ARTICLE IN PRESS

Neurocomputing (xxxx) xxxx–xxxx

Contents lists available at ScienceDirect

Neurocomputing

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

A Novel multiple kernel-based dictionary learning for distributive and collective sparse representation based classifiers

Tahereh Zare*, Mohammad Taghi Sadeghi

Electrical Engineering Department, Yazd University, Yazd, Iran

ARTICLE INFO

Communicated by Jun Yu

Keywords:
Sparse representation classification
Dictionary learning
Multiple kernel learning
K-SVD algorithm
Kernel K-SVD algorithm

ABSTRACT

In recent years, sparse representation theory has attracted the attention of many researchers in the signal processing, pattern recognition and computer vision communities. The choice of dictionary matrix plays a key role in the sparse representation based methods. It can be a pre-defined dictionary or can be learned via an optimization procedure. Furthermore, the dictionary learning process can be extended to a non-linear setting using an appropriate kernel function in order to handle non-linear structured data. In this framework, the choice of kernel function is also a key step. Multiple kernel learning is an appealing strategy for dealing with this problem. In this paper, within the framework of kernel sparse representation based classification, we propose an iterative algorithm for coincident learning of the dictionary matrix and multiple kernel function. The weighted sum of a set of basis functions is considered as the multiple kernel function where the weights are optimized such that the reconstruction error of the sparse coded data is minimized. In our proposed algorithm, the sparse coding, dictionary learning and multiple kernel learning processes are performed in three steps. The optimization process is performed considering two different structures namely distributive and collective for the sparse representation based classifier. Our experimental results show that the proposed algorithm outperforms the other existing sparse coding based approaches. These results also confirm that the collective setting leads to better results when the number of training examples is limited. On the other hand, the distributive setting is more appropriate when there are enough training samples.

1. Introduction

In recent years, sparse representation based algorithms have become increasingly important in machine learning applications such as image denoising, object recognition and classification. The Sparse Representation based Classifier (SRC) is among these fruitful algorithms [1]. The success of sparse representation based classification stems from the fact that high-dimensional signals and images are naturally sparse in an appropriate feature space. So, such signals can be represented by a few samples (atoms) of a proper set of exemplars (dictionary). Although the SRC was originally introduced for classification task, its ability to represent an input signal with a few training samples causes to be used in different applications such as image retrieval [2], click prediction [3] and human pose recovery [4].

In the last decade, various studies have been conducted to improve the performance of this classifier. Most of the researches have focused on two aspects of the SRC. One is optimizing of the sparse coding function and the other is improving the dictionary matrix. In [4], the authors incorporated a local similarity preserving term into the objective function of sparse coding which groups similar silhouettes to alleviate the instability of sparse codes. Wang et. al. [5] demonstrated that similar inputs are represented by the similar atoms. So, they introduced Locality-constrained Linear Coding (LLC) which implements locality by projecting each descriptor into its local-coordinate system. In [6], the authors presented a method which makes use of the Histogram Intersection Kernel (HIK) technique within the LLC framework.. In the proposed method, using the feature space induced by the HIK, the dictionary is learned and the local sparse codes of the input histograms are computed.

Dictionary matrix plays a key role in sparse representation of signals. The feature vectors derived from a training data set can be directly used in order to determine the dictionary elements. Alternatively, a learning process can be applied to form a learned dictionary. It has been shown that applying a proper learning process significantly improves the results [7–9].

There are several known algorithms for dictionary learning. Among them the K-SVD algorithm is widely used due to its effectiveness in practical applications [10]. The main goal of the K-SVD algorithm is to

* Corresponding author.
E-mail addresses: t.zare@stu.yazd.ac.ir (T. Zare), m.sadeghi@yazd.ac.ir (M.T. Sadeghi).

http://dx.doi.org/10.1016/j.neucom.2016.12.056
Received 14 July 2016; Received in revised form 14 December 2016; Accepted 20 December 2016
0925-2312/

find an overcomplete dictionary matrix which contains K atoms such that the reconstruction error of the resulted sparse representation is minimized. The algorithm uses an iterative two steps algorithm in which the sparse representation of the training data and the associated dictionary elements are iteratively updated. In the K-SVD algorithm, the feature vectors extracted from a training data set are linearly combined in order to design the dictionary. However, because of the non-linear structure of some real world data, such a linear combination is not always efficient. Nonlinear transformation using kernel methods is a well-known technique widely used for generalization of linear methods. In the non-linear version of the K-SVD algorithm, the Kernel K-SVD (KK-SVD) algorithm, the data points are implicitly mapped into a new high dimensional feature space. The sparse coding and dictionary learning steps are then performed in this new feature space. The authors in [11] believe that iterative dictionary learning algorithms such as the K-SVD are computationally expensive and proposed a new method called Orthogonal Projective Sparse Coding (OPSC). This algorithm integrates the manifold learning and sparse coding techniques.

It has been shown that the kernel based sparse representation algorithms can provide better results compare to their linear counterparts [12]. However, the type of kernel function and the value(s) of the kernel parameter(s) have to be selected appropriately.

A typical solution to the problem of kernel selection is to apply the cross-validation technique in order to find the best kernel function among a set of candidates. However, this procedure is time-consuming. Moreover, there is no guarantee that the best possible solution is found. The other solution to this problem is to use an appropriate combination of different kernel functions [13]. Up to now, there are only very few works for using multiple kernels in the sparse representation field [14–17]. In [14], the authors proposed a multiple kernels Sparse Representation Classification (SRC) algorithm in which two basis kernel matrices are combined by the weighted sum rule. The resulted matrix is considered as the dictionary matrix. For each test sample, the sparse coefficients and the kernel weights are iteratively updated. This method is obviously not suitable for real-time applications. The other problem is that in their proposed method, the required two basis functions are selected by applying the cross validation procedure to a set of Gaussian and polynomial kernels. This process sounds in contradiction of the multiple kernel learning concept. In [15] also, the dictionary matrix is supposed to be the weighted sum of different kernel matrices. However, compare to [14], the weights are determined in a training phase where an iterative process is applied for determining the weights. In [16], a same structure has been considered for the multiple kernel function and the kernel weights and dictionary elements are learned using a three steps algorithm. In these three steps, the kernel weights are optimized based on graph embedding principles, the sparse coding is performed using a simple level wise pursuit scheme and the dictionary elements are learned using multiple levels of 1-D subspace clustering successively. Different image descriptors such as color, shape and texture along with a kernel mapping function have been used for generating the basis kernels.

The authors in [17] proposed a multiple instance learning algorithm using a multiple kernel dictionary learning framework in weakly supervised condition where the labels are in the form of positive and negative bags. Their proposed algorithm is composed of four steps in them the kernels weights, sparse codes of the positive and negative bags data, positive and negative dictionaries and sample selection matrix are optimized. The sample selection matrix is used for selecting a true positive sample from the related positive bag. They have defined a cost function for optimizing the kernels weights that increases discrimination between the positive and negative bags.

In this paper, we propose an iterative multiple kernel-based dictionary learning algorithm. Each iteration consists of three steps wherein them the sparse representation of the training samples, the kernels weights and the dictionary matrix elements are updated. In each step, it is supposed.

that all the other parameters except the step related ones are fixed. The optimization process is performed within the framework of the SRC algorithm considering two different structures for dictionary matrix namely the distributive and collective schemes [12]. The key contributions of our work are:

- We incorporate multiple kernel learning process into dictionary learning framework.
- We propose a new three steps algorithm for sparse coding, multiple kernel learning and dictionary learning. Our main work is on the multiple kernel learning stage.
- We introduce distributive and collective points of view in the sparse representation based classification and formulate the proposed algorithm considering these structures.
- We find an analytical solution to the multiple kernel learning stage which speeds up the learning process.
- We demonstrate the effectiveness of our proposed algorithms on three popular datasets.

1.1. A. Notations

Vectors are denoted by lowercase bold letters and matrices by uppercase bold letters. Scalars and matrix elements are shown by non-bold letters. The ℓ_0-pseudo-norm, denoted by $\|\|_2$, is the number of non-zero elements of a vector and $\|\|_F$ shows the Frobenius norm. T_0 refers to the sparsity level.

1.2. B. Paper organization

This paper is organized as follow: Section 2 defines and formulates the sparse representation problem in the kernel space. Section 3 presents the proposed multiple kernel based dictionary learning in two scenarios of collective and distributive SRC algorithm. The classification procedures for both distributive and collective structures are presented in Section 4. The experimental setups and results are presented in Section 5. Finally, Section 6 concludes the paper with the summary and suggests some future research directions.

2. Sparse representation and dictionary learning in the kernel space

Given a set of training samples, $\mathbf{Y} = [\mathbf{y}_1, \ldots, \mathbf{y}_N] \in \mathbb{R}^{N\times n}$, the goal is to learn a dictionary $D \in \mathbb{R}^{N\times K}$ with K atoms that leads to the best representation of the training samples. By best representation, we mean the one that leads to the least reconstruction error. The optimization problem for achieving this goal is as follows:

$$\begin{aligned} &\arg\min_{\mathbf{D},\mathbf{X}} \quad \|\mathbf{Y} - \mathbf{DX}\|_F^2 \\ &\text{s.t} \quad \|\mathbf{x}_i\|_0 \le T_0, \quad \|\mathbf{d}_j\|_2 = 1, \quad \forall\ i, j. \end{aligned} \tag{1}$$

where $\mathbf{x}_i$ is the sparse representation of the i-th training sample, $\mathbf{y}_i$. That is the i-th column of the sparse coefficient matrix $\mathbf{X}$ with maximum number of T_0 non-zero entries. Similarly, $\mathbf{d}_j$ is the j-th column of the dictionary matrix, $\mathbf{D}$, which is referred to as the j-th atom of $\mathbf{D}$. Two well-known algorithms for solving the above problem are the method of optimal direction (MOD) [18] and the KSVD algorithm [10].

In the above mentioned formulation of the problem, the sparse representation of the samples in the original feature space is calculated. Other feature spaces can also be taken into account. Let $\phi: \mathbb{R}^n \to H \subset \mathbb{R}^{\tilde{n}}$ be a non-linear mapping function that maps the data samples from the original feature space $\mathbb{R}^n$ into a dot product space (Hilbert space H). It is worth noting that the dimensionality of the new feature space, $\tilde{n}$, is often much larger than n. It can possibly be infinite. Using this non-linear mapping function, one can generate the non-linear form of the sparse representation problem in (1) as follows: