Neurocomputing 175 (2016) 198–205

Contents lists available at ScienceDirect

Neurocomputing

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

A regularized least square based discriminative projections for feature extraction

Wankou Yang [a,b,c], Changyin Sun [a,b,*], Wenming Zheng [c]

[a] School of Automation, Southeast University, Nanjing 210096, China
[b] Key Lab of Measurement and Control of Complex Systems of Engineering, Ministry of Education, Southeast University, Nanjing 210096, China
[c] Key Laboratory of Child Development and Learning Science of Ministry of Education, Southeast University, Nanjing 210096, China

ARTICLE INFO

Article history:
Received 11 August 2015
Received in revised form
10 October 2015
Accepted 21 October 2015
Communicated by Su-Jing Wang
Available online 31 October 2015

Keywords:
Regularized least square
Sparse representation
Collaborative representation
Feature extraction

ABSTRACT

In this paper, we present a regularized least square based discriminative projections (RLSDP) method for feature extraction. First, we show that both sparse representation based classifier (SRC) and collaborative representation based classification (CRC) are regularized least square in nature. Second, a regularized least square based graph embedding framework (RLSGE) is constructed. Third, a RLSGE based feature extraction method is given, named regularized least square based discriminant projections (RLSDP). In RLSDP, the within-class compactness information is characterized by the reconstruction residual from the same class, which is consistent with the idea of reconstruction; the between-class separability information is characterized by the between-class scatter matrix like Fisher LDA. RLSDP is much faster than SPP since RLSDP adopts the L2 norm constraint while SPP adopts the L1 norm constraint. The experimental results on AR face database, FERET face database, and the PolyU FKP database demonstrate that RLSDP works well in feature extraction and has a great recognition performance.

1. Introduction

There are much high-dimensional data in numerous real-world applications and it is difficult to directly process the high dimensional data. Feature extraction (i.e. dimensionality reduction) [1,2] aims to project the high dimensional data into an intrinsic meaningful low-dimensional subspace. Feature extraction is an effective approach to extract the distinctive features from the original data in pattern classification [1]. Principal components analysis (PCA) [3] and Fisher linear discriminant analysis (FLDA) [4] are the two typical feature extraction methods.

Feature extraction methods could be divided into two categories: (1) linear methods, (2) nonlinear methods. Typical linear methods contain PCA, Fisher LDA and others. Fisher LDA often meets small sample size problem (SSS) [2] in which the within-class is singular. Many LDA variants are developed in the past decades [5–16]. Since linear methods could not process the nonlinear variants (e.g. illumination, gesture and so on). Then nonlinear methods are presented. The nonlinear methods include (1) kernel methods, (2) manifold learning based methods. Kernel methods have kernel PCA [17], kernel Fisher LDA [18] and so on. Typical manifold learning based methods include isometric feature mapping (ISOMAP) [19], local linear embedding (LLE) [20], and Laplacian eigenmap [21], Locality Preserving Projections (LPP) [22], Neighborhood Preserving Embedding (NPE) [23], Unsupervised Discriminant Projections (UDP) [24], marginal Fisher analysis (MFA) [25], spectral regression discriminant analysis (SRDA) [26,27], group sparse multiview patch alignment framework (GSM-PAF) [28], locality preserving discriminant projections (LPDP) [29], orthogonal discriminant local tangent space alignment (OD-LTSA) [30] and etc. [31–33]. PCA, LPP, UDP and KPCA et al. are the unsupervised methods. LDA, MFA and KLDA etc. are the supervised methods. Usually supervised methods and nonlinear methods could surpass unsupervised methods and linear methods, respectively.

Yan et al. gave a graph embedding framework (GE) [25] which could unify all these mentioned feature extraction methods. The key step in graph embedding framework is to construct the graph. The classical graph construction methods are k nearest neighbor method and ε-ball based method: (1) connect the samples in k nearest neighbor or ε-ball, (2) calculate the graph edge weight using Gaussian-kernel or L2-reconstrucition. It is difficult to manually set the nearest neighbor number, ball radius size and Gaussian-kernel parameter in high dimensional space. In short, both methods are not datum adaptive. Recently, sparse representation has been successfully applied in the construction of graph in the dimension reduction, with promising performance. For instance, sparse representation was used to generate a L1-

* Corresponding author.
E-mail address: cysun@seu.edu.cn (C. Sun).

http://dx.doi.org/10.1016/j.neucom.2015.10.049
0925-2312/

graph [39,40] and a discriminative dimension projection was developed based on sparse representation [39–41].

Ma et al. [37] gave a sparse representation based classification (SRC) which got promising results in face recognition. Sparse representation (SR) has attracted much attention [34–36]. In SRC, a testing sample is reconstructed by sparse linear combination of all the training samples and assigned the class label with minimum sparse reconstruction error. Representative vector machines (RVM) [38] was proposed as a unified framework for classical classifiers, such as the nearest neighbor (NN) classifier, support vector machine (SVM), and sparse representation based classification (SRC). The success of SRC has triggered many sparse representation based dimension reduction methods. Qiao et al. [39] developed a sparsity preserving projections (SPP) for dimensionality reduction based on SRC. In SPP, the graph (i.e. reconstruction relationships or affine matrix) was automatically constructed based on sparse representation. Chen et al. [40] presented a L1-graph construction method. Their ideas are similar. Yang et al. [41] gave the SRC steered discriminative projections (SRC-DP) for face recognition, which aimed to maximize the ratio of the between-class reconstruction residual to the within-class reconstruction residual in the reduced subspace. Gui et al. [42] developed a discriminant sparse neighborhood preserving embedding method using SR and maximum margin criterion. Wang [43] developed a structured sparse extension to linear graph embedding framework by introducing a structured sparsity-inducing norm. The L1-norm minimization is very slow, which makes the sparse representation based methods (e.g. SRC and SRC-DP) have a high computation complexity. So both SPP and L1-graph are slow.

Following SRC, Zhang et al. [44] studied the working mechanism of SRC and pointed out that the key part in SRC was collaborative representation and developed a collaborative representation based classification (CRC) for image recognition. CRC is much faster than SRC but has similar performance. The key in SRC and CRC is the collaborative representation. The difference between SRC and CRC is only the regularized term. Both sparse representation and collaborative representation belong to regularized least square methods in nature. Least square regression and regularized least square regression are efficient statistical methods and have been successfully applied in pattern recognition and machine learning [45–49]. Saunders et al. [45] studied a dual version of the ridge regression procedure and gave a way to perform nonlinear regression by constructing a linear regression function in a high dimensional feature space. An et al. [46] presented a ridge regression (RR) and kernel ridge regression (KRR) techniques for multivariable labels and applied the methods to face recognition. Cai et al. [47] gave a spectral regression method for dimension reduction. Fernando [48] presented a least squares framework to formulate component analysis methods. Xiang et al. [49] proposed a discriminative least square model for multiclass classification and feature selection. More information about ridge regression could be find in [50].

Inspired by SPP, least square regression theory and supervised feature extraction methods, we present a regularized least square based graph embedding (RLSGE) framework for feature extraction. In RLSGE, the regularized least square based reconstruction coefficients are used to characterize the relationships between samples (i.e. In RLSGE, the graph is constructed based on regularized least square). Under RLSGE framework, we give a regularized least square based discriminative projection (RLSDP) method for feature extraction. In RLSDP, two quantities for classification are considered: the within-class compact information and between-class separability information in the modeling process. (1) According to the classification idea of SRC and CRC, the reconstruction residual of the test sample on the same class should be minimized. In RLSDP subspace, we also aim to minimize the reconstruction residual of the test sample on the same class (i.e. minimize the within-class compact information). (2) For classification, we want to maximize the separability between classes in RLSDP subspace (i.e. maximize the between-class separability information). Based on the within-class compact information and the between-class separability information, we give a discriminative criterion which maximizes the between-class separability and minimizes the within-class compact simultaneously. The proposed criterion, similar to the classical Fisher criterion, is a Rayleigh quotient form and can be calculated via generalized eigenvalue decomposition.

The rest of the paper is organized as follows: In Section 2, we briefly review SRC, CRC and SPP. In Section 3, we describe the RLSGE framework and RLSDP in detail. In Section 4, experiments are conducted to evaluate the proposed method. Conclusions are given in Section 5.

2. Related works

2.1. SRC and CRC

Suppose that $X_i = [x_{i,1}, x_{i,2}, \cdots, x_{i,n_i}] \in R^{m \times n_i}$ is the set of training samples of the i^{th} class, where $x_{i,j} (j = 1,2,\ldots,n_i)$ is the j^{th} training samples of the i^{th} class, $X = [X_1, X_2, \cdots, X_c]$ is the total training samples set and $n = n_1 + n_2 + \ldots + n_c$. A testing sample $y \in R^m$ from the i^{th} class should be well approximated by linear combination of the sample of the same class X_i, i.e., $y = \sum_{j=1}^{n_i} x_{i,j} \cdot w_{i,j} = X_i w_i$, $w_i = \left[w_{i,1}, w_{i,2}, \cdots, w_{i,n_i}\right]^T$ is the coefficient. The linear representation of y can be rewritten based on all training samples as $y = Xw$, where $w = [w_1; \cdots; w_i; \cdots; w_c] = [0, \cdots, 0, w_{i,1}, w_{i,2}, \cdots, w_{i,n_i}, 0, \cdots, 0]^T$.

The SRC [37] algorithm is as follows:

Step1. Normalize. Normalize the columns of X using L2-norm;

Step2. Sparse Code. Solve the L1-minimization problem,

$$\widehat{w} = \arg\min_{w} \{ \|y - Xw\|_2^2 + \lambda \|w\|_1 \} \tag{1}$$

where λ is a positive scalar number that balances the reconstructed error and coefficients' sparsity.

Step3. Calculate the residuals:

$$r_i(y) = \|y - X\delta_i(\widehat{w})\|_2, \text{for } i = 1, \ldots, c. \tag{2}$$

where $\delta_i(\bullet) : R^n \to R^n$ is the characteristic function which selects the coefficients associated with the i^{th} class.

Step4. Classify. Output the label as $\text{label}(y) = \arg\min_i r_i(y)$.

The SRC scheme mainly consists of two steps: a sparse code step and the classification step.

In CRC [44], the regularized term is replaced with L2 norm. The objective function of the CRC is Eq. (3):

$$\widehat{w} = \arg\min \{ \|y_0 - Xw\|_2^2 + \lambda \|w\|_2^2 \} \tag{3}$$

where $\lambda > 0$ is the regularized parameter. The regularization has two roles: (1) it makes the least square solution stable, (2) it also introduces a certain amount of sparsity to the solution w, yet this sparsity is much weaker than that by L1-norm. The solution of the CRC is Eq. (4)

$$\widehat{w} = \left(X^T X + \lambda I\right)^{-1} X^T y_0 \tag{4}$$

2.2. Least square regression

Let we have a training set (x_i, y_i) $(i = 1, 2, \cdots, n,\ x_i \in R^m, y_i \in R)$ and $y_i = w \cdot x_i$ $(w \in R^m)$. Least squares method calculate w by

$$\min_{w} \left\{ \frac{1}{2} \sum_{i=1}^{n} (y_i - w \cdot x_i)^2 \right\} = \min_{w} \left\{ \frac{1}{2} \|y - Xw\|^2 \right\} \tag{5}$$