Neurocomputing 151 (2015) 223–230

Contents lists available at ScienceDirect

Neurocomputing

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

An algorithm framework of sparse minimization for positive definite quadratic forms

Si-Bao Chen [a,*], Chris H.Q. Ding [b], Bin Luo [a]

[a] Key Lab of Intelligent Computing and Signal Processing of Ministry of Education, School of Computer Science and Technology, Anhui University, Hefei, Anhui, China

[b] Department of Computer Science and Engineering, University of Texas at Arlington, Engineering Research Building, Room 529, 500 UTA Blvd, Arlington, TX 76019, USA

ARTICLE INFO

Article history:
Received 30 April 2014
Received in revised form
15 August 2014
Accepted 28 September 2014
Communicated by Feiping Nie
Available online 7 October 2014

Keywords:
Sparse minimization
l_1-norm
Global minimum
Positive definite quadratic form
Pattern recognition

ABSTRACT

Many well-known machine learning and pattern recognition methods can be seen as special cases of sparse minimization of Positive Definite Quadratic Forms (PDQF). An algorithm framework of sparse minimization is proposed for PDQF. It is theoretically analyzed to converge to global minimum. The computational complexity is analyzed and compared with the state-of-the-art Fast Iterative Shrinkage-Thresholding Algorithm (FISTA). Some well-known machine learning and pattern recognition methods are illustrated to be optimized by the proposed algorithm framework. Illustrative experiments show that Sparse Representation Classification (SRC) and Least Absolute Shrinkage and Selection Operator (LASSO) via the proposed method converges much faster than several classical methods.

1. Introduction

Positive Definite Quadratic Forms (PDQF) are very common in machine learning and pattern recognition. Many methods, such as Least Squares (LS) [13], linear regression [9] and ridge regression [19], can be seen as special cases of PDQF. There is close-form solution for minimization of plain PDQF. However, if some additional constraints are imposed on the solution of PDQF, one may need to resort to optimization computation, such as quadratic programming [4] or more general convex optimization [5].

Recently, sparsity constraint, which makes most of elements of solution vector be zero, aroused great research interests due to its meaningful application backgrounds. Least Absolute Shrinkage and Selection Operator (LASSO) [18] added sparse minimization constraint (l_1-minimization) on the solution of regression between predictors and response to achieve the goal of variable selection. Elastic Net [27] added l_2-smooth term in LASSO to make the sparse regression coefficient more stable. Nie et al. [15] generalized the sparse minimization to $l_{2,1}-$norm for robust feature selection. Generalized LASSO [17] considered LASSO in kernel space. Sparse representation classification (SRC) [24] pursued sparse linear reconstruction of test sample with training samples for classification. Kernel SRC [25,10] performed SRC in kernel space. Sparse coding [14] learnt succinct representations of stimuli with sparse constraint on coefficients. There are also lots of related work of l_1-norm based optimization for linear dimensionality reduction, such as l_1-norm based Principal Component Analysis (PCA_L1) [12], l_1-norm based Linear Discriminant Analysis (LDA_L1) [21], l_1-norm based Two-Dimensional PCA (2DPCA_L1) [22] and robust sparse-preserved learning [23].

Many of these methods can be cast into sparse minimization of PDQF. Generally, quadratic programming [4] or more general convex optimization [5] can be applied to these quadratic problems. However, the optimization procedure is complicated and the computational complexity is very high. Specifically, much simpler gradient projection algorithms [8] and iterative shrinkage-thresholding algorithms (ISTA) [6], where each iteration only involves matrix-vector multiplication followed by a shrinkage/soft-threshold step, are proposed to solve the linear inverse problems with sparsity constraints. However, such methods are also known to converge quite slowly. Recently, Beck and Teboulle [3] proposed a fast iterative shrinkage-thresholding algorithm (FISTA) for linear inverse problems, which preserves the computational simplicity of ISTA but with a global rate of convergence which is proven to be significantly better. There also appear other optimization methods for solving

* Corresponding author. Tel.: +86 551 65108507; fax: +86 551 65108445.
E-mail addresses: sbchen@ahu.edu.cn (S.-B. Chen), CHQDing@uta.edu (C.H.Q. Ding), luobin@ahu.edu.cn (B. Luo).

http://dx.doi.org/10.1016/j.neucom.2014.09.049
0925-2312/

sparse minimization problems recently, such as Coordinate Gradient Descent (CGD) [26] and Alternating Direction Method of Multipliers (ADMM) [20]. However, we found that these methods converge still very slowly in practice.

In this paper, we propose a simpler and efficient iteration algorithm framework of sparse minimization for PDQF. It is proved to converge to global minimum. The computational complexity of the proposed algorithm is analyzed and compared with classical FISTA. Some well-known machine learning and pattern recognition methods are illustrated to be optimized by the proposed algorithm framework. Experiments of Sparse Representation Classifications (SRC) [24] and Least Absolute Shrinkage and Selection Operator (LASSO) via the proposed method are implemented and compared with those via classical FISTA, CGD and ADMM to show the superiority of the proposed method.

2. Sparse minimization for positive definite quadratic form

2.1. The formulation

Let the positive definite quadratic form $f(\mathbf{w})$ be

$$f(\mathbf{w}) = \mathbf{w}^\top \mathbf{A}\mathbf{w} - 2\mathbf{w}^\top \mathbf{b} + c, \tag{1}$$

where parameter $\mathbf{w} = (w_1, w_2, \ldots, w_n)^\top \in \mathbb{R}^n$, square symmetric quadratic term coefficient $\mathbf{A} \in \mathbb{R}^{n\times n}$ is strictly positive definite, satisfying $\mathbf{v}^\top \mathbf{A}\mathbf{v} > 0$ for any n-dimensional nonzero vector $\mathbf{v}$, first-order coefficient $\mathbf{b} \in \mathbb{R}^n$ and constant term $c \in \mathbb{R}$. c is omitted since it has no effect on the solution of optimizing (1). The objective function $f(\mathbf{w})$ in (1) is strictly convex with respect to $\mathbf{w}$ and is bounded below. In particular, it has one global minimum and no local minima. Usually, the solution of minimizing $f(\mathbf{w})$ directly is not sparse.

There are many circumstances where the solution of minimizing (1) should be sparse with many entries being zeros, i.e., $\|\mathbf{w}\|_0 \le T$. The l_0-norm $\|\mathbf{w}\|_0$ of vector $\mathbf{w}$ is the number of nonzero entries. The sparse-constrained minimization of $f(\mathbf{w})$ is formulated as

$$\min_{\mathbf{w}} f(\mathbf{w}), \ s.t. \ \|\mathbf{w}\|_0 \le T, \tag{2}$$

or in Lagrange multiplier formulation

$$\min_{\mathbf{w}} f(\mathbf{w}) + \lambda\|\mathbf{w}\|_0. \tag{3}$$

where $\lambda > 0$ is a tuning parameter. However, since the l_0-norm $\|\mathbf{w}\|_0$ is counting number of non-zero entries and is not differentiable, it is hard to minimize the optimization problem of (3).

To obtain a sparse solution of minimizing $f(\mathbf{w})$, the l_1-norm minimization of $\mathbf{w}$ is commonly adopted to be added as a constraint

$$\min_{\mathbf{w}} L(\mathbf{w}) = f(\mathbf{w}) + \lambda\|\mathbf{w}\|_1, \tag{4}$$

where l_1-norm $\|\mathbf{v}\|_1$ of an n-dimensional vector $\mathbf{v} = (v_1, v_2, \ldots, v_n)^\top$ is the sum of its absolute elements, $\|\mathbf{v}\|_1 = \sum_{i=1}^n |v_i|$. Based on the theory of sparse representation and compressed sensing [7], the solution of l_1-norm minimization of (4) is equivalent to that of l_0-norm minimization of (3).

Since both of the two items of (4) are convex with the first one strictly convex, then the whole formula in (4) is also strictly convex with respect to $\mathbf{w}$. Therefore, there exists a unique global minimum for sparse minimization of (4) with no local minima.

2.2. Optimization algorithm

To obtain the global minimization solution of (4), we propose an iterative algorithm, which can be summarized as in Algorithm 1. In each iteration step, diagonal matrix $\mathbf{M}$ is calculated with the current $\mathbf{w}$ as in (5), and then coefficient vector $\mathbf{w}$ is updated based on the just calculated $\mathbf{M}$ as in (6). The iteration procedure between (5) and (6) is repeated until the algorithm converges.

Algorithm 1. Procedure of l_1-minimization for positive definite quadratic form.

```
1: Input: Coefficients of positive definite quadratic form
   $\mathbf{A} \in \mathbb{R}^{n\times n}$ and $\mathbf{b} \in \mathbb{R}^n$, initial non-zero solution $\mathbf{w}^{(0)} \in \mathbb{R}^n$,
   tuning parameters $\lambda > 0$, maximum number of iterations
   $t_{\max}$ or relative residual error tolerance $\varepsilon > 0$;
2: Set $t=0$;
3: Update diagonal matrix
   $$\mathbf{M}^{(t)} = diag\left(\sqrt{|w_1^{(t)}|}, \sqrt{|w_2^{(t)}|}, \ldots, \sqrt{|w_n^{(t)}|}\right); \tag{5}$$
4: Update solution
   $$\mathbf{w}^{(t+1)} = \mathbf{M}^{(t)}\left[\mathbf{M}^{(t)}\mathbf{A}\mathbf{M}^{(t)} + \frac{\lambda}{2}\mathbf{I}_n\right]^{-1}\mathbf{M}^{(t)}\mathbf{b}, \tag{6}$$
   where $\mathbf{I}_n$ is an identity matrix of order $n$;
5: If $t > t_{max}$ or $|L(\mathbf{w}^{(t+1)}) - L(\mathbf{w}^{(t)})|/L(\mathbf{w}^{(t)}) < \varepsilon$, go to step 6,
   otherwise, let $t = t+1$ and go to step 3;
6: Output: The optimal solution $\mathbf{w}^* = \mathbf{w}^{(t+1)}$.
```

2.3. Justification

In this section, we will see that Algorithm 1 does converge to the unique global minimum of l_1-minimization problem (4). Let $L(\mathbf{w})$ denote the objective function of l_1-minimization in (4). Note that $L(\mathbf{w})$ is a strictly convex function of $\mathbf{w}$ and there are no local minima. Strictly convex functions could not have interior points being global maximum, local maxima or saddle points. Then only global minimum could satisfy the Karush-Kuhn-Tucker (KKT) first order necessary conditions (for a solution in nonlinear programming to be optimal) [5].

Therefore, we only need to prove two things. The first one is that the objective function value $L(\mathbf{w})$ is decreasing along with each iteration in Algorithm 1, which is summarized in Theorem 1. The second one is that the limit solution obtained by the iteration of Algorithm 1 satisfies the KKT conditions, which is summarized in Theorem 4.

Theorem 1. *The objective function value $L(\mathbf{w})$ in l_1-minimization (4) is decreasing, $L(\mathbf{w}^{(t+1)}) \le L(\mathbf{w}^{(t)})$, along with each alternate iteration of formulae (5) and (6) in Algorithm 1. The equality holds only at convergence.*

To prove Theorem 1, we need the help of the following two Lemmas, which are needed to be proved firstly.

Lemma 2. *Let index set $\mathcal{C}_t = \{i | w_i^{(t)} \ne 0, i = 1, 2, \ldots, n\}$. Define an auxiliary function*

$$G(\mathbf{w}, \mathbf{w}^{(t)}) = f(\mathbf{w}) + \lambda \sum_{i \in \mathcal{C}_t} \frac{w_i^2}{2|w_i^{(t)}|}. \tag{7}$$

Along with the solution sequence $\{\mathbf{w}^{(t)}, t = 0, 1, 2, \ldots\}$ obtained in Algorithm 1, the following inequality holds,

$$G(\mathbf{w}^{(t+1)}, \mathbf{w}^{(t)}) \le G(\mathbf{w}^{(t)}, \mathbf{w}^{(t)}). \tag{8}$$

The equality holds only at convergence.

Proof. Since the two terms in minimizing auxiliary function $G(\mathbf{w}, \mathbf{w}^{(t)})$ are both semi-definite programming (SDP) problems, we can obtain the unique global optimal solution of minimizing $G(\mathbf{w}, \mathbf{w}^{(t)})$ by taking the derivatives and letting them equal to zero.