Knowledge-Based Systems 79 (2015) 43–50

Contents lists available at ScienceDirect

Knowledge-Based Systems

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

Particle Swarm Optimization based dictionary learning for remote sensing big data

Lizhe Wang [a,*], Hao Geng [b], Peng Liu [a], Ke Lu [c], Joanna Kolodziej [d], Rajiv Ranjan [e], Albert Y. Zomaya [f]

[a] Institute of Remote Sensing and Digital Earth, Chinese Academy of Sciences, PR China
[b] Department of Electrical Engineering and Information Sciences, University of Science and Technology of China, PR China
[c] University of the Chinese Academy of Sciences, Chinese Academy of Sciences, PR China
[d] Institute of Computer Science, Cracow University of Technology, Poland
[e] The Commonwealth Scientific and Industrial Research Organisation, Australia
[f] School of Information Technologies, The University of Sydney, Australia

ARTICLE INFO

Article history:
Received 7 April 2014
Received in revised form 3 October 2014
Accepted 6 October 2014
Available online 23 October 2014

Keywords:
Online dictionary learning
Particle Swarm Optimization
Sparse representation
Big data
Machine learning

ABSTRACT

Dictionary learning, which is based on sparse coding, has been frequently applied to many tasks related to remote sensing processes. Recently, many new non-analytic dictionary-learning algorithms have been proposed. Some are based on online learning. In online learning, data can be sequentially incorporated into the computation process. Therefore, these algorithms can train dictionaries using large-scale remote sensing images. However, their accuracy is decreased for two reasons. On one hand, it is a strategy of updating all atoms at once; on the other, the direction of optimization, such as the gradient, is not well estimated because of the complexity of the data and the model. In this paper, we propose a method of improved online dictionary learning based on Particle Swarm Optimization (PSO). In our iterations, we reasonably selected special atoms within the dictionary and then introduced the PSO into the atom-updating stage of the dictionary-learning model. Furthermore, to guide the direction of the optimization, the prior reference data were introduced into the PSO model. As a result, the movement dimension of the particles is reasonably limited and the accuracy and effectiveness of the dictionary are promoted, but without heavy computational burdens. Experiments confirm that our proposed algorithm improves the performance of the algorithm for large-scale remote sensing images, and our method also has a better effect on noise suppression.

1. Introduction

Recently, sparse representation has become a very popular topic in the area of remote sensing image processing. In many tasks related to remote sensing images, such as image segmentation, fusion, classification, reconstruction, and change detection, sparse representation is frequently employed to improve the performance of the algorithms. Modeling data as sparse combinations of atoms, which are the elements of a dictionary, can manifest the important intrinsic characteristics of remote sensing images.

There is a long research history on how to sparsely represent a signal or data by a set of bases. We also call this set of bases a dictionary. There are two different classes of dictionary: analytic and non-analytic. Many of the earlier studies on sparse representation focused on analytic dictionaries. Different bases, such as Fourier transformations, wavelets [1], curvelet [2], bandelet [3], directionlet [4], and grouplet [5], were proposed in different periods. The development of analytic dictionaries went through several stages, such as multi-resolution, localization, anisotropy, and adaptation. Another large class of dictionary is non-analytical. Unlike decompositions based on a predefined analytic base (such as a wavelet) and its variants, we can also learn a hyper complete dictionary without analytic form, which has neither fixed forms of atoms nor requires base vectors to be orthogonal. The basic assumption behind the learning approach is that the structure of complex incoherent characters can be more accurately extracted directly from the data than by using a mathematical description.

A non-analytic dictionary learning problem apparently can be modeled as a constraint-optimization problem. The optimization of both the dictionary and coefficients is non-convex, but alternative optimization is convex. Therefore, many algorithms consist of two stages: atom updating and sparse coding. The main differences between most methods, such as the method of optimal directions

* Corresponding author.
E-mail address: Lizhe.Wang@gmail.com (L. Wang).

http://dx.doi.org/10.1016/j.knosys.2014.10.004
0950-7051/

(MOD) [6], generalized PCA (GPCA) [7], and K-SVD [8], are their atom-updating stages. Obviously, we hope that dictionary learning is as efficient as possible. The direct optimization method (DOM) [9] denoted the algorithm as a one-step block-coordinate proximal gradient descent. It is more efficient than alternating optimization algorithms. On the other hand, the effectiveness of sparse representation is also important. The Fenchel duality method [10] solved this problem in a dual space and promoted the effectiveness of the dictionary. Another idea is to use a first-order series expansion instead of the dictionary-coefficient matrix product [11]. Doing so improves performance while adding only a small additional computational load. Apart from constraint optimization, we can also model the dictionary problem as a stochastic process. Non-parametric Bayesian dictionary learning (NBDL) [12] employs a truncated beta-Bernoulli process to infer an appropriate dictionary, and obtains significant improvements in image recovery [12]. Furthermore, multi-scale dictionary learning can also be presented as a fully Bayesian model [13].

Non-analytic dictionary learning is very efficient in data representation; however, it also introduces many new problems. First, the relationship between the over-complete atoms attracts much research attention. The atoms in the dictionary could be incoherent [14], multi-model [15,16], multi-dictionary [17,18], multi-scale [19,20], or hierarchical [21,22]. Second, the dictionary problem is also a supervised versus unsupervised issue. In the early research, most of the dictionary learning methods were unsupervised. Recently, with its wide applications to many different areas, using discriminative information [23–26] in the dictionary learning process has become popular. Supervised dictionary learning [27] makes the atoms more sophisticated and more flexible. Furthermore, the conception of task-driven dictionary learning was proposed [28].

For large groups of data, it is very hard to take all the data into the computation model at once. Therefore, in addition to the batch-based methods mentioned above, a group of online learning methods, such as recursive least squares dictionary learning (RLS) [29], online dictionary learning (ODL) [30], and the non-parametric Bayesian method (NBDL) [12], were developed in recent years. However, these online learning tools also led to some new problems and concerns, such as how to introduce the data into the training process in a smooth and orderly manner, how to perform dimension reduction [31], and how to optimize the structure of the dictionarys atoms. More importantly, experiments show that the accuracy of sparse representation of the dictionary produced by online learning is decreased because of the complexity of the large data sets of remote sensing images. The reason for this is that the strategy of updating atoms in ODL, RLS, or NBDL is unreasonable when handling large data. When taking both dictionaries, D, and coefficients, α, as variables, it is difficult to optimize them at the same time because of the non-convex character of the object function. Alternative optimization of atoms and coefficients decreases the accuracy of dictionary learning, especially when the data set is very large. It is easy to stop at a local extreme value under the influence of the continuous computation manner, noise, and complexity of the data. However, for large remote sensing data sets, there are often many other priors than the sparsity that we can utilize in the dictionary-learning process.

First, for a certain area, the remote sensing images for a terrestrial object at different times always show some similarities because the changing of land covers is usually slow. Second, for the same scene, the remote sensing images from different sensors often share similar textures to some extent. Therefore, when the location is given, we can usually use the history or multi-source data to guide the direction of the optimization in the atom-updating stage of dictionary learning. For example, in the atom-updating stage of the ODL method, the gradient direction can be easily guided by reference data or an existing reference dictionary.

In this paper, to effectively and sparsely represent the large remote sensing image set, we use reference data as priors and introduce PSO [32] into the atom-updating stage of the ODL algorithm. In the iteration, special atoms in the current dictionary are selected as the particles in the PSO model. In order to reduce the dimensions of the particles, every selected atom is represented by the linear combination of a reference image and the remaining atoms. To make the optimization more efficient, in PSO, the flying directions are limited to the few dimensions that are estimated by considering the relationship between the different subspaces of the atoms. Furthermore, for the redundant and cluster characters of the textures of the large remote sensing data set, the features of the reference data constrain and guide the ranges and directions of random particle movement in PSO. As a result, the flying of the particles in PSO is semi-random. This proposed semi-random PSO promotes the accuracy of the atom updating, and does not result in heavy computational burdens because of the guidance of the reference data. In the following sections, we first summarize the ODL algorithm and then propose our method based on the new atom-updating scheme.

2. Online dictionary learning based on gradient descent

The non-analytic sparse representation uses a hyper-complete dictionary matrix $D \in \mathbb{R}^{m\times n}$, which includes n atoms for columns to represent a signal $x \in \mathbb{R}^m$ as a sparse linear combination of these atoms. The representation of sample data x can be written as the approximate $x \approx D\alpha$, which satisfies $||x - D\alpha||_p \leqslant \varepsilon$. Here, the typical norm for sparse representation is l^p-norms, and usually is true in the case of $p = 2$. Dictionary learning is an optimization problem written as:

$$\arg\min_{D,\alpha} \frac{1}{2}\|X - D\alpha\|_2^2 + \lambda||\alpha||_1. \tag{1}$$

For convenience, $X \in \mathbb{R}^{m\times q}(m \ll q)$ is the training data set and $x_i \in \mathbb{R}^m$ is the ith column of training data matrix X. The dictionary is denoted by $D = \{d_1, \cdots, d_j, \cdots, and\ d_n\}$, and d_j stands for the jth column of D. λ is a regularization parameter. α is the coefficient of sparse representation. The Frobenius norm of a matrix X in $\mathbb{R}^{m\times p}$ here can be denoted by $||X||_F \triangleq (\sum_{i=1}^{m}\sum_{j=1}^{q}X[i,j]^2)^{1/2}$. The object function to be minimized in Eq. (1) is not jointly convex in α and D, but it becomes convex in one variable, keeping the other fixed. Thus, the ODL algorithm can be divided into two steps that alternately solve the optimization problem in an iterative loop. One is keeping D fixed and finding α, which is called the sparse coding stage. The other is keeping α fixed and finding D, which is called the atom-updating stage.

In the first stage, the ODL uses LARS [33] or orthogonal matching pursuit (OMP) [34] to find α:

$$\alpha_t = \arg\min_{\alpha} \frac{1}{2}||x_t - D_{t-1}\alpha||_2^2 + \lambda||\alpha||_1, \tag{2}$$

where the subscript t means the tth iteration of the ODL procedure.

In the second step, the original objective function is:

$$D = \arg\min_{D\in C} \frac{1}{t}\sum_{i=1}^{t}\frac{1}{2}||x_i - D\alpha_i||_2^2 + \lambda||\alpha_i||_1. \tag{3}$$

As we know, the optimization variable in the object function is the dictionary D; in the meantime, we use the matrix form of X instead of the vector α. The Eq. (3) can thus be rewritten as:

$$D_t = \arg\min_{D\in C} \frac{1}{2}||X - D\alpha||_F^2. \tag{4}$$