Neurocomputing 167 (2015) 517–527

Contents lists available at ScienceDirect

Neurocomputing

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

Kernel flexible manifold embedding for pattern classification

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

[a] Doctoral School of Sciences and Technology, Lebanese University, Tripoli, Lebanon
[b] University of the Basque Country UPV/EHU, San Sebastian, Spain
[c] IKERBASQUE, Basque Foundation for Science, Bilbao, Spain

ARTICLE INFO

Article history:
Received 8 December 2014
Received in revised form
10 March 2015
Accepted 16 April 2015
Communicated by Zhi yong Liu
Available online 24 April 2015

Keywords:
Manifold learning
Kernel methods
Semi-supervised learning
Graph-based embedding
Out-of-sample extension
Pattern classification

ABSTRACT

Flexible Manifold Embedding (FME) has been recently proposed as a semi-supervised graph-based label propagation method. It aims at estimating simultaneously the optimal prediction labels and its linear regression. It integrates the label fitness, the manifold smoothness and a flexible term that forces the linear regression to be as close as possible to nonlinear one. Despite its good performance compared to its counterparts, FME may lead to poor performance when the geometrical structure of data is highly nonlinear. In this paper, we propose a Kernel version of the Flexible Manifold Embedding (KFME). As in classical FME, KFME uses labeled and unlabeled data to estimate the embedding of unlabeled data and its regression function that can map new data samples. Extensive experiments carried out on eight benchmark datasets show that the proposed KFME can outperform FME as well as many state-of-the-art semi-supervised learning methods.

1. Introduction

Nowadays massive high dimensional data are invading the informatics world. The processing and the manipulation of these massive data are not an easy task. Here lies the importance of manifold learning and dimensionality reduction methods [20,30,27]. These latters attempt to build a model of data (e.g, estimating subspaces) that will be compact without having a significant loss. Dimensionality reduction can be achieved by either feature selection or feature extraction. Indeed, feature selection methods choose the most important features of the data and throw the less important ones. Whereas feature extraction methods reduce features by a transformation or by making combinations of the different features [18].

Principal Component Analysis (PCA) [25] and Local Discriminant Analysis (LDA) [10] are the best known methods for dimensionality reduction. PCA is an unsupervised method that projects samples according to the direction of the maximal variance, preserving Euclidean distances. In contrast to PCA, LDA is a supervised method that helps in classifying data, and in reducing the dimension at the same time. This global method searches for axes that minimize the distance between samples sharing the same label and maximize the distance between samples having different labels.

In order to maintain the intrinsic structure of information, the nonlinear methods like Locally Linear Embedding (LLE) [19], Laplacian Eigenmaps (LE) [3] and isometric mapping (ISOMAP) [24] have been recently developed. Although ISOMAP and LE can overcome the linear limitation of PCA and LDA, they suffer from the out-of-sample problem, i.e., they cannot map new (unseen) examples to their reduced subspace. Usually supervised methods outperform unsupervised ones due to their additional information provided by the labels [14]. However, in many experimental works, collecting labeled data is not a trivial task. Labeled samples can be expensive or unavailable. Here comes the importance of semi-supervised methods that use both labeled and unlabeled samples as training data [6,22]. Semi-supervised graph-based methods are the focus of many research works because of their increasing success [4,32,31,23,26,29]. They profit from labeled information to gather samples of the same class and to separate those of different classes, and profit at the same time from labeled and unlabeled data to maintain the geometric data structure [17,13]. In this context, Cai et al. [5] extended LDA to Semi-supervised Discriminant Analysis (SDA) by integrating the locality preserving criterion (depending on labeled and unlabeled samples) and adding a regularizer that controls the learning complexity. In the same way, Huang et al. [11] extended Local Discriminant Embedding (LDE) to Semi-supervised Discriminant Embedding (SDE). Recently, a new projection function, which can easily transform samples to their lower dimensional representation, got a significant attention because it does not need to solve the out-of-sample problem. In [17], the authors propose a label

* Corresponding author.
E-mail address: fadi.dornaika@ehu.es (F. Dornaika).

http://dx.doi.org/10.1016/j.neucom.2015.04.042
0925-2312/

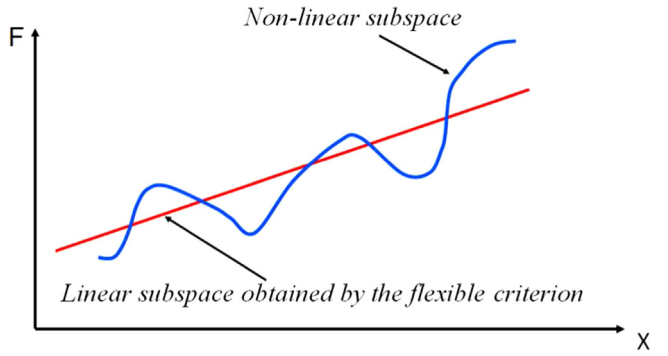

Fig. 1. An illustration of a linear subspace obtained by the Flexible Manifold Embedding (FME).

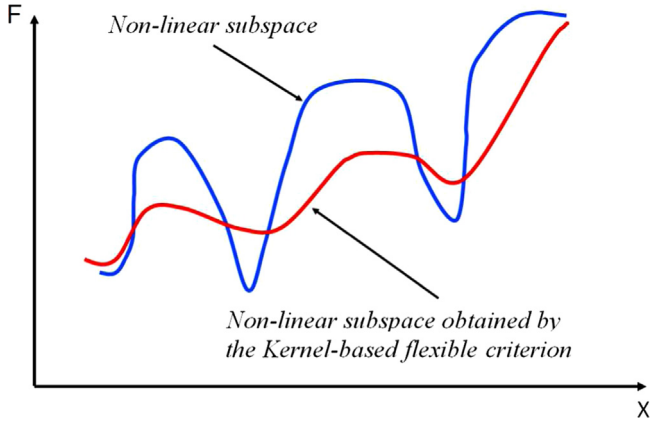

Fig. 2. An illustration of a nonlinear subspace obtained by the proposed Kernel Flexible Manifold Embedding (KFME).

propagation framework called Flexible Manifold Embedding (FME) which estimates simultaneously the optimal nonlinear prediction labels, the linear regression function and the regression residue. The regression function is estimated by adding a flexible term in order to force the linear embedding to be as close as possible to the nonlinear one. On the other hand, the first use of Kernel paradigm in machine learning was in 1990 [9]. Kernel PCA was introduced by Scholkopf in 1998 [21]. A kernelized version of LDA, Kernel Fisher Discriminant (KFD), was also proposed by Mika et al. [15] for two classes case and was extended by Baudat and Anouar [1] to multi-class case. Yang et al. [28] proposed a Complete Kernel Fisher Discriminant framework (CKFD) which has been used to improve discriminant analysis by adding a regular and an irregular subspaces.

In case of highly nonlinear structure of data, classical FME is not able to make a good discrimination. To deal with this problem, we propose, in this paper, a Kernel version of FME (KFME). In this latter, we extend the FME objective function to its kernelized version. The proposed method is also flexible because it looks for a nonlinear manifold that is the closest to the Kernel-based embedding.

Our proposed framework is characterized by the following features:

- Unlike a lot of other frameworks, ours can easily transform unseen data samples to the new subspace. These samples will be as close as possible to their Kernel-based ones.
- KFME is based on a Kernel projection, which aims to solve the data nonlinearity to some extent. However, the KFME formulation includes the Laplacian smoothness term that makes the label inference locally smooth. In other words, KFME attempts to preserve the local data distribution via the pairwise similarity matrix. We stress the fact that these two objectives are not contradictory since overcoming the data nonlinearity is performed globally while preserving data distribution works locally.
- The proposed KFME does not need a classifier since the classification is included in the embedding.

The remainder of the paper is organized as follows. In Section 2, we review the main semi-supervised methods including FME method. Section 3 introduces our proposed KFME method. In Section 4, we present experimental results and we make a comparison of our proposed framework with different semi-supervised methods applied on eight benchmark databases. Our conclusion is presented in Section 5.

2. Related work

In this section, we introduce some state-of-the-art semi-supervised learning methods, namely Semi-supervised Discriminant Analysis (SDA) [5], Semi-supervised Discriminant Embedding (SDE) [11], Laplacian Regularized Least Square (LapRLS) [4], and Flexible Manifold Embedding (FME) [17].

For this purpose, we will first present the notations that will be used later in this paper. We define the data matrix $\mathbf{X}=[\mathbf{x}_1,\mathbf{x}_2,\ldots,\mathbf{x}_l,\mathbf{x}_{l+1},\ldots,\mathbf{x}_{l+u}]\in\mathbb{R}^{D\times(l+u)}$ where D denotes the number of features and $N=l+u$ the total number of samples. $\mathbf{x}_i|_{i=1}^{l}$ 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. 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, and $y_i\in\{1,2,\ldots,C\}$ the label of the sample $\mathbf{x}_i$, where C is the total number of classes. Let n_c denote the total number of labeled samples that belong to the cth class, and let $\mathbf{S}\in\mathbb{R}^{N\times N}$ be the graph similarity matrix. The element $\mathbf{S}(i,j)$ corresponds to the score $sim(\mathbf{x}_i,\mathbf{x}_j)$ that represents the similarity between samples $\mathbf{x}_i$ and $\mathbf{x}_j$. There are several ways for computing the graph similarity matrix. The score $sim(\mathbf{x}_i,\mathbf{x}_j)$ can be simply the inverse of the Euclidean distance between $\mathbf{x}_i$ and $\mathbf{x}_j$, or it can be a more complex measure of distance such as the Gaussian [16]. Furthermore, the graph similarity matrix can be estimated using data self-representativeness techniques (e.g., [19,7]). The Laplacian matrix $\mathbf{L}\in\mathbb{R}^{N\times N}$ corresponding to the similarity matrix $\mathbf{S}$ is defined by $\mathbf{L}=\mathbf{D}-\mathbf{S}$ where $\mathbf{D}$ is a diagonal matrix whose elements are the row sums of $\mathbf{S}$ (or column sums since $\mathbf{S}$ is symmetric).

We will consider special graphs for labeled samples: let $\mathbf{S}_w$ and $\mathbf{S}_b\in\mathbb{R}^{l\times l}$ be the similarity matrices representing within and between class graphs, respectively. Thus $\mathbf{S}_w(i,j)=sim(\mathbf{x}_i,\mathbf{x}_j)$ if $\mathbf{x}_i$ and $\mathbf{x}_j$ share the same label, and $\mathbf{S}_w(i,j)=0$ otherwise. And $\mathbf{S}_b(i,j)=sim(\mathbf{x}_i,\mathbf{x}_j)$ if $\mathbf{x}_i$ and $\mathbf{x}_j$ have different labels, and $\mathbf{S}_b(i,j)=0$ otherwise. The Laplacian matrices of $\mathbf{S}_w$ and $\mathbf{S}_b$ will be defined in the same way as that of $\mathbf{S}$ by $\mathbf{L}_w$ and $\mathbf{L}_b$, respectively.

Let $\mathbf{Y}\in\mathbb{B}^{N\times C}$, where $\mathbb{B}=\{0,1\}$, be the binary label matrix associated with the samples so that $\mathbf{Y}(i,j)=1$ if $y_i=j$ and $\mathbf{Y}(i,j)=0$ otherwise. We also consider a matrix of labels $\mathbf{F}\in\mathbb{R}^{N\times C}$. In the sequel $\mathbf{0}$ and $\mathbf{1}\in\mathbb{R}^{N\times 1}$ represent two column vectors with all elements equal to 0 and 1, respectively.

2.1. Semi-supervised discriminant analysis

By keeping the foundation of LDA presented by the research of axes on which samples from the same class get similar representations and samples from different classes will be far from each other as much as possible, Cai et al. extended LDA to SDA [5] by adding a geometrically based regularizer. LDA can be seen as a particular case of graph-based