Neural Networks 102 (2018) 36–47

Contents lists available at ScienceDirect

Neural Networks

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

Inter-class sparsity based discriminative least square regression

Jie Wen [a,b], Yong Xu [a,b,*], Zuoyong Li [c], Zhongli Ma [c,d], Yuanrong Xu [a,b]

[a] Bio-Computing Research Center, Shenzhen Graduate School, Harbin Institute of Technology, Shenzhen, 518055, Guangdong, China
[b] Shenzhen Medical Biometrics Perception and Analysis Engineering Laboratory, Harbin Institute of Technology (Shenzhen), Shenzhen, 518055, Guangdong, China
[c] Fujian Provincial Key Laboratory of Information Processing and Intelligent Control, Minjiang University, Fuzhou, 350121, Fujian, China
[d] College of Automation, Harbin Engineering University, Harbin, 150001, Heilongjiang, China

ARTICLE INFO

Article history:
Received 31 August 2017
Received in revised form 9 December 2017
Accepted 2 February 2018
Available online 21 February 2018

Keywords:
Least square regression
Inter-class sparsity
Multi-class classification
Supervised learning

ABSTRACT

Least square regression is a very popular supervised classification method. However, two main issues greatly limit its performance. The first one is that it only focuses on fitting the input features to the corresponding output labels while ignoring the correlations among samples. The second one is that the used label matrix, *i.e.*, zero–one label matrix is inappropriate for classification. To solve these problems and improve the performance, this paper presents a novel method, *i.e.*, inter-class sparsity based discriminative least square regression (ICS_DLSR), for multi-class classification. Different from other methods, the proposed method pursues that the transformed samples have a common sparsity structure in each class. For this goal, an inter-class sparsity constraint is introduced to the least square regression model such that the margins of samples from the same class can be greatly reduced while those of samples from different classes can be enlarged. In addition, an error term with row-sparsity constraint is introduced to relax the strict zero–one label matrix, which allows the method to be more flexible in learning the discriminative transformation matrix. These factors encourage the method to learn a more compact and discriminative transformation for regression and thus has the potential to perform better than other methods. Extensive experimental results show that the proposed method achieves the best performance in comparison with other methods for multi-class classification.

1. Introduction

Least squares regression (LSR) has been proved to be an effective technique in the community of pattern classification and computer vision, such as face recognition (Xu et al., 2014), microarray gene classification (Li & Ngom, 2013), cancer classification (Guyon, Weston, Barnhill, & Vapnik, 2002), speech recognition (Kim & Gales, 2011), and image retrieval (Feng, Zhou, & Lan, 2016). LSR aims at learning a transformation to connect the source data and target data with the minimum regression errors. In the past decades, various LSR based methods have been proposed, such as partial LSR (Abdi, 2010), local LSR (Ruppert, Sheather, & Wand, 1995), locally weighted LSR (Ruppert & Wand, 1994), kernel LSR (Gao, Shi, & Liu, 2007), and support vector machine (SVM) (Chang & Lin, 2011; Cherkassky & Ma, 2004). Besides, some representation based classification methods, such as linear regression based classification (LRC) (Naseem, Togneri, & Bennamoun, 2010) and sparsity representation based classification (SRC) (Wright, Yang, Ganesh, Sastry, & Ma, 2009), can be also regarded as the LSR based methods since they use the LSR technique to learn the representation coefficient for classification. Moreover, the very popular subspace learning methods, such as principle component analysis (PCA) , linear discriminant analysis (LDA), locality preserving projections (LPP), and spectral clustering (SC) can be also extended to the LSR framework (Cai, He, & Han, 2007; De la Torre, 2012; Tibshirani, 2011; Wang & Gao, 2015; Wen et al., 2018; Ye, 2007; Zou, Hastie, & Tibshirani, 2006). Compared with the conventional subspace learning methods, the LSR-type methods are more favorable since it is flexible to introduce various meaningful regularizations to improve their interpretability and performances. Moreover, the LSR-type methods can overcome the small-sample-size problem and greatly improve the computational efficiency (Fang, Xu, Li, Lai, Teng et al., 2017; Tibshirani, 2011).

Linear regression (LR) is one of the most popular supervised LSR methods. It has been applied in various classification tasks owing to its good performance and computational efficiency. For multi-class classification tasks, the standard LR first defines a label matrix according to the class labels and then seeks for a transformation matrix that can perfectly transform the samples into their corresponding labels. Under a mild condition, LR is equivalent to the well-known discriminative feature extraction method, *i.e.*, LDA, for multi-class classification (Ye, 2007). LDA seeks for a linear

* Corresponding author at: Bio-Computing Research Center, Shenzhen Graduate School, Harbin Institute of Technology, Shenzhen, 518055, Guangdong, China.
E-mail address: yongxu@ymail.com (Y. Xu).

https://doi.org/10.1016/j.neunet.2018.02.002
0893-6080/

projection that can pull the samples of same class together and push the samples of different classes far away in the discriminative subspace (Li, Chen, Nie, & Wang, 2017; Wang, Meng, & Li, 2017). Compared with LDA, LR is more flexible and efficient. For example, the sparsity techniques, such as the lasso constraint (l_1 norm) and row-sparsity constraint ($l_{2,1}$ norm) can be simply introduced into the model of LR to improve the interpretability and effectiveness. Introducing the sparsity technique also allows the learned transformation matrix to select the most discriminative features for classification, which is beneficial to improve the performance (Tibshirani, 2011; Xiang, Nie, Meng, Pan, & Zhang, 2012).

However, many issues still exist in the above LR based methods. The first issue is that the target matrix, *i.e.*, zero–one label matrix, is inappropriate for classification (Cai, Ding, Nie, & Huang, 2013; Wang & Pan, 2017; Xiang, Nie et al., 2012; Zhang, Lai et al., 2017; Zhang, Wang, Xiang, & Liu, 2015). For the strict zero–one label matrix, the Euclidean distances of regression responses between samples from different classes are a constant value, *i.e.*, $\sqrt{2}$. This is contrary to the expectation that samples from different classes should be as far as possible after transformation. The second issue is that these LR based methods only focus on fitting the samples to the corresponding labels while ignoring the relationships among samples, which may destroy the underlying structure of data and lead to the overfitting problem (Argyriou, Evgeniou, & Pontil, 2008; Bunea, She, & Wegkamp, 2011; Cai, Ding et al., 2013; Xiang, Zhu, Shen, & Ye, 2012). To solve these problems, many methods have been developed. For example, many researchers proposed to perform the regression on the relaxed label matrix rather than the strict zero–one matrix, in which the most representative works are the discriminative LSR (DLSR) (Xiang, Nie et al., 2012), margin scalable discriminative LSR (MSDLSR) (Wang, Zhang, & Pan, 2016), and retargeted LSR (ReLSR) (Zhang et al., 2015). DLSR introduces the ε-dragging technique to enlarge the distances of regression targets of different classes (Xiang, Nie et al., 2012). Based on DLSR, MSDLSR further imposes a l_1 norm constraint on the dragging matrix to explicitly control the margin of DLSR (Wang et al., 2016). Different from DLSR and MSDLSR, ReLSR does not use ε-dragging technique to relax the label matrix. It directly leans the regression targets from the data by introducing a margin constraint, where the margin between the true and false targets are enforced to be larger than one (Zhang et al., 2015). To emphasize the correlations among samples, the graph regularization term is introduced to the LR, which allows to learn a more compact representations and avoids the overfitting problem (Fang, Xu, Li, Lai, Wong et al., 2017; Xue, Chen, & Yang, 2009). Some researchers also proposed the low-rank linear regression (LRLR) models, in which the rank constraint, *i.e.*, nuclear norm, is imposed on the transformation matrix to explore the correlations among samples (Argyriou et al., 2008; Bunea et al., 2011; Cai, Ding et al., 2013; Xiang, Zhu et al., 2012).

Both the techniques mentioned above are useful and have the potential to improve the classification performance. However, relaxing the label matrix by introducing the ε-dragging technique or margin constraint will also enlarge the distances of the regression responses between samples from the same class, which is harmful to the classification. In this paper, a new relaxed label regression method named inter-class sparsity based discriminative least square regression (ICS_DLSR) is proposed to learn a more discriminative transformation. Different from the above methods, ICS_DLSR aims to preserve the row-sparsity consistency property of samples from the same class such that the distances of regression responses between samples from the same class can be greatly reduced, and thus can obtain a better performance. To this end, a novel inter-class sparsity regularization term is imposed on the transformation. Meanwhile, a sparsity error term with $l_{2,1}$ norm is introduced to the LSR model to relax the strict target label matrix for regression. Several experimental results show that ICS_DLSR can greatly improve the classification accuracies in comparison with the state-of-the-art methods. In brief, the proposed method has the following properties.

(1) The inter-class sparsity constraint is for the first time integrated into the LSR to exploit the relationships among samples. In particular, ICS_DLSR can learn a more compact and discriminative transformation that allows the transformed samples to have a common structure in each class.

(2) ICS_DLSR introduces a sparsity error term with $l_{2,1}$ norm to compensate the regression errors, which is beneficial to learn a more flexible transformation.

The rest of the paper is organized as follows. Section 2 introduces some notations and related works. In Section 3, the formulation and the optimal solution of the proposed method are presented. Then we analyze the proposed method in Section 4. Some experiments are conducted in Section 5 to prove the effectiveness of the proposed method. Section 6 offers the conclusion.

2. Related work

This section briefly introduces some related linear regression methods. For convenience, we first introduce some notations which are used throughout the paper. Let $X = [x_1, x_2, \ldots, x_n] \in R^{m\times n}$ be the training set with n training samples from c classes, where m is the feature dimension of each sample. We use $X_i \in R^{m\times n_i}$ and n_i to denote the sub-training set and the number of samples of the ith class, respectively. For a vector $z = [z_1, z_2, \ldots, z_n]$, its l_2 norm is calculated as $\|z\|_2 = \sqrt{\sum_{i=1}^{n} z_i^2}$. For a matrix $W \in R^{c\times m}$, its l_1-norm, $l_{2,1}$-norm, and 'Frobenius' norm (l_F-norm) are calculated as $\|W\|_1 = \sum_{j=1}^{c}\sum_{i=1}^{m}|W_{ij}|$, $\|W\|_{2,1} = \sum_{i=1}^{c}\sqrt{\sum_{j=1}^{m} W_{ij}^2}$, $\|W\|_F^2 = \sum_{i=1}^{c}\sum_{j=1}^{m} W_{ij}^2$, respectively. $W_{i,j}$ denotes the (i,j)th element of matrix W. W^{-1} denotes the inverse matrix of matrix W. W^T is the transposed matrix of matrix W. We use a zero–one matrix $Y = [y_1, y_2, \ldots, y_n] \in R^{c\times n}$ to represent the label matrix corresponding to the training set X, where each column vector $y_i \in R^{c\times 1}$ is simply defined as follows: if training sample x_i comes from the kth class, then the kth element of column vector y_i is 1 while the remaining elements are 0. I is the identity matrix. Note that, the matrix based features such as image are pre-transformed into the column vector by stacking the matrix columns.

2.1. Standard LR (StLR) and low-rank LR (LRLR)

Given a training set $X \in R^{m\times n}$ and the corresponding label matrix $Y \in R^{c\times n}$, StLR aims at jointly learning a projection that can well transform the given training samples into their respective class labels as follows:

$$\min_{Q} \|Y - QX\|_F^2 + \lambda \|Q\|_F^2 \tag{1}$$

where $Q \in R^{c\times m}$ is the transformation matrix, λ is the regularization parameter with a small positive value. Problem (1) can be easily solved and has a closed solution as $Q = YX^T\left(XX^T + \lambda I\right)^{-1}$. For a test sample z, StLR predicts its label as $k = \mathrm{argmax}_i(Qz)_i$, where $(Qz)_i$ denotes the ith element of vector Qz.

To exploit the correlations reside in the high-dimensional data, LRLR replaces the l_F-norm regularization term with a low-rank constraint as follows:

$$\min_{Q} \|Y - QX\|_F^2 + \lambda \|Q\|_* \tag{2}$$

where $\|Q\|_*$ denotes the nuclear norm (trace norm) of matrix Q and is calculated as the sum of all singular values of matrix Q (Cai, Ding et al., 2013; Zhang, Lai et al., 2017). Compared with StLR, LRLR can discover the low-rank structures of data such that a more discriminative and compact transformation can be learned, and thus has the potential to obtain a better performance.