Artificial Intelligence 224 (2015) 28–50

Contents lists available at ScienceDirect

Artificial Intelligence

www.elsevier.com/locate/artint

Efficient nonconvex sparse group feature selection via continuous and discrete optimization ☆

Shuo Xiang [a,b], Xiaotong Shen [c], Jieping Ye [a,b,*]

[a] Center for Evolutionary Medicine and Informatics, Arizona State University, Tempe, AZ 85287, United States
[b] Computer Science and Engineering, Arizona State University, Tempe, AZ 85287, United States
[c] School of Statistics, University of Minnesota, Minneapolis, MN 55347, United States

ARTICLE INFO

Article history:
Received 15 January 2014
Received in revised form 11 January 2015
Accepted 15 February 2015
Available online 10 March 2015

Keywords:
Nonconvex optimization
Error bound
Discrete optimization
Application
EEG data analysis

ABSTRACT

Sparse feature selection has proven to be effective in analyzing high-dimensional data. While promising, most existing works apply convex methods, which may be suboptimal in terms of the accuracy of feature selection and parameter estimation. In this paper, we consider both continuous and discrete nonconvex paradigms to sparse group feature selection, which are motivated by applications that require identifying the underlying group structure and performing feature selection simultaneously. The main contribution of this article is twofold: (1) computationally, we develop efficient optimization algorithms for both continuous and discrete formulations, of which the key step is a projection with two coupled constraints; (2) statistically, we show that the proposed continuous model reconstructs the oracle estimator. Therefore, consistent feature selection and parameter estimation are achieved simultaneously. Numerical results on synthetic and real-world data suggest that the proposed nonconvex methods compare favorably against their competitors, thus achieving desired goal of delivering high performance.

1. Introduction

During the past decade, sparse feature selection has been extensively investigated, on both optimization algorithms [1] and statistical properties [39,52,4]. When the data possesses certain group structure, sparse modeling has been explored in [49,30,21,47] for group feature selection. The group lasso [49,40] proposes an L_2-regularization method for each group, which ultimately yields a group-wisely sparse model. The utility of such a method has been demonstrated in detecting splice sites [48]—an important step in gene finding and theoretically justified in [21]. The sparse group lasso [17] enables to encourage sparsity at the level of both features and groups simultaneously. In the literature, most approaches use convex methods to pursue the grouping effect due to globality of the solution and tractable computation. However, this may lead to suboptimal results. Recent studies demonstrate that nonconvex methods [14,42,8,20,22], particularly the truncated L_1-penalty [35,29,51], may deliver superior performance than the standard L_1-formulation. In addition, [36] suggests that a constrained nonconvex formulation is slightly more preferable than its regularization counterpart in terms of the capability of feature selection. In this paper, we investigate the sparse group feature selection through a constrained nonconvex

☆ This paper is an invited revision of a paper first published at "The 30th International Conference on Machine Learning (ICML 2013)".

* Corresponding author.
E-mail address: jieping.ye@asu.edu (J. Ye).

http://dx.doi.org/10.1016/j.artint.2015.02.008
0004-3702/

formulation. Ideally, we wish to optimize the following L_0-model:

$$\begin{aligned} \underset{\boldsymbol{x}}{\text{minimize}} \quad & \frac{1}{2}\|\boldsymbol{A}\boldsymbol{x}-\boldsymbol{y}\|_2^2 \\ \text{subject to} \quad & \sum_{j=1}^{p} I(|x_j|\neq 0) \leq s_1 \\ & \sum_{j=1}^{|G|} I(\|\boldsymbol{x}_{G_j}\|_2 \neq 0) \leq s_2, \end{aligned} \tag{1}$$

where $\boldsymbol{A}$ is an n by p data matrix with its columns representing different features. $\boldsymbol{x}=(x_1,\cdots,x_p)$ is partitioned into $|G|$ non-overlapping groups $\{\boldsymbol{x}_{G_i}\}$ and $I(\cdot)$ is the indicator function. The advantage of the L_0-model (1) lies in its complete control on two levels of sparsity (s_1, s_2), which are the numbers of features and groups respectively. However, problems such like (1) are known to be NP-hard [31] because of the discrete nature.

We develop two methods for the sparse group feature selection problem. The first method makes use of a continuous computational surrogate of the L_0-method described above, and has theoretically guaranteed performance. On the contrary, the second proposed method retains the discrete nature and the key is to solve a sparse group subset selection problem via dynamic programming. We develop efficient algorithms for both methods. In addition, we explore the statistical properties of the first method; specifically we show that the proposed method retains the merits of the L_0-approach (1) in the sense that the oracle estimator can be reconstructed, which leads to consistent feature selection and parameter estimation. An earlier version of this paper [45] containing only the first approach was accepted by the 30th International Conference on Machine Learning (ICML).

The rest of this paper is organized as follows. Section 2 presents our continuous optimization approach, in which a nonconvex formulation with its optimization algorithm and theoretical properties are explored. The discrete optimization approach is discussed in Section 3, where we transform the key projection into a discrete sparse group subset selection problem and develop a dynamic programming algorithm to compute the globally optimal solution. The significance of this work is presented in Section 4. Section 5 demonstrates the efficiency of the proposed methods as well as the performance on real-world applications. Section 6 concludes the paper with a discussion of future research.

2. Continuous optimization approach

One major difficulty of solving (1) comes from nonconvex and discrete constraints, which require enumerating all possible combinations of features and groups to achieve the optimal solution. Therefore we approximate these constraints by their continuous computational surrogates:

$$\begin{aligned} \underset{\boldsymbol{x}}{\text{minimize}} \quad & \frac{1}{2}\|\boldsymbol{A}\boldsymbol{x}-\boldsymbol{y}\|_2^2 \\ \text{subject to} \quad & \sum_{j=1}^{p} J_\tau(|x_j|) \leq s_1 \\ & \sum_{i=1}^{|G|} J_\tau(\|\boldsymbol{x}_{G_i}\|_2) \leq s_2, \end{aligned} \tag{2}$$

where $J_\tau(z)=\min(|z|/\tau,1)$ is a truncated L_1-function approximating the L_0-function [35,50], and $\tau>0$ is a tuning parameter such that $J_\tau(z)$ approximates the indicator function $I(|z|\neq 0)$ as τ approaches zero.

To solve the nonconvex problem (2), we develop a Difference of Convex (DC) algorithm [38] based on a decomposition of each nonconvex constraint function into a difference of two convex functions:

$$\sum_{j=1}^{p} J_\tau(|x_j|) = S_1(\boldsymbol{x}) - S_2(\boldsymbol{x}),$$

where

$$S_1(\boldsymbol{x}) = \frac{1}{\tau}\sum_{j=1}^{p}|x_j|, \quad S_2(\boldsymbol{x}) = \frac{1}{\tau}\sum_{j=1}^{p}\max\{|x_j|-\tau,0\}$$

are convex in $\boldsymbol{x}$. Then each trailing convex function, say $S_2(\boldsymbol{x})$, is replaced by its affine minorant at the previous iteration

$$S_1(\boldsymbol{x}) - S_2(\hat{\boldsymbol{x}}^{(m-1)}) - g^T(\boldsymbol{x}-\hat{\boldsymbol{x}}^{(m-1)}), \tag{3}$$