Pattern Recognition 60 (2016) 613–629

Contents lists available at ScienceDirect

Pattern Recognition

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

Cost-sensitive dictionary learning for face recognition

Guoqing Zhang [a,*], Huaijiang Sun [a,*], Zexuan Ji [a], Yun-Hao Yuan [b], Quansen Sun [a]

[a] School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing, Jiangsu 210094, China
[b] School of Internet of Things, Jiangnan University, Wuxi 214122, China

ARTICLE INFO

Article history:
Received 19 June 2014
Received in revised form
7 April 2016
Accepted 17 June 2016
Available online 18 June 2016

Keywords:
Cost sensitive
Sparse representation
Dictionary learning
Face recognition

ABSTRACT

As one of the most popular research topics, sparse representation and dictionary learning technique has received an increasing amount of interest in recent years. Sparse representation based classification (SRC) has been shown to be an effective method and produce impressive performance on face recognition. SRC directly used the entire set of training samples as the dictionary for sparse coding. Recent research has shown that learning a dictionary from the training samples instead of using a predefined one can produce state-of-the-art results. However, all of these dictionary learning methods are designed to achieve low classification errors and implicitly assumes that the losses of all misclassification are the same. In many real-world face recognition applications, this assumption may not hold as different misclassifications could lead to different losses. Motivated by this concern, in this paper we propose a cost-sensitive dictionary learning algorithm for SRC, by which the designed dictionary is able to produce cost-sensitive sparse coding, resulting in improved classification performance in such scenarios. Our method considers the cost information during the sparse coding stages. Specifically, we introduce a new "cost" penalizing matrix and enforce the cost-sensitive requirement throughout the learning process. The optimal solution is efficiently obtained following the alternative optimization method. Experimental results demonstrate the effectiveness of the proposed method.

1. Introduction

Face recognition is a challenging computer vision task that has been studied over 30 years [1].Many successful face recognition systems have been developed, such as Eigenfaces based on Principal component analysis (PCA) [2], Fisherfaces based on Linear Discriminate Analysis (LDA) [3] and Laplacianfaces using locality preserving projection (LPP) [4]. Those methods usually involve two stages: feature extraction and classification. Recently, sparse representation technique has been applied in a variety of applications in computer vision and pattern recognition [5–9]. Wright et al. [10] developed a sparse representation based classification (SRC) and obtained promising results on face recognition. In SRC, a testing image was encoded by sparse linear combination of all the training samples and classified into the class with minimum sparse reconstruction residual. Unlike conventional methods such as Eigenfaces and Fisherfaces, SRC does not need an explicit feature extraction stage.

As well known, Wright et al. [10] directly employ entire training samples as a dictionary for discriminative sparse coding. However, it has been demonstrated that learning a dictionary from original training samples instead of using a predefined one such as wavelets [11], can lead to much better results [12–22]. In [14], a dictionary learning algorithm, K-SVD, is introduced that generalizes k-means clustering and efficiently learns an overcomplete dictionary from a set of training samples. Lee et al. [15] proposed an efficient reconstructive dictionary learning method which shows promising results in self-taught learning tasks [16] and image categorization [17]. Yang et al. [18] proposed a metaface learning (MFL) algorithm to represent training samples by a series of "metafaces" learned from each class. Based on K-SVD, Zhang et al. [19] developed D-KSVD algorithm by simultaneously learning a linear classifier. Jiang et al. [20,21] introduced a label consistent regularization to enforce the discrimination of coding vectors. The so-called LC-KSVD algorithm exhibits good classification results. Yang et al. [22] proposed a Fisher discrimination dictionary learning method, where the category-specific strategy is adopted for learning a structural dictionary and the Fisher discrimination criterion is imposed on the coding vectors to enhance class discrimination. Moreover, there are many other efforts which have been devoted to the learning of a proper dictionary for particular applications, i.e. image denoising [23,24], image inpainting [25], and image classification [26–30].

However, current sparse representation and dictionary learning based methods only target at low recognition errors and implicitly

* Corresponding authors.
E-mail addresses: xiayang14551@163.com (G. Zhang), sunhuaijiang@njust.edu.cn (H. Sun).

http://dx.doi.org/10.1016/j.patcog.2016.06.012
0031-3203/

assume that the losses of all misclassification are the same. Although this assumption is widely taken, we argue that it is not really reasonable because, for most real-world applications, different kinds of mistakes generally lead to different amounts of losses. For example, consider a door locker based on a face recognition system for a certain group (e.g., family members or company employees); it may cause inconvenience to a gallery person who is misrecognized as an impostor and not allowed to enter the room, but may result in a serious loss or damage if an impostor is misrecognized as a gallery person and allowed to enter the room. From the example above, clearly we can conclude that face recognition is a cost-sensitive pattern classification problem, which has been neglected by most existing face recognition algorithm.

Cost-sensitive learning is one important topic in the data mining and machine learning community [31–35]. In such settings, cost information is introduced to measure the importance of different samples in different classes, and different costs reflect different amounts of losses. The purpose of cost-sensitive learning is to minimize the total cost rather than total error. Generally, there are two kinds of misclassification cost. The first is class-dependent, where the costs of misclassifying any example in class A to class B are the same. The second is example-dependent, where the costs of classifying examples in class A to class B are different. In this paper, we focus on the former one because face recognition is generally a class-dependent cost-sensitive problem.

There have been several cost-sensitive learning algorithms proposed in the literature. Such as cost-sensitive boosting [32,34], cost-sensitive SVM [31], cost-sensitive semi-supervised learning [31], and cost-sensitive neural networks [33]. Zhang et al. [35] presented a multiclass cost-sensitive learning framework for face recognition which aims at minimizing the total loss of misclassifications instead of classification errors. Lu et al. [36,37] introduced the cost information into four popular and widely used linear subspace learning algorithms and devised the corresponding cost-sensitive methods, namely CSPCA, CSLDA, CSLPP and CSMFA. Lu et al. [38] proposed a cost-sensitive semi-supervised discriminant analysis method by making use of both labeled and unlabeled sample and exploring different cost information of all the training samples simultaneously. To our best knowledge, [39] should be the first work that formally introduced the cost-sensitive idea into sparse representation and presented a sparse cost-sensitive classifier (SCS-C), SCS-C utilizes the probabilistic model of sparse representation to estimate the posterior probabilities of a testing samples, and calculates all the misclassification losses via the posterior probabilities. Finally, the testing sample is assigned to the class with minimal loss. Note that, SCS-CS uses all the training samples to form a dictionary for sparse representation. However, original training samples may contain some redundant information, noise or even other trivial information that obstructs the correct recognition. Intuitively, a more accurate and discriminative representation can be obtained if we could optimize a dictionary from original training samples. Motivated by the above concerns, in this paper we propose a novel cost-sensitive dictionary learning approach for sparse representation based classification. Our method considers the cost information during sparse coding stage. We introduce a new "cost" penalizing matrix and enforce the cost-sensitive constraint throughout the learning process. The learned dictionary which is able to produce cost-sensitive sparse coding and encourages the samples from the same class to have similar sparse codes and those from different classes to have dissimilar sparse codes. To our best knowledge, our work is the first attempt to introduce cost information into dictionary learning technique.

The rest of this paper is organized as follows: Section 2, reviews related works on sparse representation and dictionary learning. In Section 3, we formulate the cost-sensitive learning for face recognition, and then we present our proposed algorithm for cost-sensitive dictionary learning in Section 4. Experimental results on real image datasets are given in Section 5. Section 6 concludes the paper.

2. Related work

In this section, we review briefly some related works on sparse representation based classification and dictionary learning.

2.1. Sparse representation based classification

The SRC scenario proposed by Wright et al. [10] uses sparse representation for robust face recognition. Suppose that there are n training images from c object classes, and each class has n_i images. Let $X = [X_1, X_2, \cdots, X_c] \in R^{m\times n}$ be the training data matrix, where $X_i = \left[x_{i1}, x_{i2}, \cdots, x_{in_i}\right] \in R^{m\times n_i}$ is the training samples of the i–th class, and $n = \sum_{i=1}^{c} n_i$. Given a testing image $y \in R^m$, we represent y in an overcomplete dictionary whose atoms are training samples themselves, i.e., $y = X\alpha$, where $\alpha = [\alpha_1; \alpha_2; \cdots; \alpha_c]$ is the coding vector and $\alpha_i \in R^{n_i}$ be the entries of α associated with class i. α can be sought by solving the following optimization problem

$$\hat{\alpha} = \arg\min_{\alpha} \|\alpha\|_0 \quad s.\ t. \quad y = X\alpha \tag{1}$$

where $\|\|_0$ is the ℓ_0-norm, which counts the number of nonzero entries in a vector.

Solving ℓ_0 optimization problem in (1), however, is NP-hard problem and extremely time-consuming. Fortunately, recent research efforts reveal that for certain dictionaries, if the solution $\hat{\alpha}$ is sparse enough, finding the solution of the ℓ_0 optimization problem is equivalent to find the solution to the following ℓ_1 optimization problem [40,41]

$$\hat{\alpha} = \arg\min_{\alpha} \|\alpha\|_1 \quad s.\ t. \quad y = X\alpha \tag{2}$$

Using the Lagrangian method, Eq. (2) can be transformed to the following unconstrained optimization problem:

$$\hat{\alpha} = \arg\min_{\alpha} \|x - X\alpha\|_2^2 + \lambda\|\alpha\|_1 \tag{3}$$

where regularization parameter λ controls the sparsity of α. Currently, there are a large number of algorithms using different techniques to solve (3) in recent literature. More details of those algorithms can be found in [40,42].

Once (3) is solved, the classification can be performed using minimum class-wise reconstruction error. We recognize the test image y as i^* according to the following rule:

$$i^* = \arg\min_{i} \|y - X_i\alpha_i\|_2^2 \tag{4}$$

The underlying assumption behind SRC is that, if the testing image y belongs to class i, it should be presented in the column space of X_i. Therefore, the non-zero elements of α will mainly be observed in α_i and thus satisfy (4).

2.2. Dictionary learning

In SRC, considering original training samples as a dictionary to represent the testing sample. However, intuitively, a more accurate and discriminative representation can be obtained if we could optimize a dictionary from original training samples. We can learn a dictionary $D = [d_1, \cdots, d_K] \in R^{m\times K}$ by solving the following optimization problem: