J. Vis. Commun. Image R. 25 (2014) 1878–1885

Contents lists available at ScienceDirect

J. Vis. Commun. Image R.

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

Locality-sensitive kernel sparse representation classification for face recognition

Shiqing Zhang, Xiaoming Zhao *

Institute of Image Processing and Pattern Recognition, Taizhou University, Taizhou 318000, PR China

ARTICLE INFO

Article history:
Received 29 March 2014
Accepted 16 September 2014
Available online 28 September 2014

Keywords:
Sparse representation
Kernal
Sparsity
Data locality
Face recognition
Discriminating
l_1-norm
Classification

ABSTRACT

In this paper a new classification method called locality-sensitive kernel sparse representation classification (LS-KSRC) is proposed for face recognition. LS-KSRC integrates both sparsity and data locality in the kernel feature space rather than in the original feature space. LS-KSRC can learn more discriminating sparse representation coefficients for face recognition. The closed form solution of the l_1-norm minimization problem for LS-KSRC is also presented. LS-KSRC is compared with kernel sparse representation classification (KSRC), sparse representation classification (SRC), locality-constrained linear coding (LLC), support vector machines (SVM), the nearest neighbor (NN), and the nearest subspace (NS). Experimental results on three benchmarking face databases, i.e., the ORL database, the Extended Yale B database, and the CMU PIE database, demonstrate the promising performance of the proposed method for face recognition, outperforming the other used methods.

1. Introduction

Face recognition is one of the most widely investigated problems in pattern recognition and computer version. Recently, sparse representation (also called sparse coding) has become a currently very active research topic in signal processing, pattern recognition and computer version due to its plausive mathematical statistical theory [1]. Sparse representation is initially developed as an extension to traditional signal representations such as Fourier and wavelet representations. So far, sparse representation has been successfully applied to solve many practical problems in signal processing, computer vision, pattern recognition, etc. For instance, in signal and image processing fields, sparse representation is used for signal recovery and acquisition [2], image representation [3], and image sequence denoising [4]. In the emerging field of compressive sensing (CS) [5,6], as a very attractive theory challenging Shannon-Nyquist sampling theorem, sparse representation aims to recover the signal from the compressed measures in a most economical way. Especially, sparse representation classification (SRC) has been successfully applied for face recognition and obtained very promising performance on face recognition [7,8]. The main idea of SRC is that a test face image is represented as a sparse linear combination over the dictionary formed with all the training face images, and the class label of the test sample is then identified by evaluating which class yields the minimum the residual between itself and the reconstruction constructed by training samples of this class. A recent survey of sparse representation can be found in [9].

The discriminating ability of SRC depends on the quality of the dictionary. Ideally, from the point of view of the l_1-norm minimization algorithms the atoms of the dictionary corresponding to different classes should be separated from each other. In other words, the effectiveness of SRC is limited by an important assumption that data points from different classes are not distributed along the same radius direction [10,11]. Note that this assumption is reasonable for face recognition since the images from the same subject under different intensity levels are still considered to belong to the same class. Nevertheless, the assumption is not always effective when dealing with some real-world data, e.g., the Iris dataset from the UCI machine learning archive, in which classes are stratified along the radius direction. Therefore, SRC could not identify a test sample if it has the same vector direction as the training samples belonging to two or more classes. In other words, SRC suffers from the drawback of losing its discriminating ability when classifying data with the same direction distribution.

To overcome the drawback of SRC, in recent years some preliminary efforts have been devoted to develop kernel sparse representation classification (KSRC) [10–14]. Essentially, KSRC [10–14] uses the kernel trick [15] to find sparse representation coefficients

* Corresponding author.
E-mail address: tzxyzxm@163.com (X. Zhao).

http://dx.doi.org/10.1016/j.jvcir.2014.09.011
1047-3203/

in a high-dimensional feature space rather than in the original feature space. It's worth pointing out that the presented KSRC in [10] is a little different from [12–14] since it integrates the described KSRC [12–14] with dimensionality reduction in the kernel feature space. In machine learning area, the kernel trick is originally used to develop support vector machines (SVM) [16], kernel Fisher discriminant analysis (KFDA) [17], and kernel principal component analysis (KPCA) [18], since it can be easily used to generalize a linear algorithm into a nonlinear algorithm. These kernel-based learning methods recall the kernel trick to map the original data into a high-dimensional feature space (also called kernel feature space) by using a nonlinear mapping, and then perform linear processing problems in this high-dimensional feature space with the inner products. The inner products in the high-dimensional feature space are computed by using some implicit mapping kernel function. Similarly, KSRC maps the data into a high-dimensional feature space by using some nonlinear mapping associated with a kernel function and then implements the SRC algorithm in this high-dimensional feature space. As a kernel method, KSRC is capable of capturing the nonlinear relationships of data, and thus performs better than SRC for classification.

Although KSRC has been successfully applied for image classification and face recognition since it integrates the kernel method and the SRC method, KSRC is not able to capture the locality structure of data. In pattern recognition area, data locality is an important issue in the problems of K-nearest-neighbor (KNN) classifier [19], data clustering [20], dimensionality reduction [21,22], image classification [23], etc. Note that, the recently reported work [22,23] have proved that in the problem of sparse coding data locality is more essential than sparsity since integrating data locality with the original sparse coding methods yields more effective sparse coding coefficients. Motivated by the advantage of data locality, in this paper we integrate KSRC with data locality in the kernel feature space rather than in the original feature space, and develop an extension of KSRC, called locality-sensitive kernel sparse representation classification (LS-KSRC).

The remaining of this paper is organized as follows: Section 2 reviews kernel sparse representation classification (KSRC) in brief. Section 3 provides the proposed LS-KSRC method in detail. Experimental results and analysis are given in Section 4. Section 5 offers the concluding remarks and discussions.

2. Review of kernel sparse representation classification

2.1. Sparse representation classification

In this section, we will simply introduce the principal of SRC [7] based on the CS theory. Essentially, SRC is based on the linearity assumption that the whole set of training samples form a dictionary, and then the recognition problem is cast as one of discriminatively finding a sparse representation of the test image as a linear combination of training images by solving the l_1-norm optimization problem.

Given a set of training samples with dimensionality d $(x_i, y_i)|x_i \in R^d, y_i \in \{1, 2, \cdots, c\}, i = 1, 2, \cdots, N$, where y_i is the class label of input data x_i, c is the number of classes, the goal of SRC is to use the given c-class training samples to exactly predict the class label y_i of x_i.

Now let the jth class training samples form columns of a matrix $\mathbf{X}_j = [\mathbf{x}_{j,1}, \mathbf{x}_{j,2}, \cdots, \mathbf{x}_{j,n_j}] \in \mathbf{R}^{d\times n_j}, j = 1, 2, \cdots, c$, where $\mathbf{x}_{j,i}$ is the ith training sample of the jth class, and n_j denotes the number of the jth class training samples. Then, for all training samples a new sample matrix $\mathbf{X}$ can be expressed as

$$\mathbf{X} = [\mathbf{X}_1, \mathbf{X}_2, \cdots, \mathbf{X}_c] \tag{1}$$

In SRC, a test sample $\mathbf{x}$ could be represented linearly by all the training samples

$$\mathbf{x} = \mathbf{X}\boldsymbol{\alpha} + \varepsilon \tag{2}$$

where $\boldsymbol{\alpha}$ is the coefficient vector and ε is the approximation error. The linearity assumption in SRC implies that the coefficient vector $\boldsymbol{\alpha}$ is expected to be zero except some of those associated with the correct class label of the test sample. To achieve the coefficient vector $\boldsymbol{\alpha}$, the following l_1-norm minimization problem should be solved.

$$\min_{\alpha} \|\boldsymbol{\alpha}\|_1, \quad \text{subject to} \quad \|\mathbf{x} - \mathbf{X}\boldsymbol{\alpha}\|_2^2 \leqslant \varepsilon \tag{3}$$

This is a convex optimization problem and can be solved by quadratic programming. So far, a variety of efficient algorithms have been proposed to solve the l_1-norm minimization of Eq. (3), such as l1-magic [24], l1-ls [25], spectral projected gradient method (SPGL1) [26] and NESTA (a shorthand for Nesterov's algorithm) [27].

Once the coefficient vector $\boldsymbol{\alpha}$ is found, the test sample $\mathbf{x}$ could be classified in terms of the reconstruction errors (residuals) between $\mathbf{x}$ and its approximations. The jth approximation of the test sample $\mathbf{x}$ is achieved by using only the coefficients belonging to the jth class. The class label of the test sample $\mathbf{x}$ is then assigned to the one with the minimum residual. The detailed classification procedure of SRC [7] is summarized in Algorithm 1:

Algorithm 1. Sparse representation classification (SRC)

(1) **Input**: the matrix of all training samples $\mathbf{X}$, and a test sample $\mathbf{x}$
(2) Solve the l_1-norm minimization problem in Eq. (3)
(3) Compute the residuals by using the samples associated with the jth class by using $r_j(\mathbf{x}) = \|\mathbf{x} - \mathbf{X}\alpha_j\|_2^2$
(4) **Output**: the class label y of the given test sample $\mathbf{x}$: $y = \arg\min_{j=1,2,\cdots,c} r_j(\mathbf{x})$

2.2. Kernel sparse representation classification

By means of integrating the kernel trick and the SRC method, kernel sparse representation classification (KSRC) [10–14] is developed as a nonlinear extension of SRC. Essentially, KSRC aims to seek sparse representation coefficients in the kernel feature space rather than in the original feature space.

Assuming that there exists a nonlinear kernel mapping ϕ for each input data point $\mathbf{x}$:

$$\phi : R^d \to \mathcal{F}, \quad \mathbf{x} \mapsto \phi(\mathbf{x}) \tag{4}$$

With the nonlinear mapping ϕ, we can map the input data point $x_i \in R^d$ into some potentially high-dimensional feature space $\mathcal{F}$. In this kernel feature space $\mathcal{F}$ an inner product $\langle,\rangle$ could be defined for a properly chosen ϕ, which gives rise to a so-called reproducing kernel Hilbert space (RKHS). In RHKS, a kernel function $k(x_i, x_j)$ is defined as:

$$k(\mathbf{x}_i, \mathbf{x}_j) = \langle \phi(\mathbf{x}_i), \phi(\mathbf{x}_j) \rangle = \phi(\mathbf{x}_i)^T \phi(\mathbf{x}_j) \tag{5}$$

where k is known as a kernel. There are three commonly used kernel functions, i.e., the linear kernel, polynomial kernel as well as the Gaussian kernel.

The linear kernel function is defined as

$$k(\mathbf{x}_i, \mathbf{x}_j) = \mathbf{x}_i^T \mathbf{x}_j \tag{6}$$

The polynomial kernel function is defined as

$$k(\mathbf{x}_i, \mathbf{x}_j) = (\mathbf{x}_i^T \mathbf{x}_j)^r \tag{7}$$