Pattern Recognition 46 (2013) 1209–1219

Contents lists available at SciVerse ScienceDirect

Pattern Recognition

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

Adaptive weighted learning for linear regression problems via Kullback–Leibler divergence

Zhizheng Liang [a,*], Youfu Li [b], ShiXiong Xia [a]

[a] School of Computer Science and Technology, China University of Mining and Technology, Xuzhou City, China
[b] Department of Manufacturing Engineering and Engineering Management, City University of Hong Kong, Hong Kong

ARTICLE INFO

Article history:
Received 5 April 2012
Received in revised form
28 September 2012
Accepted 24 October 2012
Available online 2 November 2012

Keywords:
Linear regression
KL divergence
Weighted learning
Alternative optimization
Image classification

ABSTRACT

In this paper, we propose adaptive weighted learning for linear regression problems via the Kullback–Leibler (KL) divergence. The alternative optimization method is used to solve the proposed model. Meanwhile, we theoretically demonstrate that the solution of the optimization algorithm converges to a stationary point of the model. In addition, we also fuse global linear regression and class-oriented linear regression and discuss the problem of parameter selection. Experimental results on face and handwritten numerical character databases show that the proposed method is effective for image classification, particularly for the case that the samples in the training and testing set have different characteristics.

1. Introduction

Image representation and classification have gained wide applications in many areas such as pattern recognition, computer vision and machine learning [1–3] during the past many years. One of the important characteristics of images is that the dimensions of images are usually high. If the images are represented in the vector form, they can be considered as the points in a very high dimensional space. However, high dimensional data suffers from the curse of dimensionality, which has a considerable impact on various classification techniques. In order to alleviate this problem, feature extraction techniques [4–6] have been proposed to obtain the effective and meaningful features from the original data. These methods not only reduce the computational complexity of algorithms, but also improve the generalization performance of classifiers.

The feature extraction methods usually search for low-dimensional representations of the original data. Principal component analysis, linear discriminant analysis, and independent component analysis [7] are typical examples of feature extraction methods. Note that classical feature extraction methods may not be robust in dealing with contaminated data. To this end, some robust feature extraction methods [8–15] have been proposed during the past several decades. The robust feature extraction methods can produce better low-dimensional representations of the contaminated data. After the reduced features of data are achieved, one still needs to resort to the classifiers to perform the classification tasks.

Note that when feature extraction regardless of robustness is used as a pre-processed step in the classification problem, a basic assumption is that the samples in the training and testing sets should have similar characteristics. Specifically, low-dimensional representations of the samples in the training set have similar characteristics with those of the samples in the testing set. However, the samples in the training and testing sets may vary. For example, in a real-time face recognition system, the images taken on-line may have a big difference with those images in the database since real-time face images may contain occlusions and are taken from different cameras under different conditions. Consequently, directly performing feature extraction on these new images may be improper and will yield undesirable results in the classification problem.

Different from feature extraction approaches where the features of samples are reduced, there is an increasing interest in the approaches based on the nearest subspace classifier in recent several years [16–18]. Generally this representing model implicitly creates virtual prototypes from the training set and a linear

Abbreviations: (KL), Kullback–Leibler; (GLR), Global linear regression; (COLR), Class-oriented linear regression; (WGLR), Weighted global linear regression; (WCOLR), Weighed class-oriented linear regression; (FLR), The fusion of GLR and COLR; (WFLR), The fusion of WGLR and WCOLR; (NN), Nearest neighbor; (RLR), Robust linear regression; (MC), Maximum correntropy; (SRC), Sparse representation classification

* Corresponding author. Tel.: +86 516 15996966232; fax: +86 516 83995918.
E-mail address: cuhk_liang@yahoo.cn (Z. Liang).

0031-3203/$ - see front matter © 2012 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.patcog.2012.10.017

combination of training samples is usually utilized in real applications. By computing the distance between the testing sample and the virtual prototype produced for each class, the testing sample is assigned to be the class with the smallest distance among all classes. In [17], Wright et al. adopted this idea for probing face representations and proposed a sparse representation-based classified scheme, with the encouraging result reported. Subsequently, a novel method dealing with face misalignments and illuminant variations is proposed in [19]. In order to deal with the occlusions in face images, a sparse correntropy framework [20] for computing robust sparse representations of face images is developed, where the half-quadratic optimization technique is used to maximize the objective function. In practice, exploring the representation of the testing sample in the above methods can be regarded as linear regression problems under proper constraints. The coefficients of combinations are usually controlled by the L1 norm in order to keep sparsity. However, if the dimension of samples is much larger than the number of the samples and there are high correlations between predictors, it has been empirically observed that the prediction performance of the Lasso is dominated by ridge regression [21]. Moreover, several researchers [22,23] also pointed out that the sparse representation on the testing image may not be crucial as the dimension of samples increases. In order to improve classical linear regression, the weighted linear regression methods [24,25] have been proposed for dealing with contaminated data. They work by incorporating extra nonnegative weights, associated with each data point, into the fitting criterion. However, one disadvantage of some weighted linear regression methods is the assumption that the weights are known exactly. This is almost never the case in real applications and the effect of using estimated weights is not easy to assess. It is also noted that some robust regression methods [25] can be solved by iteratively doing weighted least squares.

In this paper, the weight is first regarded as the probability distribution of some random variable. Then we use the KL divergence [26] to measure the difference between the true distribution and the estimated distribution. Thus the weights can be automatically learned by using the KL divergence. As a result, applying our method not only avoids the determination of the weights in advance, but also improves the robustness of linear regression. It is also found that ridge regression is an extreme case of our method. The experimental results show that our method is particularly effective in the case that the samples in the training and testing sets have different characteristics. Overall, the main contributions of this paper include

(1) We propose adaptive weighted learning for linear regression problems based on the KL divergence.
(2) We solve the proposed method using the alternative optimization algorithm and prove the convergence of the algorithm.
(3) We fuse global linear regression and class-oriented linear regression and give the weighted classification rule.
(4) We conduct the experiments on various image data sets to give a comparative analysis on several methods and discuss the criterion for parameter selection.

The rest of this paper is organized as follows: The related work is briefly reviewed in Section 2. In Section 3, we propose adaptive weighted learning for linear regression problems via the KL divergence and use the alternative optimization technique to solve the proposed method. In Section 4, we carry out the experiments on face and handwritten numerical character databases and discuss the problem of parameter selection. Conclusions and further work are given in the final section.

2. Related work

2.1. Linear regression

The classical linear model can be described as follows:

$$\boldsymbol{y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \tag{1}$$

where $\boldsymbol{X}$ is an $m \times n$ matrix, $\boldsymbol{\beta}$ is an n-dimensional vector of unknown parameters, $\boldsymbol{y}$ is an m-dimensional observation vector, and $\boldsymbol{\varepsilon}$ is the error with independent identically distributed components.

In linear regression, the standard estimator of $\boldsymbol{\beta}$ can be obtained by solving the following ordinary least squares problem:

$$\min_{\boldsymbol{\beta}} \ (\boldsymbol{y}-X\boldsymbol{\beta})^T(\boldsymbol{y}-X\boldsymbol{\beta}), \tag{2}$$

where $(\ )^T$ denotes the transpose of a vector. If $\boldsymbol{X}$ is a full-rank matrix, there exists a unique minimizer to Eq. (2), i.e., $\boldsymbol{\beta}^* = \left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T\boldsymbol{y}$. It is also noted that the ordinary least squares method is equivalent to the maximum likelihood estimator if the error in Eq. (1) follows the normal distribution. If the matrix $\boldsymbol{X}^T\boldsymbol{X}$ is singular, there exist infinitely many solutions to Eq. (2). In order to avoid this problem, an effective strategy is to impose a penalty on the size of the coefficients. The popular method is to add the L2 norm penalty to the objective function of Eq. (2). Thus the optimization problem is

$$\min_{\boldsymbol{\beta}} \ (\boldsymbol{y}-X\boldsymbol{\beta})^T(\boldsymbol{y}-X\boldsymbol{\beta}) + \lambda\boldsymbol{\beta}^T\boldsymbol{\beta}, \tag{3}$$

where λ is a regularization parameter. The L2 norm regularized linear regression is also called ridge regression. From Eq. (3), one can obtain the closed form solution for ridge regression, denoted by

$$\hat{\boldsymbol{\beta}} = \left(\boldsymbol{X}^T\boldsymbol{X} + \lambda\boldsymbol{I}_n\right)^{-1}\boldsymbol{X}^T\boldsymbol{y}. \tag{4}$$

Weighted linear regression introduces the additional weights to measure the residuals and its optimization model is denoted by

$$\min_{\boldsymbol{\beta}} \ (\boldsymbol{y}-X\boldsymbol{\beta})^T\boldsymbol{W}(\boldsymbol{y}-X\boldsymbol{\beta}), \tag{5}$$

where $\boldsymbol{W}$ is a diagonal matrix whose diagonal elements are the weights. If the variance of the residual of each observation is known, the weight is set to be inversely proportional to the variance. In fact, the exact variances are never known in real applications and some weighted functions such as Huber, cauchy, bisquare, logistic, and Welsch functions [27] are often adopted.

2.2. Sparse linear regression

In fact, one can replace the L2 norm in Eq. (3) by the general norm. Thus one can obtain the corresponding optimization problem

$$\min_{\boldsymbol{\beta}} (\boldsymbol{y}-X\boldsymbol{\beta})^T(\boldsymbol{y}-X\boldsymbol{\beta}) + \lambda\|\boldsymbol{\beta}\|_p \tag{6}$$

where $\|\boldsymbol{\beta}\|_p$ denotes the Lp norm. In fact $\|\boldsymbol{\beta}\|_p$ is not a true norm in a strict sense when $p < 1$. When $p \le 1$, solving Eq. (6) obtains the sparse solution of $\boldsymbol{\beta}$. The optimization problem of Eq. (6) based on the L0 norm is NP-hard. Instead, the L1 norm is often used to obtain the sparse solution of $\boldsymbol{\beta}$. Now a number of algorithms for solving L1 norm linear regression [28] have been proposed, including proximal gradient, augmented Lagrange multiplier and Homotopy.