J. Vis. Commun. Image R. 38 (2016) 11–17

Contents lists available at ScienceDirect

J. Vis. Commun. Image R.

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

Spare L1-norm-based maximum margin criterion ☆

Gui-Fu Lu [a,b,*], Ganyi Tang [a], Jian Zou [a]

[a] *School of Computer and Information, AnHui Polytechnic University, WuHu, AnHui 241000, China*
[b] *Key Laboratory of Intelligent Perception and Systems for High-Dimensional Information, Ministry of Education, Nanjing University of Science and Technology, Nanjing 210094, China*

ARTICLE INFO

Article history:
Received 10 October 2015
Accepted 5 February 2016
Available online 11 February 2016

Keywords:
Feature extraction
Linear discriminant analysis
L1-norm
L2-norm
Maximum margin criterion
Elastic net
Sparse subspace
Face recognition

ABSTRACT

Maximum margin criterion (MMC) is a popular method for dimensionality reduction or feature extraction. MMC can alleviate the small size sample (SSS) problem encountered by linear discriminant analysis (LDA) and extract more discriminant vectors than LDA. However, the objective function of MMC is derived from L2-norm, which makes MMC be sensitive to noise and outliers. Besides, the basis vectors of MMC are dense, which makes it hard to explain the obtained features. To address the drawbacks of MMC, in this paper, we propose a novel sparse L1-norm-based maximum margin criterion (SMMC-L1). L1-norm rather than L2-norm is used in the objective function of SMMC-L1. Besides, L1-norm is also used as a lasso penalty to regularize the basis vectors. An iterative algorithm for solving SMMC-L1 is proposed. Experiment results on some databases show the effectiveness of the proposed SMMC-L1.

1. Introduction

Dimensionality reduction plays a core role in many machine learning and pattern recognition problems [1]. In the past decade years, a lot of dimensionality reduction methods have been proposed in the literatures [2]. Among the various methods, principal component analysis (PCA) [1,3] and linear discriminant analysis (LDA) [1,3,4] are the two most famous ones.

PCA is an unsupervised dimensionality reduction approach which aims to find a representative projection transformation matrix such that the variance of the given data is maximized. On the contrary, LDA is a supervised dimensionality reduction approach which aims to find a discriminative projection transformation matrix on which the between-class distance is maximized and meanwhile the within-class distance is minimized. For classification problem, it is generally believed that LDA can obtain better classification performances than PCA.

LDA, however, suffers from the so-called small sample size (SSS) problem or undersampled problem when the number of samples is smaller than the dimensionality of samples. To address this problem, many extensions to LDA, e.g., Fisherfaces [4], null space LDA [5], complete LDA [6] and maximum margin criterion (MMC) [7–10], etc., have been developed in the recent years.

Among these LDA extensions, MMC is an effective one. Different from LDA, which uses the generalized Rayleigh quotient as the discriminant criterion, MMC uses the difference of the between-class scatter and within-class scatter as the discriminant criterion. Then MMC can alleviate the SSS problem since it does not compute the inverse matrix. Besides, MMC can obtain more projection vectors than LDA. Motivated by the large-margin principle, some other large-margin based learning method have been proposed, e.g. the large-margin based weakly supervised dimensionality reduction method [11] which integrates two aspects of the large principle (angle and distance), the large-margin multiview information bottleneck (LMIB) algorithm [12], and the large-margin multi-label causal feature learning method (LMCF) [13].

The traditional PCA, LDA and MMC methods ignore the possible nonlinearity inherent in data. The manifold learning algorithms, however, can discover the underlying manifold structure hidden in the data space. Many manifold-based learning methods and their extensions have been proposed, e.g. Isomap [14], Laplacian Eigenmap [15], Hessian regularized support vector machines (SVM) [16], and multiview Hessian regularized logistic regression (mHLR) [17].

A common property of aforementioned approach is that all these methods are derived from L2-norm. L2-norm based dimensionality reduction approaches, however, are sensitive to noised and outliers since the square operation in L2-norm will magnified the effect of noise and outliers [18]. In [19,20], Liu and Tao pointed out that L1-norm based methods gives a small weight to a large

☆ This paper has been recommended for acceptance by M.T. Sun.

* Corresponding author at: School of Computer and Information, AnHui Polytechnic University, WuHu, AnHui 241000, China.

E-mail address: luguifu_tougao@163.com (G.-F. Lu).

http://dx.doi.org/10.1016/j.jvcir.2016.02.004
1047-3203/© 2016 Elsevier Inc. All rights reserved.

error training sample and a large weight to a small error training sample during the optimization procedures. Generally, L1-norm is much more robust to outliers than L2-norm. Then, researchers utilize L1-norm instead of L2-norm to develop robust dimensionality reduction methods [18,21–30]. For example, there are some L1-norm-based PCA approaches, which include L1-PCA [23], R1-PCA [22], PCA-L1 [18,25], 2DPCA-L1 [21] and L1-norm-based tensor PCA (TPCA-L1) (TPCA-L1), etc., have been developed in the literature. Similarly, some L1-norm-based LDA [26–30] are also proposed in recent years. These L1-norm-based dimensionality reduction methods have demonstrated encouraging performances on some data sets. In some situations, sample labels rather than samples are corrupted by noise. In [31], Liu and Tao discussed the classification problem where sample labels are randomly corrupted, and addressed two fundamental problems in the scenario. Xu et al. [32] pointed that Cauchy loss function is also robust to outliers and advantageous over least squared loss function and least absolute function.

The projection vectors learned by the above methods, however, are still dense and then it is difficult to explain the obtained features. To address the issue, sparse methods have been received increasing attention and many sparse dimensionality reduction approaches have been proposed in recent years [33]. By first reformulating the conventional PCA as a regression optimization problem and then using the elastic net to penalize the basis vectors, Zou et al. [34] proposed the sparse PCA (SPCA) method. The structured sparse PCA, which is a generalization of SPCA, is further proposed by Jenatton et al. [35]. Liu et al. [36] proposed an efficient and paralleled method of SPCA using graphics processing units (GPUs), which can process large blocks of data in parallel. By using the graph embedding framework [37] and spectral regression, Cai et al. [38,39] proposed a unified sparse subspace learning (USSL) framework which first cast various dimensionality reduction methods into regression problem and then use L1-norm to regularize the basis vectors. Similarly, by using patch alignment technique, Tao et al. [40] also proposed a unified sparse dimensionality reduction framework called manifold elastic net. By using the structured sparsity-inducing norm to penalty the basis vector learned by linear graph embedding, Wang [41] proposed structured sparse linear graph embedding (SSLGE), which is also a sparse learning framework. Meng et al. [42] and Wang et al. [43], respectively, extended PCA-L1 and 2DPCA-L1 with sparsity. By using an overcomplete dictionary, sparse coding [44] can represent a signal sparsely. Recently, Liu et al. [45] proposed a novel sparse coding method, called multiview Hessian discriminative sparse coding (mHDSC). mHDSC integrates Hessian regularization with discriminative sparse coding for multiview learning problems.

In this paper, we propose a novel spare L1-norm-based maximum margin criterion (SMMC-L1). We first replace L2-norm in conventional MMC with L1-norm, and then utilize the elastic net to penalize the projection vectors. The role of L1-norm in the proposed SMMC-L1 method is twofold. One is the robust measurement of the between-class dispersion and the within-class dispersion. The other is used as penalty by which the spare basis vectors can be obtained. We also propose an iterative algorithm to solve SMMC-L1.

The remainder of the paper is organized as follows. The conventional LDA and MMC are briefly reviewed in Section 2. In Section 3, we present the SMMC-L1 method, including its objective function and algorithmic procedure. The experiment results are reported in Section 4. Finally, we conclude the paper in Section 5.

2. Outline of LDA and MMC

Let $X = \{\mathbf{x}_j^i, j = 1,2,\ldots,n_i; i = 1,2,\ldots,k\} \in R^{d\times n}$ be the given training samples, where $\mathbf{x}_j^i$ is the jth samples of the ith class, k is the number of the classes, n_i is the number of the samples of ith class, d is the dimensionality of the training samples and $n = \sum_{i=1}^{k} n_i$ is the number of the data set. In LDA (termed as LDA-L2), between-class scatter matrix and within-class scatter matrix, are respectively defined as follows:

$$S_b = \sum_{i=1}^{k} n_i(\mathbf{m}_i - \mathbf{m})(\mathbf{m}_i - \mathbf{m})^T, \tag{1}$$

$$S_w = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(\mathbf{x}_j^i - \mathbf{m}_i)(\mathbf{x}_j^i - \mathbf{m}_i)^T, \tag{2}$$

where $\mathbf{m}_i = (1/n_i)\sum_{j=1}^{n_i}\mathbf{x}_j^i$ is the mean of the ith class and $\mathbf{m} = (1/n)\sum_{i=1}^{k}\sum_{j=1}^{n_i}\mathbf{x}_j^i$ is the global mean of the data set.

The optimal projection vector $\mathbf{w} \in R^d$ of LDA can be obtained by maximizing the following so-called Fisher criterion:

$$J(\mathbf{w}) = \frac{\mathbf{w}^T S_b \mathbf{w}}{\mathbf{w}^T S_w \mathbf{w}}. \tag{3}$$

If the matrix S_w is nonsingular, the optimal projection vector $\mathbf{w}$ is the leading eigenvector of $S_w^{-1}S_b$.

The conventional LDA cannot work when the matrix S_w is singular. To address this issue, maximum margin criterion (MMC) (termed as MMC-L2) [7–10] has been proposed. The discriminant criterion based on MMC is defined as follows

$$J(\mathbf{w}) = \mathbf{w}^T S_b \mathbf{w} - \mathbf{w}^T S_w \mathbf{w}. \tag{4}$$

The optimal projection vector $\mathbf{w}$ is the leading vector of $S_b - S_w$.

3. Sparse L1-norm-based maximum margin criterion (SMMC-L1)

3.1. Problem formulation

In this subsection, we will present our proposed sparse L1-norm-based maximum margin criterion.

Let

$$H_b = [\,\sqrt{n_1}(\mathbf{m}_1 - \mathbf{m}), \quad \sqrt{n_2}(\mathbf{m}_2 - \mathbf{m}), \quad \ldots, \quad \sqrt{n_k}(\mathbf{m}_k - \mathbf{m})\,], \tag{5}$$

$$H_w = \left[\mathbf{x}_1^1 - \mathbf{m}_1, \quad \ldots \quad \mathbf{x}_{n_1}^1 - \mathbf{m}_1, \quad \ldots, \quad \mathbf{x}_1^k - \mathbf{m}_k, \quad \ldots, \quad \mathbf{x}_{n_k}^k - \mathbf{m}_k\right]. \tag{6}$$

By simply transforming, Eq. (4) can be reformulated as

$$J(\mathbf{w}) = \|\mathbf{w}^T H_b\|_2^2 - \|\mathbf{w}^T H_w\|_2^2 \tag{7}$$

where $\|\cdot\|_2$ denotes L2-norm. From Eq. (7) we can find that the objective function of MMC-L2 is derived from L2-norm. However, L2-norm is more sensitive to noise and outliers than L1-norm since the square operation in L2-norm will magnify the effects of the noise and outliers. Then L1-norm based approaches are believed to be more robust to noise and outliers than L2-norm based ones. Besides, sparse basis vectors, which can be obtained by using L1-norm penalty, can encode semantic information and obtain more discriminant information than condense basis ones. Motivated by these ideas, we propose to maximize the objective function as follows:

$$F(\mathbf{w}) = \|\mathbf{w}^T H_b\|_1 - \|\mathbf{w}^T H_w\|_1 - \lambda\|\mathbf{w}\|_1 - \frac{\eta}{2}\|\mathbf{w}\|_2^2 \tag{8}$$

where $\lambda > 0$ and $\eta > 0$ are tuning parameters. It is difficult to solve Eq. (8) directly and obtain a global optimal solution due to the abso-