Bayesian sample steered discriminative regression for biometric image classification

Guangwei Gao [a], Jian Yang [b], Songsong Wu [c], Xiaoyuan Jing [c], Dong Yue [a],*

[a] Institute of Advanced Technology, Nanjing University of Posts and Telecommunications, Nanjing 210023, PR China
[b] School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing 210094, PR China
[c] School of Automation, Nanjing University of Posts and Telecommunications, Nanjing 210023, PR China

ABSTRACT

Regression techniques, such as ridge regression (RR) and logistic regression (LR), have been widely used in supervised learning for pattern classification. However, these methods mainly exploit the class label information for linear mapping function learning. They will become less effective when the number of training samples per class is small. In visual classification tasks such as face recognition, the appearance of the training sample images also conveys important discriminative information. This paper proposes a novel regression based classification model, namely Bayesian sample steered discriminative regression (BSDR), which simultaneously exploits the sample class label and the sample appearance for linear mapping function learning by virtue of the Bayesian formula. BSDR learns a linear mapping for each class to extract the image class label features, and classification can be simply done by nearest neighbor classifier. The proposed BSDR method has advantages such as small number of mappings, insensitiveness to input feature dimensionality and robustness to small sample size. Extensive experiments on several biometric databases also demonstrate the promising classification performance of our method.

© 2015 Elsevier B.V. All rights reserved.

1. Introduction

Discriminant feature extraction [1,2] is a critical step in many multi-class visual classification tasks, such as face recognition [3], etc. How to extract low dimensional discriminative features from high dimensional visual data (e.g., images and videos) is becoming a hot topic in computer vision, pattern recognition and machine learning areas. In addition, it is highly desired that the feature extraction process can be linear, for example, via linear mappings (or projections). This can make the feature extraction process very efficient and practical to use.

Many subspace learning and manifold learning methods have been proposed to achieve the above goals, such as the well-known principal component analysis (PCA) [4–8], linear discriminant analysis (LDA) [9–12], unsupervised manifold learning [13,14], supervised manifold learning [15–21], etc. However, the performance of these methods will degrade much when the number of training samples per class is small. For example, LDA will not work if each class has only one training sample. On the other hand, these methods learn projections mainly based on the spatial distribution

of training samples. They do not exploit effectively the appearance information of samples in the original high dimensional space, resulting in limited classification performance.

The nearest neighbor (NN) classifier is most widely used in couple with the aforementioned dimensionality reduction methods for visual classification tasks. In case that the number of training samples can be very limited, the nearest feature line [22,23], nearest feature plane [24], and nearest subspace (NS) [25] classifiers were proposed to use two, three or all the training samples in each class to compute the distance from the query sample to each class. Recently, the sparse representation based classification (SRC) [26] has been proposed as a new classifier to code the query sample using the training samples from all classes via l_1-norm minimization. Then, SRC uses the coding coefficients associated with each class to calculate the distance from the query sample to each class. Many following works of SRC have been reported, for example, robust sparse coding [27], class l_1-optimizer classifier [28], half-quadratic (HQ) framework [29] and fisher discrimination dictionary learning [30], etc. While most of these works owe the success of SRC and its variants to the role of l_1-norm minimization, in [31] it is argued that the collaborative representation of query sample by all the training samples across classes is more crucial. By using the non-sparse l_2-norm to regularize the coding process, the authors [31] presented a collaborative representation based classification

* Corresponding author. Tel.: +86 13814188988.
E-mail address: medongy@vip.163.com (D. Yue).

(CRC) model which could lead to similar face recognition results to SRC but with significantly less computational complexity. Up to now, many following works of CRC have been reported, such as kernel collaborative representation [32], multiple collaborative representations [33], image set based collaborative representation [34], etc.

Both SRC and CRC, however, have two problems. First, the number of training samples per class cannot be too small. Otherwise, SRC/CRC's performance will deteriorate because the distance between the query sample and each class cannot be stably computed in such cases. Second, SRC and CRC are representation based classifiers, so that the dimensionality of the sample feature cannot be too low. Otherwise, their performance will drop because the coding process will become less stable with low dimensional features. Linear regression techniques have been applied to overcome the difficulties in pattern classification with a small training sample size as well as low feature dimensionality. For example, by using ridge regression the authors of [35,36] learned a projection matrix to map the training samples onto the vertices of a regular c-simplex, which represents the labels of different classes. The logistic regression [37–39] has been widely used by researchers to solve linear classification problems, where the logistic function is utilized to model the posterior probability of training samples to predict the discrete binary response.

Most of these regression methods were not originally designed for visual data classification, and hence they ignore the appearance of training samples and mainly employ the class label to guide the learning of linear mapping functions. They will become less effective when the number of training samples per class is small and the number of classes is big. Considering that in visual classification tasks the appearance of training samples also conveys important information for pattern discrimination, this paper proposes to incorporate the appearance and the class label of training samples into linear mapping function learning process. Fortunately, it can be found that the class label and the appearance of each sample could be naturally and simultaneously exploited based on the Bayesian formula. Thus a Bayesian sample steered discriminative regression (BSDR) model can be developed for pattern classification. The proposed BSDR model learns a linear discriminant mapping for each class, and performs classification by the simple NN classifier with cosine distance. It has the following advantages. First, it offers a small number of linear mappings (equals to the number of classes), leading to small feature storage space and fast feature extraction speed. Second, it is insensitive to the dimensionality of the input feature, and can achieve good classification rates at both high and low dimensional compared to other methods. Third, it works well with small sample size, achieving higher recognition rates than other state-of-the-art methods.

The rest of the paper is organized as follows. Section 2 briefly reviews representative regression based classification models. Section 3 presents the proposed BSDR method. Section 4 conducts extensive experiments and Section 5 concludes the paper.

2. Regression for classification

This section briefly reviews some representative linear regression models for classification. Suppose that there are K classes of subjects. Let $X = [x_1, x_2, \ldots, x_N] \in \Re^{d \times N}$ be the dataset of all training samples, where d is the dimensionality of the sample feature, N is the total number of training samples.

2.1. Ridge regression (RR)

Ridge regression (RR) is generalized for face recognition in [35,36]. RR uses regular simplex vertices to represent the multiple target class labels. Let $T_k \in \Re^{K-1}$ ($k = 1, 2, \ldots, K$) be the vertices of one regular K-simplex and let $T = [T_1, T_2, \ldots, T_K]$. RR constructs T as follows:

1. Let $T_1 = [1, 0, \ldots, 0]^T$ and $T_{1,i} = -1/(K-1)$, $i = 2, \ldots, K$.
2. For $1 \le k \le K - 2$,

$$T_{k+1,k+1} = \sqrt{1 - \sum_{i=1}^{k} T_{i,k}^2}$$

$$T_{k+1,j} = -\frac{T_{k+1,k+1}}{K-k-1}, \quad j = k+2, \ldots, K \qquad (1)$$

$$T_{i,k+1} = 0, \quad k+1 < i \le K-1$$

RR treats the classification problem as a ridge regression problem, and aims to find a matrix W that can map x_i to label Y_i, where $Y_i = T_j$ if x_i belongs to the jth individual, $i = 1, 2, \ldots, N, j = 1, 2, \ldots, K$. W can be obtained by

$$W_{RR} = (XX^T + \lambda \cdot I)^{-1} XY^T \qquad (2)$$

where $Y = [Y_1, Y_2, \ldots, Y_N]$.

For a test image y, RR compares the distance $W^T y$ and the individual target T_i and identifies x as the class with the minimal distance.

2.2. Logistic regression (LR)

Logistic regression has been successful used in pattern recognition tasks [37–39].

Suppose we have N independent observations $(y_1, x_1), \ldots, (y_N, x_N)$, where y_i is a binary variable with probability $p_i = P(y_i = 1)$ and $x_i \in \Re^d$ are some explanatory variables. Fitting a linear regression model for a binary response can give predicted values beyond $(0, 1)$ that are theoretically inadmissible. Logistic regression models are widely used to model binary response using the following formulation:

$$\log \frac{p_i}{1 - p_i} = w^T x_i \qquad (3)$$

where $w \in \Re^d$ are unknown regression coefficients and is often estimated using maximum likelihood estimation (MLEs):

$$L(w) = \prod_{i=1}^{n} p_i^{y_i} (1 - p_i)^{1-y_i} \qquad (4)$$

Eq. (3) can be rewritten as

$$p_i = \frac{\exp(w^T x_i)}{1 + \exp(w^T x_i)} \qquad (5)$$

Thus, the likelihood function (4) is a function of w.

Maximizing $L(w)$ is equivalent to maximizing the log-likelihood function:

$$l(w) = \sum_{i=1}^{n} [y_i x^T x_i + \log(1 - \mu(x^T x_i))] \qquad (6)$$

where μ represents sigmoid function.

According to the optimality condition of unconstrained optimization, the parameter that maximizes $l(w)$ is the solution to the following likelihood equation:

$$l'(w) = \frac{\partial l(w)}{\partial w} = \sum_{i=1}^{n} [y_i - \mu(x^T x_i)] x_i = 0 \qquad (7)$$

In general, the Newton–Raphson method is used to compute the optimal w.