Pattern Recognition 83 (2018) 1–13

Contents lists available at ScienceDirect

Pattern Recognition

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

Collinear groupwise feature selection via discrete fusion group regression

Younghoon Kim, Seoung Bum Kim*

School of Industrial Management Engineering, Korea University, 145 Anam-dong, Seongbuk-gu, Seoul 02841, Republic of Korea

ARTICLE INFO

Article history:
Received 6 August 2017
Revised 12 February 2018
Accepted 13 May 2018
Available online 15 May 2018

Keywords:
Multiple linear regression
Machine learning
Feature selection
Multicollinearity
Mixed-integer quadratic programming
Best subset selection

ABSTRACT

We propose a method to select the subset of features in multiple linear regression models that considers the collinearity between features. The proposed method first detects collinear groups of features and then uses collinear groupwise feature selection constraints to estimate the coefficients of the regression model. The constraints simultaneously control the number of features selected and predefined collinear feature groups. We manage the multicollinearity in the regression model by controlling the parameters of the fusion group constraint. To address the NP-hard problem of the proposed method, we propose a modified discrete first-order algorithm. We use simulation and real-world data to demonstrate the usefulness of the proposed method by comparing it to existing regularization and discrete optimization-based methods in terms of predictive accuracy, bias, and variance. The comparison confirms that the proposed method outperforms the alternatives.

1. Introduction

Today, hundreds or even thousands of features can characterize the data sets in applications. In many situations, the number of features exceeds the number of samples. In these highly dimensional situations, dimensionality reduction via feature selection is necessary to simplify the entire modeling process and reduce the computational loads [1–4]. Feature selection is of interest especially with large numbers of candidate features and the possibility that many redundant or irrelevant features are present. Unnecessary features increase the size of the search space and thus degrade the generalization performance of the predictive model. The predictive performance of linear regression methods is especially dependent on how efficiently the methods learn patterns between the response variables and a set of predictors. Redundant features, which are highly correlated with each other, complicate the extraction of the meaningful patterns necessary for prediction. Further, a greater number of features heightens the risk of overfitting. Many regularization approaches have been proposed to reduce high dimensionality [5]. These approaches can be categorized into two groups: feature extraction and feature selection. Feature extraction methods transform the original features to extract a new set of features [6,7]. These are efficient to reduce the dimension of the features; however, the transformed features lose interpretability.

Unlike feature extraction, feature selection returns a subset of the original features without transformation. In the process of selecting features for regression, given a set of training data, selection involves identifying a subset of features that lead to an "optimal" characterization of the different responses [2]. Li et al. [8] comprehensively surveyed feature selection algorithms including filter, wrapper, and embedded methods. In the present study, we focus on embedded methods that capitalize on both the filter and wrapper methods.

Embedded regression models such as least absolute shrinkage and selection operator (LASSO) [9] simultaneously perform feature selection and prediction. LASSO uses an L_1 penalty of the coefficient vector to induce sparsity in the obtained coefficient vector, which also selects features. However, in the presence of highly correlated features, LASSO tends to select only one, or seldom more than a small number, of these features, which leads to an increased variance of coefficients and impaired predictive performance. To compensate for this limitation, methods have been proposed to select groups of features and simultaneously select groups of highly correlated features. Elastic net [10] is a representative method of feature group selection that uses both L_1 and L_2 penalties for the residual norm. These penalties combine the advantages of both LASSO and ridge regression [11], which adopts L_2 penalties to reduce the variance of the coefficients. The L_1 and L_2 penalties in the elastic net can control not only sparsity and shrinkage but also the degree of collinearity between selected features. A sparse group LASSO capable of using predefined feature group information to select features was proposed [12]. Recently, the octagonal shrink-

* Corresponding author.
E-mail address: sbkim1@korea.ac.kr (S.B. Kim).

https://doi.org/10.1016/j.patcog.2018.05.013
0031-3203/

age and clustering algorithm for regression (OSCAR) [13], which combines an L_1 penalty and pairwise L_∞ penalties on each pair of features, has demonstrated superior performance in prediction and selecting feature groups. Moreover, as an extension of OSCAR, the hexagonal operator for regression with shrinkage and equality selection (HORSES) [14] has been proposed to further improve predictive performance and feature selection. In the remote sensing area, the dual clustering-based hyperspectral band selection method was proposed [15]. This method clusters the band with a dual clustering algorithm with the contextual information and selects representative bands. Wang et al. [16] proposed salient band selection that groups original band images into subsets and uses a manifold ranking method to rank and select the salient bands.

Previous embedded regression methods are based on continuous convex optimization. These types of methods are widely used because of the ease of computation [8]. Nevertheless, discrete optimization-based methods for best subset selection that select features more sparsely than LASSO-type methods and with lower coefficient bias have been recently proposed [17–21]. Based on the principle of parsimony [22], if models perform comparably, those models with the least number of features are preferred. These discrete optimization-based methods use nonconvex cardinality constraints that directly control the number of features selected instead of using a convex L_1 or L_2 penalty that shrinks coefficients to zero. Thus, discrete optimization-based methods can select a smaller number of features with lower bias in estimating coefficients [19]. Miyashiro & Takano [17] attempted to select the best subset of features in multiple linear regression models. They proposed a second-order cone programming approach to select the best subset of features with respect to adjusted R^2, Akaike information criterion (AIC), and Bayesian information criterion (BIC). In the same year, they also proposed a mixed-integer quadratic programming-based subset selection method. It operates by minimizing Mallows' C_p, one of the goodness of fit measures [18]. This formulation is more efficient than previous AIC and BIC-based formulations. Bertsimas et al. [19] proposed a mixed-integer optimization (MIO) approach to solve the classical best subset selection problem of choosing k out of p features in linear regression. They developed a discrete extension of modern first-order continuous optimization methods to achieve high quality feasible solutions. Kimura & Waki [20] proposed a branch-and-bound search algorithm for mixed-integer nonlinear programming to minimize AIC. The method selects important features more sparsely than previous methods. Mazumder & Radchenko [21] proposed the discrete Dantzig selector, which minimizes the number of nonzero regression coefficients with a budget constraint on the maximal absolute correlation between features and residuals. Although all of these methods yielded reliable results in their settings, none considered the collinearity between features (i.e., multicollinearity), which is a critical issue to overcome to improve predictive performance in regression problems. Selecting the important features while also considering multicollinearity can enhance the performance of the feature selection model in practice [10,23,24].

In this paper, we propose a discrete optimization-based method to select the best subset of features in multiple linear regression models while simultaneously considering correlation between the features. To control the collinearity between the selected features, we propose using discrete fusion group regression, a new method for discrete optimization-based regularization and best subset selection. We illustrate the overall framework of the proposed discrete fusion group regression method in Fig. 1.

The proposed method first detects collinear groups of features by finding isolated feature subgraphs in which edges have larger correlation values than the minimum correlation threshold. Then, the method estimates the coefficients of the regression model with collinear groupwise feature selection constraints. The constraints simultaneously control the number of features selected and predefined collinear feature groups. The variance of coefficients can be controlled by the parameters of the fusion group constraint, which limits the difference between coefficients within the feature group. Hence, the variance of the estimated coefficients can be reduced. Moreover, we propose a modified version of the discrete first-order algorithm formulated by mixed-integer quadratic programming to efficiently solve the best subset selection problem. Note that the best subset selection problem is nondeterministic polynomial time (NP)-hard [19]. The discrete first-order algorithm iteratively updates current solution with gradient descent and projects the solution to constraint space to produce the high quality and near-optimal solution.

The main contributions of this paper can be summarized as follows:

(1) We propose the discrete fusion group regression method that minimizes the least square error subject to collinear groupwise feature selection constraints. The constraints control the number of selected features and predefined collinear feature groups. We formulate the method with mixed-integer quadratic programming to reduce the bias of the estimated coefficients. Unlike conventional convex optimization-based methods, the bias can be diminished because the coefficients are not reduced.
(2) We propose using the fusion group constraint to address the problem of the variance increase of coefficients caused by using discrete optimization. The fusion group constraint controls the sum of pairwise differences between coefficients within a collinear group. By constraining the intragroup differences, the constraint prevents certain coefficients from becoming overly large or excessively small compared to the other coefficients in the same group.
(3) Motivated by recent algorithmic developments in first-order convex optimization, we develop a modified version of the discrete first-order method to produce high quality and near-optimal solutions for discrete fusion group regression problems. Although the proposed method does not provide global optimal solutions, it can solve the NP-hard best subset selection problem with acceptable speed and accuracy.
(4) To demonstrate the usefulness and applicability of the proposed method, we use simulated and real-world highly dimensional data to compare the proposed method with existing regularization methods in terms of predictive accuracy and efficiency in selecting features. The results confirm that the proposed method outperforms the alternatives.

The remainder of this paper is organized as follows. In Section 2, we review existing regularization and MIO-based best subset selection methods associated with the present study. In Section 3, we present details of the proposed method for collinear groupwise regression. Section 4 presents a simulation study to examine the performance of the proposed method and compare it with other methods under different scenarios. Section 5 presents a case study to demonstrate the applicability of the proposed method. Finally, Section 6 offers our concluding remarks.

2. Review of existing regularization methods

In this section, we review five regularization methods (LASSO, elastic net, sparse group LASSO, OSCAR, and HORSES). Further, we review MIO best subset selection (MBSS), which is one of the representative discrete optimization-based methods. Later in this paper, we compare the proposed method with these six methods. We summarize the formulations of the existing regularization methods in Table 1.

LASSO selects features sparsely using L_1 penalty. The sparsity is controlled by the parameter λ_1. As λ_1 becomes larger, features are