf{x}_i|_{i=l+1}^{l+u}$ are the labeled and unlabeled samples, respectively, with l and u being the total numbers of labeled and unlabeled samples and D being the feature dimension. Let $N = l + u$ denote the total number of training samples and n_c be the total number of labeled samples in the cth class and let $\mathbf{X}_{\mathcal{L}} = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_l] \in \mathbb{R}^{D\times l}$ be the labeled samples matrix, with the label of $\mathbf{x}_i$ as $y_i \in \{1, 2, \ldots, C\}$, where C is the total number of classes. Let $\mathbf{S} \in \mathbb{R}^{(l+u)\times(l+u)}$ denote the graph similarity matrix with $S(i, j)$ representing the similarity between $\mathbf{x}_{\mathrm{i}}$ and $\mathbf{x}_{\mathrm{j}}$, i.e., $S(i, j) = sim(\mathbf{x}_i, \mathbf{x}_j)$. In a supervised context, one can also consider two similarity matrices $\mathbf{S}_w$ and $\mathbf{S}_b$ that encode the within class and between class graphs, respectively. $\mathbf{S}_w$ encodes the pairwise similarities among samples having the same label. Thus, $S_w(i, j) = sim(\mathbf{x}_i, \mathbf{x}_j)$ if $\mathbf{x}_i$ and $\mathbf{x}_j$ have the same class label; $S_w(i, j) = 0$, otherwise. Similarly, $\mathbf{S}_b$ encodes the pairwise similarities among samples having different labels. Thus, $S_b(i, j) = sim(\mathbf{x}_i, \mathbf{x}_j)$ if $\mathbf{x}_i$ and $\mathbf{x}_j$ have different labels; $S_b(i, j) = 0$, otherwise. The function $sim(.,.)$ can be any symmetric function that measures the similarity between two samples. This can be given by the cosine or the Gaussian kernel.

For each similarity matrix $\mathbf{S}$, a Laplacian matrix $\mathbf{L}$ can be computed. This latter is given 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 $\mathbf{S}$ matrix. Similar expression can be found for $\mathbf{L}_b$ and $\mathbf{L}_{\mathrm{w}}$. The normalized Laplacian $\hat{\mathbf{L}}$ is defined by $\hat{\mathbf{L}} = \mathbf{I} - \mathbf{D}^{-1/2}\,\mathbf{S}\,\mathbf{D}^{-1/2}$ where $\mathbf{I}$denotes the identity matrix.

We also define a binary label matrix $\mathbf{Y} \in \mathbb{B}^{N\times C}$ associated with the samples with $Y(i, j) = 1$ if $\mathbf{x}_i$ has label $y_i = j$; $Y(i, j) = 0$, otherwise. Similarly, we can define an unknown label matrix denoted by $\mathbf{F} \in \mathbb{R}^{N\times C}$. In a semi-supervised setting, $\mathbf{F} = \begin{pmatrix} \mathbf{F}_{\mathcal{L}} \\ \mathbf{F}_{\mathcal{U}} \end{pmatrix}$ where $\mathbf{F}_{\mathcal{L}} = \mathbf{Y}_{\mathcal{L}}$.