Signal Processing 152 (2018) 255–264

Contents lists available at ScienceDirect

Signal Processing

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

Learning a discriminative dictionary for classification with outliers[☆]

Jiaming Qi, Wei Chen[*]

State Key Laboratory of Rail Traffic Control and Safety, Beijing Jiaotong University, China

ARTICLE INFO

Article history:
Received 31 January 2018
Revised 25 May 2018
Accepted 4 June 2018

Keywords:
Dictionary learning
Classification
Sparse representation

ABSTRACT

Dictionary learning aims to find a dictionary where signals in some ensemble have sparse representations, and has been successfully applied for classification. However, traditional dictionary learning methods for classification assume there is no outlier in the training data, which may not be the case in practical applications. In this paper, we propose a new discriminative dictionary learning framework for classification, which simultaneously learns a discriminative dictionary and detects outliers in the data. The dictionary learning framework is formulated into an optimization problem with designed regularizers to promote both the discrimination and outlier-detection capability. We demonstrate the superior performance of the proposed approach in comparison with state-of-the-art alternatives by conducting extensive experiments on various image classification tasks.

1. Introduction

The theory and algorithms of sparse signal representation have made a rapid development in the past years. They have been playing a central role in signal compression and proved to be very useful in many applications in the filed of signal processing, such as signal acquisition, denoising, and classification [2–8].

In the perspective of the sparse signal representation, a signal can be seen as a linear combination of a few atoms selected from a complete or over-complete dictionary, where each atom represents a column of the dictionary. Traditionally, signal dictionaries have relied heavily on explicit analytic bases constructed in nonlinear approximation and harmonic analysis, ranging from the classic Fourier to modern multidimensional, multidirectional, multiresolution bases (e.g., wavelets, curvelets, ridgelets) [9–11]. However, these dictionaries are often restricted to certain classes of signals, and it is uncertain if they are suitable to describe signals in new domains.

To overcome the shortcomings of the analytic dictionaries, Olshausen and Field exploit the principle of encoding signals with few atoms from a learned dictionary [12]. The learned dictionary is good at capturing the characteristics of signals on hand, and this finding has attracted significant attentions in the past two decades, leading to fruitful results involving both algorithmic developments and successful applications [6,13,14]. In addition, some theoretical study of dictionary learning had been conducted. In [15], Aharon et al. provided an algorithmic procedure to correctly find the underlying dictionary, where the algorithm requires exponential number of training signals and has exponential running time [16]. Recent work [17] provided a polynomial-time algorithm that provably recovers most over-complete dictionaries under certain conditions.

1.1. Dictionary learning for sparse representation

The goal of dictionary learning is to learn a basis from a collection of signals so that they can be sparsely represented. The typical dictionary learning problem can be described as

$$\min_{\mathbf{D},\mathbf{A}} \|\mathbf{X}-\mathbf{D}\mathbf{A}\|_F^2 + \lambda \sum_{i=1}^{N} \|\boldsymbol{a}_i\|_0, \tag{1}$$

where $\mathbf{X} \in \mathbb{R}^{m\times N}$ denotes a collection of N signals of dimension m, $\mathbf{D} \in \mathbb{R}^{m\times K}$ denotes a complete ($m = K$) or over-complete ($m < K$) dictionary that contains K atoms, and $\mathbf{A} = [\boldsymbol{a}_1, \boldsymbol{a}_2, \ldots, \boldsymbol{a}_N] \in \mathbb{R}^{K\times N}$ denotes a collection of sparse vectors corresponding to the signals in $\mathbf{X}$. $\|\cdot\|_0$ denotes the ℓ_0 norm, i.e., the number of non-zero elements in a vector. The first term and the second term of (1) denote the approximation error and the sparsity of the signal representation, respectively, and λ is a parameter that controls the balance between approximation accuracy and sparsity. However, this dictionary learning problem with the ℓ_0 norm is NP hard. Alternatively, one can either replace the non-convex ℓ_0 norm regularizer with the convex ℓ_1 norm regularizer to encourage sparsity [18],

☆ This work is supported by the Natural Science Foundation of China (61671046, 61725101); the Beijing Natural Science Foundation (4182050); the State Key Laboratory of Rail Traffic Control and Safety (RCS2018ZZ002) of Beijing Jiaotong University. Parts of this work have been presented at 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) [1].

* Corresponding author.
E-mail address: weich@bjtu.edu.cn (W. Chen).

https://doi.org/10.1016/j.sigpro.2018.06.005
0165-1684/

or use greedy algorithms, e.g., orthogonal matching pursuit [19] to solve (1). Classic dictionary learning methods for sparse representation include K-SVD [20] and the method of optimal direction (MOD) [21].

In recent years, various methods have been proposed, either to improve the performance of dictionary learning for sparse representation or to consider constraints in practical applications. In [22], a tree-structured sparse regularization is employed to learn structured dictionaries. In [23], Dai et al. propose a new algorithm, which allows to update an arbitrary set of dictionary atoms and the corresponding sparse coefficients simultaneously. Sulam et al. propose an Online Sparse Dictionary Learning (OSDL) algorithm to handle the dictionary learning problem for high dimensional image data [24]. Raja et al. propose a method, namely cloud K-SVD, which collaboratively learns a common over-complete dictionary for distributed data [25]. In [26–28], dictionary learning is designed with the consideration of compressed sensing applications.

1.2. Dictionary learning for classification

By using an over-complete dictionary, a signal is transformed into a new representation in a higher dimensional space, where somewhat challenging problems, e.g., classification, may become easier. In contrast to the sparse representation task which concerns the approximation accuracy, the goal of classification is to determine the correct class label for the query signal. Therefore, it would be beneficial to make the learned dictionary have discriminative capability. Existing discriminative dictionary learning approaches in literature can be roughly divided into two categories.

Approaches in the first category learn a class-specific subdictionary for each signal class [29–33], and these sub-dictionaries together constitute the complete dictionary. Yang et al. propose the first method that learns a sub-dictionary for each class, and apply it to image classification [30]. In [31], Gao et al. propose an improved method that learns both the category-specific subdictionaries for each category and a shared dictionary for fine-grained categories by imposing incoherence constraints. In [32], Suo et al. incorporate the group structure with the dirty model in the training process, and achieve a hierarchical structure of each sparse representation.

Approaches in the second category learn a dictionary that is shared by all classes [6,8,34–36]. Jiang et al. propose a method, namely LC-KSVD, which uses a binary class label sparse code matrix to encourage signals from the same class with similar sparse representations [6]. In [8], Liu et al. propose a support discrimination dictionary learning (SDDL) method for image classification, where the common support of the images from the same class is automatically identified in the dictionary design. Yang et al. propose a Fisher discrimination dictionary learning method (FDDL), where the Fisher discrimination criterion is imposed on the dictionary atoms to enhance class discrimination [34]. All of these methods exploit all training samples to learn a dictionary.

Other dictionary learning approaches with a low-rank model can be found in [37–39]. As a relevant note, deep neural networks demonstrate good classification performance in many signal processing and machine learning applications. Some recent literature [40,41] suggests the combination of deep learning and dictionary learning, which improved performance compared to using deep learning alone.

1.3. Motivations and contributions

Generally, it is more likely to obtain a better discriminative dictionary if more training data are given and the learning algorithm is designed appropriately. In most of the existing work on dictionary learning for classification, it is assumed there is no outlier in the training data, which may not be the case in practical applications. In this paper, we propose a new discriminative dictionary learning framework, which simultaneously learns a discriminative dictionary and detects outliers in the data. In addition, to enhance discrimination, we force the supports of signals of different classes to have as less overlap as possible. It is worth noting that the number of supports for each class does not need to be predefined, and the proposed approach is able to automatically identify the number of dictionary atoms for each class. We also would like to emphasize that the proposed framework allows signals of different classes to have partly shared support, and thus is scalable even with a large number of classes.

Robust dictionary learning has been considered in literature [42–45]. Gribonval et al. theoretically analyze the local minimum property of dictionary learning approaches in various cases, where they assumes a complete dictionary for detecting outliers [42]. Deng et al. propose to use an intra-class variant matrix as a subdictionary to capture unbalanced lighting changes, exaggerated expressions, or occlusions [44]. Instead of learning, the intra-class variant matrix is obtained by subtracting the natural image from other images of the same class. Zhou et al. utilize the low-rank model with the convex relaxed nuclear matrix norm to capture the inherent subspace structure of clean data in [43]. However, for multiple classes with large variations, the data may not have a low rank. In [45], Xu et al. propose a two-phase test sample representation method for face recognition. In the first phase, each test sample is expressed as a linear combination of all the training samples and then several nearest neighbors of the test sample is selected; in the second phase, each test sample is represented as a linear combination of the determined nearest neighbors and then exploits the representation result to perform classification. There are clear differences between the previous robust dictionary learning methods and the proposed one in terms of the problem formulation.

The contributions of this paper are summarized as following:

1. We propose a novel dictionary learning framework for classification with outliers. The proposed approach can simultaneously identify the dictionary atoms for each class without predefining the number of atoms and detect outliers in the training data.
2. We formulate the dictionary learning framework into a nonconvex optimization problem with designed regularizers to promote a structure for conducting the two tasks, i.e., learning discriminative dictionary and detecting outliers. A coordinate decent algorithm is employed to solve the optimization problem.
3. We evaluate the classification performance of the proposed dictionary learning approach by using various real datasets. Our experimental results show that the proposed method achieves superior performance in comparison to state-of-the-art dictionary learning methods.

1.4. Notation

Here, we introduce notations that are used throughout the paper. We use bold lower-case letters such as $\mathbf{x}$ to represent vectors, bold upper-case letters such as $\mathbf{A}$ to denote matrices, bold lower-case letter with subscript such as $\mathbf{a}_j$ to represent columns of the matrix $\mathbf{A}$, and bold upper-case with subscript such as $\mathbf{A}_i$ to represent a sub-matrix of the matrix $\mathbf{A}$. $\mathbf{I}$ denotes the identity matrix with appropriate dimension inferred from the context. For any vector $\mathbf{x}$, we use $\|\mathbf{x}\|_p$ to represent its ℓ_p norm, i.e., for $p > 0$, $\|\mathbf{x}\|_p = (\sum_{i=1} x_i^p)^{1/p}$, and $\|\mathbf{x}\|_0$ calculates the number of non-zero elements of $\mathbf{x}$. For a matrix $\mathbf{X}$, we use $\|\mathbf{X}\|_{2,1}$ to denote the joint ℓ_2/ℓ_1 norm of $\mathbf{X}$, i.e., $\sum_{k=1}^{K} \left\|\mathbf{x}^k\right\|_F$, where $\mathbf{x}^k$ is the kth row of $\mathbf{X}$. diag(·) transform a vector into a diagonal matrix. $(\cdot)^T$, $(\cdot)^{-1}$ and tr(·) represent the transpose matrix, inverse matrix and trace of a