Neurocomputing 205 (2016) 165–181

Contents lists available at ScienceDirect

Neurocomputing

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

Supervised kernel nonnegative matrix factorization for face recognition

Wen-Sheng Chen [a,b], Yang Zhao [a,b], Binbin Pan [a,b,*], Bo Chen [a,b]

[a] College of Mathematics and Statistics, Shenzhen University, Shenzhen 518160, China
[b] Shenzhen Key Laboratory of Media Security, Shenzhen University, Shenzhen 518160, China

ARTICLE INFO

Article history:
Received 19 July 2015
Received in revised form
3 February 2016
Accepted 8 April 2016
Communicated by Ran He
Available online 10 May 2016

Keywords:
Face recognition
Kernel nonnegative matrix factorization
Supervised method

ABSTRACT

Nonnegative matrix factorization (NMF) is a promising algorithm for dimensionality reduction and local feature extraction. However, NMF is a linear and unsupervised method. The performance of NMF would be degraded when dealing with the complicated nonlinear distributed data, such as face images with variations of pose, illumination and facial expression. Also, the available labels could potentially improve the discriminant power of NMF. To overcome the aforementioned limitations of NMF, this paper proposes a novel supervised and nonlinear approach to enhance the classification power of NMF. By mapping the input data into a reproducing kernel Hilbert space (RKHS), we can discover the nonlinear relations between the data. This is known as the kernel methods. At the same time, we make use of discriminant analysis to force the within-class scatter small and between-class scatter large in the RKHS. It theoretically shows that the proposed approach can guarantee the non-negativity of the decomposed components and the objective function is non-increasing under the update rules. The proposed method is applied to face recognition. Compared with some state-of-the-art algorithms, experimental results demonstrate the superior performance of our method.

1. Introduction

Over the past decades, face recognition (FR) techniques have attracted much attention in the community of pattern classification and computer vision. The crucial stage of FR lies in the facial feature extraction. To achieve the goal, one should model a group of facial basis images as projection directions and use their coefficients (features) to represent original facial data. The popular linear methods for feature extraction are principal component analysis (PCA) [1], linear discriminant analysis (LDA) [2] and nonnegative matrix factorization (NMF) [3,4]. PCA aims to project the primitive samples into a low dimensional feature space such that the projected samples have largest variance. The PCA basis images, also called eigenfaces in FR, are the leading eigenvectors of the covariance matrix of the training samples. LDA focuses on extracting features for classification purpose. It maps the samples into a linear subspace by maximizing the ratio of the inter-class distance to the intra-class distance. In face recognition, LDA basis images are known as Fisherfaces. Different from PCA and LDA approaches, NMF does not allow subtraction operations because NMF is implemented under nonnegative constraints. The objective of NMF is to approximatively decompose the nonnegative data matrix into two matrices with nonnegative elements, namely basis matrix and coefficient feature matrix. As we know, the pixels of facial image are nonnegative as well, so NMF can be used to learn part-based representation of the facial data. It is interesting that the parts of face represented by the basis matrix are the nose, eyes, mouth, etc. Conversely, these parts can be non-negatively combined to reconstruct the whole facial image. In recent years, NMF, as an unsupervised feature extraction method, has been successfully applied to face recognition. A large number of NMF variants have been proposed [5–17]. For example, literature [16] incorporated the label information and local geometric structure into the factors of NMF. Zheng et al. learned the structured sparse basis images of NMF by using the pixel dispersion penalty [17].

Nevertheless, owing to the variations of illumination, facial expression, pose and so on, the distribution of facial images is very complicated and thus nonlinear in facial data space. This means that the performances of linear methods, such as PCA, LDA and NMF, will be degraded for nonlinear classification tasks. A commonly used method to tackle nonlinear problem is the kernel method, which has shown to be an effective technique for nonlinear feature extraction [18]. The basic idea of kernel method is to find a nonlinear mapping φ, from original pattern space to a reproducing kernel Hilbert space (RKHS) $\mathcal{F}$, such that the mapped samples are linearly separated, and then execute the linear methods in RKHS. However, the dimensionality of $\mathcal{F}$ is usually

* Corresponding author at: College of Mathematics and Statistics, Shenzhen University, Shenzhen 518160, China.
E-mail addresses: chenws@szu.edu.cn (W.-S. Chen), pbb@szu.edu.cn (B. Pan), chenbo@szu.edu.cn (B. Chen).

http://dx.doi.org/10.1016/j.neucom.2016.04.014
0925-2312/

very large, and may even be infinite. Also, how to obtain the nonlinear mapping is still a problem. Fortunately, these problems can be overcome using kernel trick. It is based on the fact that the kernel based algorithms just have relationship with the inner product $\langle\varphi(x),\varphi(y)\rangle_{\mathcal{F}}$ which can be replaced by a kernel function k, that is,

$$k(x,y)=\langle\varphi(x),\varphi(y)\rangle_{\mathcal{F}}, \tag{1}$$

where x and y are in the input space. This indicates that the nonlinear mapping is implicitly applied in RKHS. At present, a number of publications have extended the linear methods to the kernel counterparts, such as kernel PCA (KPCA) [19], kernel LDA (KLDA) [20] and kernel NMF (KNMF) [21]. These kernel based findings generally surpass their linear versions in face recognition.

The existing kernel based NMF approaches mainly attempt to discover the nonnegative pre-images in the input space and the nonnegative coefficient matrix such that the mapped samples could be expressed as linear combinations of the mapped pre-images under the non-negative constraints. Because the mapped samples are unknown in RKHS, it is infeasible to factorize the mapped samples matrix directly. To avoid this obstacle, Buciu et al. proposed a polynomial kernel NMF (PNMF) approach using the cost function with Frobenius norm [21], which can be easily expressed via a kernel function. To ensure that the factorization is nonnegative, PNMF only adopts the polynomial kernel, but fails to use other kernel functions. Another limitation of PNMF is that the limit point produced by the optimization algorithm cannot guarantee to be a stationary point. To remedy the above limitations of PNMF, the projected gradient kernel NMF (PGKNMF) was proposed to use arbitrary kernel functions [22]. Moreover, the limit point is ensured to be a stationary point of the optimization procedure. In order to find the pre-images of basis, both PNMF and PGKNMF needs to get back from the feature space to the input space. This is the curse of pre-image problem as raised in [23]. The authors in [23] derived the pre-images directly in the input space. It is noteworthy that [24,25] found the basis in the feature space, not the pre-images in the input space. By restricting the basis lying in the linear space spanned by the mapped training data, [24] obtains a new multiplicative update rules expressed with kernel matrix. By forcing the non-negativity of mapped training data in the RKHS, [25] performs NMF in the feature space directly. It is shown in [25] that the non-negativity of mapped training data could produce sparser basis, thus is more suitable for face recognition. Although these variants of KNMF [21–25] are able to model the nonlinear structure of the samples, one unappealing aspect of the current KNMF methods is that they do not utilize the class label information of the training data and thus are unsupervised learning methods for nonlinear feature extraction. We know that the supervised method usually has more discriminant power than the unsupervised counterpart. Therefore, it would be better for KNMF to take full advantage of the class label information.

In summary, there are two major issues in NMF including nonlinear problem and supervised learning problem. These problems will negatively affect the performance of NMF. In order to enhance the discriminant power of NMF method, we propose a novel supervised kernel NMF (SKNMF) approach in this paper by making use of the kernel theory and discriminant analysis method. Firstly, a nonlinear mapping is implicitly exploited to embed the samples into a RKHS, and then a nonlinear objective function is established to find nonnegative pre-images and nonnegative coefficient matrix such that the mapped samples are expressed as the non-negatively linear combinations of the mapped pre-images. To utilize the class label information, two quantities, namely within-class scatter and total-class scatter, are incorporated into the objective function as well. We minimize the objective function using gradient descent method and derive out SKNMF update formulae. It theoretically proves that our objective function is non-increasing under SKNMF update rules. This implies the convergence of our algorithm. The proposed SKNMF method is finally tested on face databases. The experimental results show that our SKNMF approach surpasses some state of the art approaches.

The rest of this paper is organized as follows: Section 2 will briefly introduce some related works, such as NMF and KNMF with polynomial kernel. Section 3 proposes our SKNMF and gives theoretic analysis. Experimental comparisons are reported in Section 4. Finally, Section 5 draws the conclusions.

2. Related works

This section describes NMF [3,4] and KNMF [21] algorithms briefly. Let m be the dimension of original feature space and $x_i \in \mathbb{R}_+^m$ $(i=1,2,\dots,n)$ be the training samples with nonnegative entries. The nonnegative training sample matrix X is denoted by $X=[x_1,x_2,\dots,x_n]\in\mathbb{R}_+^{m\times n}$. The ith row and jth column element of a matrix A is denoted as $[A]_{ij}$.

2.1. NMF

NMF aims to approximately factorize the training sample matrix X into two nonnegative matrices $W\in\mathbb{R}_+^{m\times r}$ and $H\in\mathbb{R}_+^{r\times n}$, i.e. $X\approx WH$. It indicates that each sample can be expressed as a linear combination of the columns of W, namely $x_j\approx Wh_j$, where h_j is the jth column of H. Matrices W and H are called basis matrix and coefficient matrix, respectively. In general, NMF needs to minimize the following objective function:

$$F_{NMF}(W,H)=\frac{1}{2}\|X-WH\|_F^2,$$

subject to the constraints $W\geq 0$[1] and $H\geq 0$. $\|\cdot\|_F$ is the Frobenius norm of a matrix.

The optimization problem can be solved using gradient descent method. Based on the objective function with Frobenius norm, NMF has the following update rules:

$$H^{(t+1)}=H^{(t)}\otimes(W^{(t)T}X)\oslash(W^{(t)T}W^{(t)}H^{(t)}), \tag{2}$$

$$W^{(t+1)}=W^{(t)}\otimes(XH^{(t)T})\oslash(W^{(t)}H^{(t)}H^{(t)T}), \tag{3}$$

$$W^{(t+1)}=W^{(t+1)}\oslash S, \tag{4}$$

where $\otimes$ and $\oslash$ denote element-wise multiplication and division, respectively, $[S]_{jr}=\sum_{i=1}^m[W]_{ir}$. The formula (4) constrains the columns of W to sum to 1. It has been shown in [3] that the objective function is non-increasing under the update rules (2)–(4).

2.2. Kernel NMF

Kernel NMF (KNMF) is also known as nonlinear NMF. The basic idea of KNMF is that it firstly maps nonnegative data into a RKHS $\mathcal{F}$ via a nonlinear mapping $\varphi:\mathbb{R}_+^m\longmapsto\mathcal{F}$, and then finds the nonnegative features and pre-images such that the mapped samples can be approximately expressed as a linear combination of the mapped pre-images in $\mathcal{F}$, namely

$$\varphi(x_j)\approx\sum_{i=1}^{r}h_{ij}\varphi(w_i),$$

where the feature h_{ij} and pre-image w_i are nonnegative. If we denote $W=[w_1,w_2,\dots,w_r]\in\mathbb{R}_+^{m\times r}$, $\varphi(X)=[\varphi(x_1),\ \varphi(x_2),\ \dots,\varphi(x_n)]$,

[1] $W\geq 0$ means that all elements of W are nonnegative.