J. Vis. Commun. Image R. 25 (2014) 1800–1809

Contents lists available at ScienceDirect

J. Vis. Commun. Image R.

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

Short Communication

Extended linear regression for undersampled face recognition

Si-Bao Chen [a,*], Chris H.Q. Ding [b], Bin Luo [a]

[a] Key Lab of Intelligent Computing and Signal Processing of Ministry of Education, School of Computer Science and Technology, Anhui University, Hefei, Anhui, China
[b] Department of Computer Science and Engineering, University of Texas at Arlington, Engineering Research Building, Room 529, 500 UTA Blvd, Arlington, TX 76019, USA

A R T I C L E I N F O

Article history:
Received 12 September 2013
Accepted 27 July 2014
Available online 9 August 2014

Keywords:
Linear regression
Intraclass variant dictionary
Singular value decomposition
Face recognition
Low rank
Undersampled classification
Supervised learning
Pattern recognition

A B S T R A C T

Linear Regression Classification (LRC) is a newly-appeared pattern recognition method, which formulates the recognition problem in terms of class-specific linear regression with sufficient training samples per class. In this paper, we extend LRC via intraclass variant dictionary and SVD to undersampled face recognition where there are very few, or even only one, training sample per class. Intraclass variant dictionary is adopted in undersampled situation to represent the possible variation between the training and testing samples. Three types of methods, quasi-inverse, ridge regularization and Singular Value Decomposition (SVD), are designed to solve low-rank problem of data matrix. Then the whole algorithm, named Extended LRC (ELRC), is presented for face recognition via intraclass variant dictionary and SVD. The experimental results on three well-known face databases show that the proposed ELRC has better generalization ability and is more robust to classification than many state-of-the-art methods in undersampled situation.

1. Introduction

In classical face recognition research, due to high-dimension of images, *curse of dimensionality* becomes a critical issue, which triggers the invention of numerous methods for dimensionality reduction, such as linear Principal Component Analysis (PCA) [1,2], Linear Discriminant Analysis (LDA) [3,4], Locality Preserving Projections (LPP) [5,6], and nonlinear ISOMAP [7], Locally Linear Embedding (LLE) [8] and Laplacian Eigenmaps [9]. De la Torre [10] presented a least-squares framework for component analysis. Apart from the classical methods, it has been shown recently that simple downsampled images and random projections can also serve well with Sparse Representation-based Classification (SRC) [11]. The choice of feature space may no longer be so critical. What really matters is the dimension of features and the design of classifier [11].

Based on the research of Wright et al. [11], Wagner et al. [12] extended SRC to handle uncontrolled illuminations, pose variations, and face alignment simultaneously by a deformable sparse recovery and classification algorithm. Naseem et al. [13] formulated the face recognition problems in terms of class-specific linear regression and proposed Linear Regression Classification (LRC) based on the assumption that samples from the same class lie on a linear subspace. LRC falls in the category of nearest subspace classification. A similar work in [14], Chai et al. introduced Locally Linear Regression (LLR) between a nonfrontal face image and its frontal counterpart to tackle the pose problem. The estimation of linear mapping is further formulated as a prediction problem with a regression based solution.

All these methods cast classification problems in a linear representation (reconstruction) of test sample by training samples. The selection of dictionary (training samples) is very critical to the classification performance. There are many dictionary learning methods appeared in the literature. Aharon et al. [15] proposed K-SVD algorithm, a generalized K-means clustering process, to iteratively alternate between sparse coding and updating dictionary atoms for sparse representation. Zhang and Li [16] incorporated classification error into the objective function of K-SVD to learn discriminative dictionary. Jiang et al. [17] introduced label consistent constraint, called discriminative sparse-code error, in learning discriminative dictionary for sparse coding. Yang et al. [18] incorporated Fisher discrimination criterion into dictionary learning for sparse representation. Ma et al. [19] learned sparse representation dictionary based on matrix rank minimization and discriminative learning. All these methods need sufficient training samples to learn a better dictionary.

However, in many applications, only a few, or even single, sample per class can be obtained for training, which is known as the undersampled situation. To overcome the difficulty of undersampled situation, some methods have been reported in

* Corresponding author. Fax: +86 551 65108445.
E-mail addresses: sbchen@ahu.edu.cn (S.-B. Chen), CHQDing@uta.edu (C.H.Q. Ding), luobin@ahu.edu.cn (B. Luo).

http://dx.doi.org/10.1016/j.jvcir.2014.07.007
1047-3203/

the literature. Deng et al. [20] proposed extended SRC via intraclass variant dictionary to undersampled face recognition, assuming that the intraclass variations of one subject can be approximated by a sparse linear combination of those of other subjects. Naseem et al. [21] solved the well-conditioned inverse problem in LRC using the robust Huber M-estimation in the presence of noise.

In this paper, we propose an Extended Linear Regression-based Classification (ELRC) for recognition in undersampled situation. We find that although training samples are insufficient in undersampled situation, there still may exist linear correlation between training samples, and the low-rank situation will lead to the failure of LRC. We propose three methods including quasi-inverse, ridge regularization and SVD for low-rank situation. Intraclass variant dictionary is introduced for undersampled situation. Then the whole ELRC algorithm is given via SVD and intraclass variant dictionary, which is effective even when there is only one training sample per class.

The remainder of the paper is organized as follows. A brief review of LRC is presented in Section 2. Section 3 extends LRC to low-rank and undersampled situation. Section 4 is devoted to experimental results and analysis on three well-known face databases, and Section 5 concludes this paper.

2. Brief review of LRC

Let there be c classes with n_i training samples from the ith class, $i = 1, 2, \ldots, c$. Suppose the training samples of each class are sampled evenly and can represent any sample of the class. Each training sample is represented as $\mathbf{x}_j^{(i)} \in \mathcal{R}^p$, $j = 1, 2, \ldots, n_i$, $i = 1, 2, \ldots, c$, and the data matrix $\mathbf{X}_i$ of the ith class is formed by stacking all the training sample vectors $\mathbf{x}_j^{(i)}$ from the same class, $\mathbf{X}_i = \left[\mathbf{x}_1^{(i)}, \mathbf{x}_2^{(i)}, \ldots, \mathbf{x}_{n_i}^{(i)}\right] \in \mathcal{R}^{p \times n_i}$, $i = 1, 2, \ldots, c$. Let $\mathbf{y} \in \mathcal{R}^p$ be a test sample vector which needs to be classified into one of the classes $i = 1, 2, \ldots, c$. Following the concept that samples from the same class lie in a linear subspace [22], LRC [13] seeks a representation of the test sample $\mathbf{y}$ by linearly combining the training samples from the same class,

$$\mathbf{y} = \mathbf{X}_i \mathbf{w}_i + \mathbf{z}_i, \quad i = 1, 2, \ldots, c, \tag{1}$$

where $\mathbf{w}_i \in \mathcal{R}^{n_i}$ is the weighting parameter, $\mathbf{z}_i \in \mathcal{R}^p$ is the residual term. If $\mathbf{y}$ belongs to the ith class, then $\mathbf{z}_i$ is noise term with bounded energy $\|\mathbf{z}_i\|_2 \leqslant \varepsilon$. When $p > n_i$, parameter $\mathbf{w}_i$ can be estimated by least-square estimation,

$$\hat{\mathbf{w}}_i = (\mathbf{X}_i^\top \mathbf{X}_i)^{-1} \mathbf{X}_i^\top \mathbf{y}. \tag{2}$$

Then, the reconstruction (approximation) of $\mathbf{y}$ by the samples of the ith class is

$$\hat{\mathbf{y}}_i = \mathbf{X}_i \hat{\mathbf{w}}_i = \mathbf{X}_i (\mathbf{X}_i^\top \mathbf{X}_i)^{-1} \mathbf{X}_i^\top \mathbf{y}, \tag{3}$$

which is actually the projection of $\mathbf{y}$ into the subspace spanned by the column vector of $\mathbf{X}_i$. The residual between original test sample $\mathbf{y}$ and the reconstruction $\hat{\mathbf{y}}_i$ by the samples of the ith class is calculated as

$$r_i(\mathbf{y}) = \|\mathbf{y} - \hat{\mathbf{y}}_i\|_2, \quad i = 1, 2, \ldots, c, \tag{4}$$

and LRC classifies test sample $\mathbf{y}$ into the class which has the minimum residual,

$$\mathbf{y} \mapsto i^* = \arg \min_{i \in \{1,2,\ldots,c\}} r_i(\mathbf{y}). \tag{5}$$

However, in the implementation, the matrix inverse operator in (2) often fails even when feature dimension p is greater than the number of samples n_i, $p > n_i$. As long as the data matrix $\mathbf{X}_i$ is not full column rank, i.e., $rank(\mathbf{X}_i) < n_i$, matrix $\mathbf{X}_i^\top \mathbf{X}_i$ must be singular, and the matrix inverse operator in (2) must fail. This situation is often encountered when there is linear correlation among data samples (column vectors of data matrix $\mathbf{X}_i$). Furthermore, zero-centralization, which is often adopted to preprocess training samples, also brings the linear correlation among data samples, $\sum_{j=1}^{n_i} \mathbf{x}_j^{(i)} = 0$, and results in the failure of matrix inverse operator in (2).

On the other hand, LRC supposes that there are sufficient training samples for each class and the training samples of each class are sampled evenly. However, in many applications, there are very few (insufficient) data samples for training, even there is only one single sample (such as a passport photograph) left for each class. In such undersampled situation, there are very limited training samples of the ith class while the test sample may contain complex variations (such as expressions, illuminations and disguises in face recognition), which will largely deviate from the linear subspace spanned by the training samples of the ith class. Then the residual term $\mathbf{z}_i$ in (1) becomes large for the correct class, even larger than that of other classes, which will result in the false classification of LRC.

Extremely, when there is only one single sample per class $\mathbf{X}_i = \mathbf{x}_i$ available for training, the parameter estimation $\hat{w}_i$ in formula (2) is reduced to a kind of scaled similarity (coefficient),

$$\hat{w}_i = \frac{\mathbf{x}_i^\top \mathbf{y}}{\mathbf{x}_i^\top \mathbf{x}_i} = \sqrt{\frac{\mathbf{y}^\top \mathbf{y}}{\mathbf{x}_i^\top \mathbf{x}_i}} \frac{\mathbf{x}_i^\top \mathbf{y}}{\sqrt{\mathbf{x}_i^\top \mathbf{x}_i}\sqrt{\mathbf{y}^\top \mathbf{y}}} = \sqrt{\frac{\mathbf{y}^\top \mathbf{y}}{\mathbf{x}_i^\top \mathbf{x}_i}} \cos(\mathbf{x}_i, \mathbf{y}), \tag{6}$$

where $\cos(\mathbf{x}_i, \mathbf{y})$ is the cosine of the angle between vectors $\mathbf{x}_i$ and $\mathbf{y}$. Then the residual in formula (4) is turned into a kind of weighted Euclidean distance:

$$r_i(\mathbf{y}) = \|\mathbf{y} - \hat{\mathbf{y}}_i\|_2 = \|\mathbf{y} - \mathbf{x}_i \hat{\mathbf{w}}_i\|_2 = \left\| \mathbf{y} - \mathbf{x}_i \sqrt{\frac{\mathbf{y}^\top \mathbf{y}}{\mathbf{x}_i^\top \mathbf{x}_i}} \cos(\mathbf{x}_i, \mathbf{y}) \right\|_2. \tag{7}$$

Therefore, in one sample training situation, LRC is reduced to nearest neighbor classifier with a kind of weighted Euclidean distance for distance metrics.

3. Extend LRC to low rank and undersampled situations

In this section, we will extend LRC to undersampled situation via intraclass variant dictionary, and to low rank situation via Singular Value Decomposition (SVD). Then the whole extended LRC algorithm is given.

3.1. Intraclass variant dictionary for undersampled situation

In undersampled situation, training samples are insufficient and test sample will largely deviate from the linear subspace spanned by the training samples of the same class. The residual term $\mathbf{z}_i$ in (1) will become large even if the test sample $\mathbf{y}$ belongs to the ith class, which will result in the false classification of decision rule (5).

Based on the observation in [20], the large deviation from the test sample to the undersampled training samples of the same class can be linearly approximated by the intraclass differences of generic classes, which are arbitrary classes different from that (classes and subjects) being recognized. Let $\mathbf{D}_I$ be a basis matrix representing the universal intraclass variant dictionary (such as unbalanced lighting, exaggerated expressions, occlusions and disguises in face recognition). When the training samples are insufficient, the representation model (1) can be modified to account for large variation between the test sample and the undersampled training samples by taking intraclass variant dictionary $\mathbf{D}_I$ into consideration:

$$\mathbf{y} = \mathbf{X}_i \mathbf{w}_i + \mathbf{D}_I \mathbf{v}_i + \mathbf{z}_i, \quad i = 1, 2, \ldots, c, \tag{8}$$