Neurocomputing 205 (2016) 222–228

Contents lists available at ScienceDirect

Neurocomputing

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

Atom Decomposition Based Subgradient Descent for matrix classification

Wenqing Chu [a], Yao Hu [a], Chen Zhao [a], Haifeng Liu [b], Deng Cai [a,*]

[a] The State Key Lab of CAD & CG, Zhejiang University, No. 388 Yu Hang Tang Road, Hangzhou 310058, China
[b] College of Computer Science, Zhejiang University, No. 388 Yu Hang Tang Road, Hangzhou 310058, China

ARTICLE INFO

Article history:
Received 30 November 2015
Received in revised form
20 March 2016
Accepted 25 March 2016
Communicated by Qingshan Liu
Available online 12 May 2016

Keywords:
Subgradient Descent
Nuclear norm minimization
Atom decomposition

ABSTRACT

Matrices are appropriate for representing a wealth of data with complex structures such as images and electroencephalogram data (EEG). To learn a classifier dealing with these matrix data, the structure information of the feature matrix is useful. In this paper, we focus on the regularized matrix classifiers whose input samples and weight parameters are both in the form of a matrix. Some existing approaches assume that the weight matrix has a low-rank structure and then utilize the popular nuclear norm of the weight matrix as a regularization term. However, the optimization methods for these matrix classifiers often involve numbers of expensive singular value decomposition (SVD) operations, which prevents scaling beyond moderate matrix sizes. To reduce the time complexity, we propose a novel learning algorithm called Atom Decomposition Based Subgradient Descent (ADBSD), which solves the optimization problem for the matrix classifier whose objective function is the combination of the Frobenius matrix norm and nuclear norm of the weight matrix along with the hinge loss function. Our ADBSD is an iterative scheme which selects the most informative rank-one matrices from the subgradient of the objective function in each iteration. We consider using the atom decomposition based methods to minimize nuclear norm because they mainly rely on the computation of top singular vector pair which leads to great advantages on efficiency. We empirically evaluate the performance of the proposed algorithm ADBSD on both synthetic and real-world datasets. Results show that our approach is more efficient and robust than the state-of-the-art methods.

1. Introduction

Linear classifiers such as support vector machines (SVM) [1] and logistic regression [2] play an important role in statistical learning. These linear classifiers often assume that the input samples are in the form of a vector. When the input samples are images or electroencephalography (EEG) datasets of alcoholism [3], a direct solution is to stack the columns of a matrix into a vector. However, this solution will bring in two drawbacks. First, naively turning a matrix into a vector destroys the structure information possessed by the matrix form. For example, the spatial relationship between some of the nearby image pixels will be eliminated [4]. And in EEG data, the correlation of the voltage values of the adjacent time points and channels will be difficult to capture [5]. Moreover, the process of vectorization produces a high-dimensional vector whose dimension is the product of the width and height of a matrix, which leads to the curse of dimensionality. Thus, it is more appropriate to represent these data as matrices.

There is a body of related work on developing effective matrix classification methods which try to capture the structure information by conveying the input samples and weight parameters through the form of a matrix. For instance, [6] proposed a bilinear classification algorithm called support tensor machine which considers a document as a second order tensor. Ref. [5] introduced a class of regularized matrix regression methods (R-GLM) based on spectral regularization, which are solved by the Nesterov optimal gradient method. Recently, [7] devised a support matrix machine model which applies the nuclear norm of the weight matrix as a convex approximation of the matrix rank. As we can see, most of these algorithms assume that the performance of the classifiers can be improved by imposing a low-rank restriction on the weight matrix. However, their methods either use a non-convex objective function or cost too much time for optimization.

* Corresponding author.
E-mail addresses: wqchu16@gmail.com (W. Chu), huyao001@gmail.com (Y. Hu), henryzhao4321@gmail.com (C. Zhao), haifengliu@zju.edu.cn (H. Liu), dengcai@cad.zju.edu.cn (D. Cai).

http://dx.doi.org/10.1016/j.neucom.2016.03.069
0925-2312/

In [8–11], rank constraint has been employed to capture the global structure of data. To leverage the structure information of the matrix data, we also consider a regularized matrix classifier based on the low-rank constraint of the weight matrix. And we utilize nuclear norm as a convex alternative of the matrix rank because nuclear norm is the tightest convex lower bound of the rank operator under certain condition. Then we formulate the objective function as the combination of the Frobenius matrix norm and nuclear norm penalty along with the hinge loss function. Ref. [7] derived the same objective function and resorted to an alternating direction method of multipliers (ADMM) to optimize it. However, their ADMM solver involves numbers of expensive singular value decomposition (SVD) operations, which brings excessive computation cost on training time and memory for large scale matrices. Motivated by this observation, we are concerned with applying greedy atom decomposition based methods to the nuclear norm minimization.

In this paper, we propose a new method called Atom Decomposition Based Subgradient Descent (ADBSD) to address the matrix classification problem. The proposed ADBSD is an iterative scheme which applies atom decomposition methods to the current subgradient matrix of the objective function and select several important bases to update the weight matrix in each step. Taking into account the speed and accuracy in the process of generating bases, a two-stage algorithm is devised. In each stage, we utilize a different method to choose the suitable bases from the current subgradient matrix and then generate the corresponding coefficients. In the first stage, we use the subgradient matrix to identify an active subspace like [12]. Then we can select several informative bases from the subspace as accurate as possible and compute the corresponding coefficients by a second order method. In the second stage, we develop a greedy atom decomposition based algorithm called Rank-one Subgradient Descent which is similar to Rank-one Descent in [13] to generate the most suitable basis which is the product of the approximate top singular-vector pair of the subgradient matrix iteratively. By alternating these two stages, our ADBSD can achieve competitive classification accuracy and extremely good computation efficiency.

The rest of the paper is arranged as follows. First, we give the notations and review the related work on nuclear norm minimization and regularized matrix classification. Then we describe our matrix classifier and the proposed learning algorithm in detail. After that, we evaluate the performance of our method on several synthetic and real-world datasets. Finally, some concluding remarks are present.

2. Notations

Before continuing, we provide here a brief summary of the notations used throughout the paper. The singular value decomposition of a matrix $\mathbf{X} \in \mathbb{R}^{p\times q}$ is $\mathbf{U}\mathbf{\Sigma}\mathbf{V}^T$, where $\mathbf{\Sigma} = diag(\sigma_1(\mathbf{X}), \ldots, \sigma_{\min(p,q)}(\mathbf{X}))$ and $\sigma_1(\mathbf{X}) \geq \sigma_2(\mathbf{X}) \geq \cdots \geq \sigma_{\min(p,q)}(\mathbf{X}) \geq 0$ are the singular values of $\mathbf{X}$. The nuclear norm is defined as $\|\mathbf{X}\|_* = \sum_{i=1}^{\min(p,q)} \sigma_i(\mathbf{X})$. In addition, the Frobenius norm is defined as $\|\mathbf{X}\|_F = \sqrt{\sum_{i,j} X_{ij}^2} = \sqrt{\sum_{i=1}^{\min(p,q)} \sigma_i(\mathbf{X})^2}$.

3. Related work

In this section, we only briefly review the recent work on nuclear norm minimization and regularized matrix classification.

3.1. Nuclear norm minimization

There has been much work on developing efficient nuclear norm minimization solvers, which fall into two categories according to the fundamental operations of these approaches. The first set of these methods involves computing expensive singular value decomposition (SVD) iteratively, including singular value thresholding (SVT) [14], Truncated Nuclear Norm Regularization (TNNR) [15] and Iteratively Reweighted Nuclear Norm (IRNN) [16]. These techniques rely on SVD calculations, which prevent their scaling to large matrices. Considering the high computational complexity of SVD, some researchers study to extend the greedy atom decomposition techniques from the vector case to the matrix case, where the atom is substituted by a rank-one matrix as one basis. The second set of these methods only requires computing the top singular vector pair (top SVD) which is used to generate the most suitable rank-one matrix as one atom [17–19]. In general, these approaches alternate between two main steps in each iteration. The first step needs to compute the most informative atom formed by the top singular vector pair of the current proximal gradient. Then the coefficients of the atoms obtained up to the current iteration should be updated in the second step. The key idea of these algorithms is the way for refining the coefficients. Although the greedy atom decomposition based methods have great advantage on efficiency, they usually bear unsatisfactory performance because some atoms selected according to local information are unimportant in the long term.

3.2. Regularized matrix classification

At first, we give the original formulation for matrix classification problem. Assuming that there is a set consisting of n training examples $S = \{\mathbf{X}_i, y_i\}_{i=1}^n$, where $\mathbf{X}_i \in \mathbb{R}^{p\times q}$ is the i-th input example and $y_i \in \{-1, 1\}$ is its corresponding category. Because the traditional linear classifiers only use a vector as the input, the convenient way is to stack the columns of $\mathbf{X}_i$ into a vector $\mathbf{x}_i$. Then one of the most successful approaches is the soft margin support vector machine (SVM) which is defined as

$$\min_{\mathbf{w},b} \frac{1}{2}(\mathbf{w}^T\mathbf{w}) + C\sum_{i=1}^{n} \left[1 - y_i(\mathbf{w}^T\mathbf{x}_i + b)\right]_+, \tag{1}$$

where $\left[1 - y_i(\mathbf{w}^T\mathbf{x}_i + b)\right]_+$ is the hinge loss, $\mathbf{w} \in \mathbb{R}^{pq}$ is a vector of regression coefficients and b is the bias. However, vectorization results in an exceedingly large dimensionality and ignores the structure information of the feature matrix.

Therefore, we express the input samples and weight parameters through the matrix form, so the objective function can be reframed as below:

$$\min_{\mathbf{W},b} \frac{1}{2}\mathrm{tr}(\mathbf{W}^T\mathbf{W}) + C\sum_{i=1}^{n} \left[1 - y_i(\mathrm{tr}(\mathbf{W}^T\mathbf{X}_i) + b)\right]_+, \tag{2}$$

where $\mathbf{W} \in \mathbb{R}^{p\times q}$ is the weight matrix and b is the bias.

Obviously, rank restriction is a general regularization way for the matrix form. Then it is natural to impose a low-rank constraint on weight matrix $\mathbf{W}$ to capture the dependence of the matrix data.

A direct regularization method is to impose a matrix norm penalty function based on the singular values of $\mathbf{W}$ on the loss function. Ref. [5] proposed a class of regularized matrix regression methods (R-GLM) based on spectral regularization:

$$\min_{\mathbf{W}} l(\mathbf{W}) + p(\mathbf{W}), \tag{3}$$

where $l(\mathbf{W})$ is a loss function and $p(\mathbf{W})$ is a penalty function based on the singular values of $\mathbf{W}$. The authors utilized the Nesterov optimal gradient method to solve (3).