Neurocomputing 000 (2017) 1–7

Contents lists available at ScienceDirect

Neurocomputing

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

Feature selection under regularized orthogonal least square regression with optimal scaling

Rui Zhang[a], Feiping Nie[a,*], Xuelong Li[b]

[a] School of Computer Science and Center for OPTical IMagery Analysis and Learning (OPTIMAL), Northwestern Polytechnical University, Xi'an, Shaanxi 710072, P.R. China

[b] Center for OPTical IMagery Analysis and Learning (OPTIMAL), State Key Laboratory of Transient Optics and Photonics, Xi'an Institute of Optics and Precision Mechanics, Chinese Academy of Sciences, Xi'an, Shaanxi 710119, P.R. China

ARTICLE INFO

Article history:
Received 23 December 2016
Revised 13 May 2017
Accepted 17 July 2017
Available online xxx

Communicated by Dr. Chenping Hou

Keywords:
Feature selection
Orthogonal least square regression
Optimal scaling
Sparse projection

ABSTRACT

Due to lack of scale change in orthogonal least square regression (OLSR), the scaling term is introduced to OLSR to build up a novel orthogonal least square regression with optimal scaling (OLSR-OS) problem in this paper. In addition, the proposed OLSR-OS problem is proven to be numerically better than the OLSR problem. In order to select relevant features under the proposed OLSR-OS problem, $\ell_{2,1}$-norm regularization is further introduced, such that row-sparse projection is achieved. Accordingly, a novel parameterized expansion balanced feature selection (PEB-FS) method is derived based on an extension balanced counterpart. Moreover, not only the convergence of the proposed PEB-FS method is provided but the optimal scaling can be automatically achieved as well. Consequently, the effectiveness and the superiority of the proposed PEB-FS method are verified both theoretically and experimentally.

1. Introduction

Serving as a crucial problem for feature selection [1–3], classification [4,5], and sparse representation [6–9], ridge regression [10,11] has numerous applications in computer science and pattern recognition. In [12], ridge regression is used for the knowledge-leveraged inductive transfer learning. In [13], kernel ridge regression is utilized to predict the internal bond strength in a medium density fiberboard process.

To retain more statistical and structural properties, ridge regression could be further restricted on the Stiefel manifold [14] as

$$\nu_{\mathbf{m,c}} = \{\mathbf{W} \in \mathbb{R}^{\mathbf{m}\times\mathbf{c}} : \mathbf{W}^{\mathrm{T}}\mathbf{W} = \mathbf{I_c}\}$$

where the Stiefel manifold $\boldsymbol{\nu}_{\mathbf{m,c}}$ is a set of the orthogonal matrices $\mathbf{W} \in \mathbb{R}^{\mathbf{m}\times\mathbf{c}}$ with $\mathbf{m} \geq \mathbf{c}$.

Consequently, ridge regression changes into the orthogonal least square regression [15–17]. Admittedly, orthogonal least square regression performs statistically better than ridge regression does due to associated orthogonality. However, it becomes extremely difficult to achieve the closed form solution with the orthogonal constraint. To cope with this issue, orthogonal least square regression is frequently related to the procrustes problem on the Stiefel manifold [18]. Nonetheless, the procrustes problem has limited applications, since it mainly admits square matrix as input. Thus, methods [19,20] for solving the orthogonal procrustes problem can not be directly utilized to solve the orthogonal least square regression. Furthermore, the issue concerning the inflexibility of the subspace, i.e., lack of scale change is inevitable in the orthogonal least square regression though the statistical structure of the data can be highly preserved during the dimensionality reduction.

To address the defects previously mentioned, we introduce the scaling term to the orthogonal least square regression to construct a novel orthogonal least square regression with optimal scaling (OLSR-OS) problem, such that scale change is taken into consideration. Additionally, $\ell_{2,1}$-norm regularization is further introduced to the proposed OLSR-OS problem, such that relevant features can be selected efficiently under row-sparse projection. Moreover, an original parameterized expansion balanced feature selection (PEB-FS) method is derived to solve regularized OLSR-OS along with achieving the optimal scaling automatically. Consequently, extensive experiments are performed to show the effectiveness and the superiority of the proposed PEB-FS method.

This paper is organized as the following order. In Section 2, we propose the OLSR-OS problem by revisiting the ridge regression (RR) and the orthogonal least square regression (OLSR). In Section 3, a novel PEB-FS method is derived to solve the proposed OLSR-OS problem with sparse-inducing regularization. Besides, convergence of the PEB-FS method is also provided. In

* Corresponding author.
E-mail addresses: ruizhang8633@gmail.com (R. Zhang), feipingnie@gmail.com (F. Nie), xuelong_li@opt.ac.cn (X. Li).

http://dx.doi.org/10.1016/j.neucom.2017.07.064
0925-2312/© 2017 Published by Elsevier B.V.

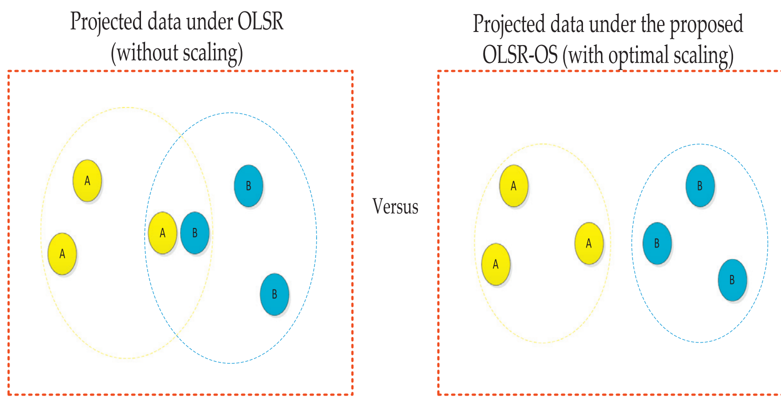

Fig. 1. OLSR versus the proposed OLSR-OS.

Section 4, the experiments are presented to show the superiority of our method. Eventually, Section 5 concludes the paper.

Notations: $\mathbf{1} = (1, 1, \dots, 1)^{\mathbf{T}} \in \mathbb{R}^{\mathbf{n}\times \mathbf{1}}$. The ℓ_p-norm of vector $\mathbf{v}$ is defined as $\|\mathbf{v}\|_{\mathbf{p}} = (\sum_{\mathbf{i}} |\mathbf{v_i}|^{\mathbf{p}})^{\frac{1}{\mathbf{p}}}$. Suppose $\mathbf{w^i}$ is the $\mathbf{i}$th row of matrix $\mathbf{W} = \{\mathbf{w_{ij}}\} \in \mathbb{R}^{\mathbf{m}\times\mathbf{c}}$, then Frobenius norm is defined as $\|\mathbf{W}\|_{\mathbf{F}} = \sqrt{\sum_{\mathbf{i=1}}^{\mathbf{m}} \sum_{\mathbf{j=1}}^{\mathbf{c}} \mathbf{w_{ij}^2}} = \sqrt{\sum_{\mathbf{i=1}}^{\mathbf{m}} \|\mathbf{w^i}\|_{\mathbf{2}}^{\mathbf{2}}}$ and the $\ell_{2,1}$-norm of matrix $\mathbf{W}$ is defined as $\|\mathbf{W}\|_{\mathbf{2,1}} = \sum_{\mathbf{i=1}}^{\mathbf{m}} \sqrt{\sum_{\mathbf{j=1}}^{\mathbf{c}} \mathbf{w_{ij}^2}} = \sum_{\mathbf{i=1}}^{\mathbf{m}} \|\mathbf{w^i}\|_{\mathbf{2}}$.

2. Orthogonal least square regression with optimal scaling

Suppose the input data $\mathbf{X} \in \mathbb{R}^{\mathbf{m}\times\mathbf{n}}$ and associated binary label matrix $\mathbf{Y} \in \mathbb{R}^{\mathbf{n}\times\mathbf{c}}$, where $\mathbf{m}$ is dimension number, $\mathbf{n}$ is data number, and $\mathbf{c}$ is class number, then ridge regression (RR) can be illustrated as the following form:

$$\min_{\mathbf{W},\mathbf{b}} \|\mathbf{X}^{\mathbf{T}}\mathbf{W} + \mathbf{1}\mathbf{b}^{\mathbf{T}} - \mathbf{Y}\|_{\mathbf{F}}^{\mathbf{2}} + \lambda\|\mathbf{W}\|_{\mathbf{F}}^{\mathbf{2}} \tag{1}$$

where $\lambda \in \mathbb{R}$ is the regularization parameter, $\mathbf{W} \in \mathbb{R}^{\mathbf{m}\times\mathbf{c}}$ is the subspace and $\mathbf{b} \in \mathbb{R}^{\mathbf{c}\times\mathbf{1}}$ is the bias.

To well preserve statistical properties of the input data and prevent tuning the regularization parameter $\boldsymbol{\lambda}$, we further restrain RR in (1) on the Stiefel manifold. Thus, the orthogonal least square regression (OLSR) could be represented as

$$\min_{\mathbf{W}^{\mathbf{T}}\mathbf{W}=\mathbf{I_c},\mathbf{b}} \|\mathbf{X}^{\mathbf{T}}\mathbf{W} + \mathbf{1}\mathbf{b}^{\mathbf{T}} - \mathbf{Y}\|_{\mathbf{F}}^{\mathbf{2}}. \tag{2}$$

Due to the orthogonal constraint $\mathbf{W}^{\mathbf{T}}\mathbf{W} = \mathbf{I_c}$, i.e., $\|\mathbf{W}\|_{\mathbf{F}}^{\mathbf{2}} = \mathbf{c}$, OLSR in (2) has fixed scale for subspace $\mathbf{W}$. To address the defect concerning lack of scale change as shown in Fig. 1, the scaling term $\boldsymbol{\gamma}$ is introduced to OLSR in (2), such that a novel orthogonal least square regression with optimal scaling (OLSR-OS) problem can be proposed as

$$\min_{\mathbf{W}^{\mathbf{T}}\mathbf{W}=\mathbf{I_c},\mathbf{b},\gamma} \|\mathbf{X}^{\mathbf{T}}(\gamma\mathbf{W}) + \mathbf{1}\mathbf{b}^{\mathbf{T}} - \mathbf{Y}\|_{\mathbf{F}}^{\mathbf{2}}. \tag{3}$$

Apparently, the proposed OLSR-OS model in (3) should be better than OLSR model in (2), since scaling $\boldsymbol{\gamma}$ is further taken into consideration. Moreover, it is interpreted both in Fig. 2 and in Theorem 2.1 that the proposed OLSR-OS in (3) is numerically better than OLSR in (2).

We will derive a novel parameterized expansion balanced method to solve the proposed OLSR-OS problem (3) with achieving the optimal scaling automatically.

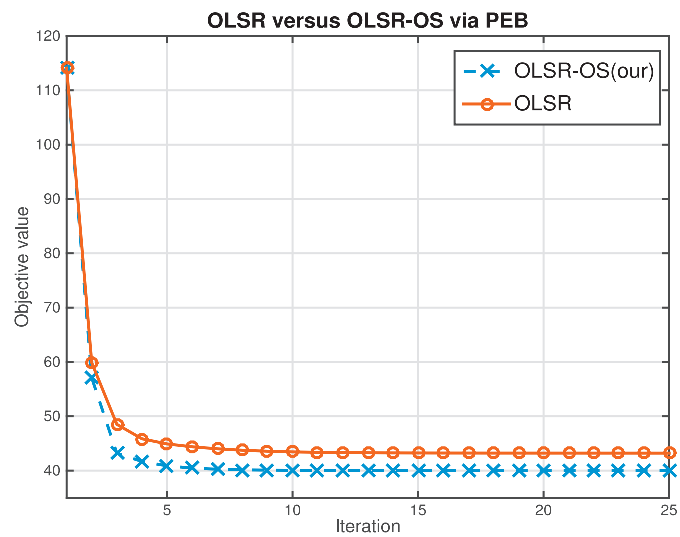

Fig. 2. The comparison of the objective value is performed for OLSR-OS in (3) and OLSR in (2) via the proposed PEB-FS method ($\lambda = \mathbf{0}$) under the same input data.

Apparently, bias $\mathbf{b}$ is free from any constraint in Eq. (3). By means of the extreme value condition w.r.t. $\mathbf{b}$, we can derive that

$$\begin{aligned}
&\frac{\partial \|\gamma\mathbf{X}^{\mathbf{T}}\mathbf{W} + \mathbf{1}\mathbf{b}^{\mathbf{T}} - \mathbf{Y}\|_{\mathbf{F}}^{\mathbf{2}}}{\partial \mathbf{b}} = \mathbf{0} \\
&\Rightarrow \frac{\partial \mathbf{Tr}(\mathbf{b}\mathbf{1}^{\mathbf{T}}\mathbf{1}\mathbf{b}^{\mathbf{T}} + \mathbf{2}(\gamma\mathbf{X}^{\mathbf{T}}\mathbf{W} - \mathbf{Y})^{\mathbf{T}}\mathbf{1}\mathbf{b}^{\mathbf{T}})}{\partial \mathbf{b}} = \mathbf{0} \\
&\Rightarrow \mathbf{b} = \frac{\mathbf{1}}{\mathbf{n}}(\mathbf{Y}^{\mathbf{T}}\mathbf{1} - \gamma\mathbf{W}^{\mathbf{T}}\mathbf{X}\mathbf{1}).
\end{aligned}$$

By substituting the above result as $\mathbf{b} = \frac{1}{\mathbf{n}}(\mathbf{Y}^{\mathbf{T}}\mathbf{1} - \gamma\mathbf{W}^{\mathbf{T}}\mathbf{X}\mathbf{1})$, the proposed OLSR-OS in (3) is reformulated into the centralized form:

$$\min_{\mathbf{W}^{\mathbf{T}}\mathbf{W}=\mathbf{I_c},\gamma} \left\| \left(\mathbf{I_n} - \frac{\mathbf{1}}{\mathbf{n}}\mathbf{1}\mathbf{1}^{\mathbf{T}}\right)(\gamma\mathbf{X}^{\mathbf{T}}\mathbf{W} - \mathbf{Y}) \right\|_{\mathbf{F}}^{\mathbf{2}}. \tag{4}$$

Accordingly, Eq. (4) can be further expanded into the following parameterized quadratic problem with orthogonal constraint:

$$\min_{\mathbf{W}^{\mathbf{T}}\mathbf{W}=\mathbf{I_c},\gamma} \mathbf{Tr}(\gamma^{\mathbf{2}}\mathbf{W}^{\mathbf{T}}\mathbf{A}\mathbf{W} - \mathbf{2}\gamma\mathbf{W}^{\mathbf{T}}\mathbf{B}) \tag{5}$$

where $\mathbf{A} = \mathbf{X}(\mathbf{I_n} - \frac{1}{\mathbf{n}}\mathbf{1}\mathbf{1}^{\mathbf{T}})\mathbf{X}^{\mathbf{T}} \in \mathbb{R}^{\mathbf{m}\times\mathbf{m}}$ and $\mathbf{B} = \mathbf{X}(\mathbf{I_n} - \frac{1}{\mathbf{n}}\mathbf{1}\mathbf{1}^{\mathbf{T}})\mathbf{Y} \in \mathbb{R}^{\mathbf{m}\times\mathbf{c}}$.

Based on Eq. (5), we have the following theorem to illustrate that the proposed OLSR-OS in (3) is numerically better than OLSR in (2).

Please cite this article as: R. Zhang et al., Feature selection under regularized orthogonal least square regression with optimal scaling, Neurocomputing (2017), http://dx.doi.org/10.1016/j.neucom.2017.07.064