Computer Vision and Image Understanding 121 (2014) 86–99

Contents lists available at ScienceDirect

Computer Vision and Image Understanding

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

Soft label based Linear Discriminant Analysis for image recognition and retrieval ☆

Mingbo Zhao [a], Zhao Zhang [b], Tommy W.S. Chow [a], Bing Li [c,*]

[a] Department of Electronic Engineering, City University of Hong Kong, Tat Chee Avenue, Kowloon, Hong Kong Special Administrative Region
[b] School of Computer Science and Technology, Soochow University, PR China
[c] School of Economics, Wuhan University of Technology, PR China

ARTICLE INFO

Article history:
Received 4 January 2013
Accepted 24 January 2014
Available online 5 February 2014

Keywords:
Linear Discriminant Analysis
Semi-supervised dimensionality reduction
Soft label
Label propagation

ABSTRACT

Dealing with high-dimensional data has always been a major problem in the research of pattern recognition and machine learning. Among all the dimensionality reduction techniques, Linear Discriminant Analysis (LDA) is one of the most popular methods that have been widely used in many classification applications. But LDA can only utilize labeled samples while neglect the unlabeled samples, which are abundant and can be easily obtained in the real world. In this paper, we propose a new dimensionality reduction method by using unlabeled samples to enhance the performance of LDA. The new method first propagates the label information from labeled set to unlabeled set via a label propagation process, where the predicted labels of unlabeled samples, called soft labels, can be obtained. It then incorporates the soft labels into the construction of scatter matrixes to find a transformed matrix for dimensionality reduction. In this way, the proposed method can preserve more discriminative information, which is preferable when solving the classification problem. Extensive simulations are conducted on several datasets and the results show the effectiveness of the proposed method.

1. Introduction

Dealing with high-dimensional data has always been a major problem in the research of pattern recognition and machine learning. Typical applications of these include face recognition, document categorization, and image retrieval. Finding a low-dimensional representation of high-dimensional space, namely dimensionality reduction is thus of great practical importance. The goal of dimensionality reduction is to reduce the complexity of input space and to embed high-dimensional space into a low-dimensional space while keeping most of the desired intrinsic information [1,2,13,14]. Among all the dimensionality reduction techniques, Principle Component Analysis (PCA) [3] and Linear Discriminant Analysis [4] are the most popular methods and have been widely used in many classification applications. The objective of PCA is to pursue a set of orthogonal basis that can capture the directions of maximum variance in the dataset for optimal reconstruction. While the objective of LDA is to find the optimal projection that maximizes the between-class scatter matrix S_b while minimizes the within-class scatter matrix S_w in the low-dimensional subspace. Given that the within-class scatter matrix is nonsingular, the optimization problem of LDA can be solved by generalized eigen-value decomposition (GEVD), i.e. to find the d largest eigenvectors corresponding to the eigenvalues of $S_w^{-1}S_b$ [27]. However, for many applications where the number of dimensionality is much larger than that of the samples, the within-class scatter matrix tends to be singular. Hence the optimal projection matrix may be found incorrect. This is the so-called small sample problem and many variants of LDA have been proposed to solve it, which include Regularized LDA [23], Null-space LDA [24], Uncorrelated LDA [25], SRDA [26].

In general, LDA is supervised, which means it requires label information. Although supervised methods generally outperform unsupervised methods, a large number of labeled samples are needed in order to achieve satisfactory results [27]. But in many cases, labeling large number of samples is time-consuming and costly. On the other hand, unlabeled samples are abundant and can be easily obtained in the real world. Thus, semi-supervised learning methods [5–12,41–45], which incorporate both labeled and unlabeled samples into learning procedure, have become more effective than only relying on supervised learning. Many semi-supervised methods have been proposed in the past few years, which include Gaussian Fields and Harmonic Functions (GFHF) [5], Learning with Local and Global Consistency (LLGC) [6],

☆ This paper has been recommended for acceptance by Chung-Sheng Li.

* Corresponding author.
E-mail addresses: mzhao4@cityu.edu.hk, zmbman@hotmail.com (M. Zhao), cszzhang@gmail.com (Z. Zhang), eetchow@cityu.edu.hk (T.W.S. Chow), lib675@163.com (B. Li).

http://dx.doi.org/10.1016/j.cviu.2014.01.008
1077-3142/

Semi-supervised Discriminant Analysis (SDA) [9] and Laplacian regularized Least Square (LapRLS) [11].

The above methods can be divided into two categories: transductive method and inductive method. Two well-known transductive methods are GFHF and LLGC. These methods work in a transductive way by propagating the label information from labeled set to unlabeled set. But they cannot predict the class labels of new-coming samples hence suffering out-of-sample problem. In contrast, inductive methods, such as SDA and LapRLS can solve such problem. These methods firstly construct a manifold regularized term to preserve the geometrical structure with both labeled and unlabeled set [17]. Then, they arm to find a transformed matrix to perform dimensionality reduction by incorporating the manifold regularized term into the original objective function of supervised algorithms. Hence, the new-coming samples can be projected into a low-dimensional subspace by using such transformed matrix and the out-of-sample problem can be naturally solved, which is more practical in real-world applications.

In this paper, we propose a new semi-supervised dimensionality reduction method to enhance conventional LDA performance by incorporating the soft labels into the scatter matrixes. The proposed method first propagates the label information from labeled set to unlabeled set via label propagation, where the predicted class labels of unlabeled samples, called soft labels, can be obtained. It then finds a transformed matrix to perform dimensionality reduction by incorporating the soft labels into the scatter matrixes. The proposed method can be viewed as a unified framework to extend many existing LDA algorithms to their semi-supervised versions. Its basic ideas are different from those semi-supervised algorithms such as SDA and LapRLS. These methods use the labeled and unlabeled samples in a simple manner to construct a manifold regularized term and add it to the objective function of supervised algorithms. But in the proposed algorithm, by incorporating the soft labels into training, it can well preserve the probability distribution of samples both in labeled and unlabeled set hence obtaining a better subspace for dimensionality reduction. We also analyze our proposed algorithm under a least square framework. It can be concluded that given a certain class indicator, the optimization problem of the proposed method can be equivalent to a weighted least square problem.

The main contribution of this paper is summarized as follows:

(1) We present an effective label propagation procedure, which is based on a new local reconstruction graph with symmetrization and normalization. The symmetrization can add more connections of a sample with its neighborhoods and the normalization can handle with the case when the density of dataset varies dramatically.
(2) The proposed SL-LDA can preserve more discriminative information in the soft labels of unlabeled samples than other methods, which is good to the performance for classification. It can also be easily extended to its kernel version by using kernel tricks [21,22].
(3) Motivated by the equivalence between LDA and the least square problem [18–20], we extend this relationship and further analyze SL-LDA under a weighted least square problem (W-LS). Based on such relationship, we then propose a more efficient approach for calculating the optimal solution of SL-LDA, which is a linear system of equation and can be performed on large-scale dataset.

This paper is organized as follows: In Section 2, we will present an effective label propagation procedure with outlier detection and dealing with noisy labels. In Section 3, we will introduce our soft label based Linear Discriminant Analysis (SL-LDA) for semi-supervised dimensionality reduction. We will also build a close relationship between SL-LDA and W-LS in this section and propose a more efficient approach for solving SL-LDA. The simulation results based on extensive datasets are shown in Section 5 and the final conclusions are drawn in Section 6.

2. Transductive learning via label propagation

Let $X = [X_l, X_u] \in R^{D\times(l+u)}$ be the labeled and unlabeled set, where each sample in X_l is associated with a class label c_i, $i \in [1, 2, \ldots, c]$. The goal of label propagation is to propagate the label information of labeled set to the unlabeled set according to the distribution associated with both labeled and unlabeled set [5–8,39,46], through which the predicted labels of unlabeled set, called soft labels can be obtained.

2.1. Graph construction

In label propagation, a similarity matrix must be defined for evaluating the similarities between any two samples. The similarity matrix can be approximated by a neighborhood graph associated with weights on the edges. Officially, let $\widehat{G} = (\widehat{V}, \widehat{E})$ denote this graph, where $\widehat{V}$ is the vertex set of $\widehat{G}$ representing the training samples, $\widehat{E}$ is the edge set of $\widehat{G}$ associated with a weight matrix containing the local information between two nearby samples. There are many strategies to define the weight matrix W, a typical one is to use Gaussian function as [5,6,39]:

$$w_{ij} = \exp(-\|x_i - x_j\|^2/\sigma) \quad x_i \in N_k(x_j) \text{ or } x_j \in N_k(x_i), \tag{1}$$

where $N_k(x_j)$ is the k neighborhood set of x_j, σ is the Gaussian function variance. However, σ is hard to be determined and even a small variation of σ can make the results dramatically different [7]. Wang and Zhang have proposed another strategy to construct $\widehat{G}$ by using the neighborhood information of samples [7]. This strategy assumes that each sample can be reconstructed by a linear combination of its neighborhoods [2], i.e. $x_i \approx \sum_{j:x_j\in N_k(x_i)} w_{ij}x_j$. It then calculates the weight matrix by solving a standard quadratic programming (QP) problem as:

$$\min \left\| x_i - \sum_{j:x_j\in N_k(x_i)} w_{ij}x_j \right\|_F^2 \quad \text{s.t. } w_{ij} \geqslant 0, \quad \sum_{j\in N_k(x_i)} w_{ij} = 1. \tag{2}$$

The above strategy is empirically better than the Gaussian function, as the weight matrix can be automatically calculated in a closed form once the neighborhood size is fixed. In addition, in order to make a connection to the normalized graph, we symmetrize W as $W \leftarrow (W + W^T)/2$ or $w_{ij} \leftarrow (w_{ij} + w_{ji})/2$. The advantage of this step is that it considers the node degree of each sample and a sample with large node degree can connect more neighborhoods. Then similar to [39], we normalize W as $\widetilde{W} = D^{-1/2}WD^{-1/2}$ or $\tilde{w}_{ij} = w_{ij}/\sqrt{d_{ii}d_{jj}}$ where D is the diagonal matrix satisfying $d_{ii} = \sum_{j=1}^{l+u} w_{ij}$. The normalization can strengthen the weights in the low-density region and weaken the weights in the high density region, which is good for handling the case that the density of dataset varies dramtically. Finally, to satisfy the sum-to-one constraint as Eq. (2), the weight matrix $\widetilde{W}$ is set as $\widetilde{W} \leftarrow \widetilde{W}\widetilde{D}^{-1}$ or $\tilde{w}_{ij} \leftarrow \tilde{w}_{ij}\big/\sum_{j=1}^{l+u}\tilde{w}_{ij}$, where $\widehat{D}$ is the diagonal matrix satisfying $\tilde{d}_{ii} = \sum_{j=1}^{l+u}\tilde{w}_{ij}$. The basic steps for graph construction can be seen in Table 1.

2.2. Label propagation process

We will then predict the labels of unlabeled samples based on a label propagation process, through which the soft labels of unlabeled set can be obtained. Let $Y = [y_1, y_2, \ldots, y_{l+u}] \in R^{(c+1)\times(l+u)}$ be