Pattern Recognition Letters 34 (2013) 679–685

Contents lists available at SciVerse ScienceDirect

Pattern Recognition Letters

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

Generalization of linear discriminant analysis using L_p-norm

Jae Hyun Oh [1], Nojun Kwak [*,2]

Department of Electrical & Computer Engineering, Ajou University, San 5, Woncheon-dong, Yeongtong-gu, Suwon 443-749, Republic of Korea

ARTICLE INFO

Article history:
Received 16 July 2012
Available online 4 February 2013

Communicated by S. Sarkar

Keywords:
LDA
Norm
Outlier
LDA-L_p

ABSTRACT

In this paper, the linear discriminant analysis (LDA) is generalized by using an L_p-norm optimization technique. Although conventional LDA based on the L_2-norm has been successful for many classification problems, performances can degrade with the presence of outliers. The effect of outliers which is exacerbated by the use of the L_2-norm can cause this phenomenon. To cope with this problem, we propose an LDA based on the L_p-norm optimization technique (LDA-L_p), which is robust to outliers. Arbitrary values of p can be used in this scheme. The experimental results show that the proposed method achieves high recognition rate for many datasets. The reason for the performance improvements is also analyzed.

1. Introduction

During the last few decades, numerous feature extraction methods have been proposed for data analysis and object classification in the computer vision and pattern recognition communities. Principal component analysis (PCA) (Fukunaga, 1990; Turk et al., 1991), independent component analysis (ICA) (Bell and Sejnowski, 1995; Kwak and Choi, 2003) and linear discriminant analysis (LDA) (Belhumeur et al., 1997; Martinez and Kak, 2001) are successful representatives of linear subspace-based feature extraction methods, and many further improvements continue to be researched. Unlike PCA and ICA, LDA is designed for supervised learning and has been widely used for classification problems. The goal of LDA is to find a series of projections that maximize the ratio of between class and within class variance, both of which are based on the L_2 norm. It is known that conventional L_2-norm based LDA is optimal if each class has the same Gaussian distribution. Although conventional LDA, based on the L_2-norm, has been successful for many problems, there are numerous problems whose class-specific distributions are far from Gaussian. For these problems, the performances of LDA could degrade with the presence of outliers because L_2-norm-based methods are dominated by samples with large norms.

As a generalized version of LDA, Yang et al. (2011) introduced a new concept of designing a discriminant analysis method and Yang and Yang (2003) suggested a complete PCA plus LDA algorithm. A new kernel Fisher discriminant analysis framework was also proposed to implement the KPCA plus LDA strategy (Yang et al., 2005). An extension of LDA to regression problems and its kernel version were also proposed in (Kwak and Lee, 2010; Kwak, 2012), respectively.

There are many studies aimed at enhancing the performance of the conventional L_2-norm-based feature extraction methods. In particular, many studies have focused on PCA algorithms based on the L_1-norm instead of the L_2-norm. L_1-norm-based PCA (L1-PCA) Ke and Kanade (2005) finds the optimal projection vectors that minimize the L_1-norm-based reconstruction error in the input space through linear or quadratic programming which is computationally expensive. Another drawback of L1-PCA is that it is not rotational invariant. Ding et al. (2006) proposed R1-PCA, which combines the merits of L2-PCA and those of L1-PCA. Unlike L1-PCA, it is rotation-invariant while it successfully suppresses the effect of outliers, as L1-PCA does. On the other hand, PCA-L1 (Kwak, 2008) maximizes L_1-norm-based dispersion in the feature space, instead of maximizing L_2-norm-based variance, to achieve robust and rotation-invariant PCA. Several extensions of PCA-L1 have been introduced recently. 2DPCA-L1 (Li et al., 2009) is an L_1-norm version of 2DPCA that is robust to outliers with very simple iteration process. In addition, Kwak and Oh (2009) proposed SL1-BDA, an L_1-norm version of biased discriminant analysis that was originally developed for one-class classification problems. It tries to reduce the negative effect of extracting features due to negative samples that are very far from the center of positive samples and utilizes the L_1-norm instead of the L_2-norm.

There are also studies that try to extend LDA using other norms than the L_2-norm. The novel rotation-invariant L_1-norm (R_1-norm)-based discriminant criterion called DCL_1, which better characterizes intra-class compactness and inter-class separability by using the rotation-invariant L_1-norm, was proposed in (Li et al., 2010).

* Corresponding author. Tel.: +82 (0) 31 219 2480; fax: +82 (0) 31 212 9531.
E-mail addresses: hyunsda@ajou.ac.kr (J.H. Oh), nojunk@ajou.ac.kr, nojunk@ieee.org (N. Kwak).
[1] Jae Hyun Oh is pursuing a Ph.D. degree at the Department of Electrical & Computer Engineering, Ajou University, Republic of Korea.
[2] Nojun Kwak is an associate professor at the Department of Electrical & Computer Engineering, Ajou University, Republic of Korea.

0167-8655/$ - see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.patrec.2013.01.016

In addition, robust L1-norm based tensor analysis (TPCA-L1) formulates the reconstruction error with the L_1-norm (Pang et al., 2010). The use of the L_1 norm makes tensor analysis robust to outliers. Moreover, the algorithm converges well in several iterations. Fast Haar transform (FHT) based PCA and FHT-based spectral regression discriminant analysis have also been proposed to solve the problem of the computationally expensive processing time of the projection process (Pang et al., 2009). Recently, we studied the generalization of the L_1 norm to an L_p norm with an arbitrary p value for PCA (Kwak, 2013). This algorithm uses a new L_p-norm optimization technique using the gradient search method.

In this paper, a method is proposed for classification, which is based on the L_p-norm optimization technique as a generalized version of LDA. We address a novel method of LDA that uses the L_p-norm instead of the L_2-norm to obtain a robust and rotation-invariant version of LDA. The objective function is formulated using the general L_p-norm in both the numerator and denominator and the optimal solution is found using the steepest-gradient method. The effect of outliers for each method is analyzed, and it is shown that the proposed LDA based on the L_p-norm is more robust to outliers. In doing so, a novel methodology for measuring the effect of outliers is also presented.

This paper is organized as follows. In Section 2, conventional LDA is overviewed, and the new algorithm LDA-Lp which uses the L_p-norm instead of the L_2-norm is presented. Section 3 shows the experimental results with an analysis on the effect of outliers. Finally, conclusions are presented in Section 4.

2. Methods

2.1. LDA (based on the L_2-norm)

LDA is one of the well-known methods of supervised dimensionality reduction for classification problems. It tries to find transformations that maximize the ratio of the between-class and the within-class scatter matrices. Consider a dataset $\{(x_i, c_i)\}_{i=1}^N$, where $x_i \in \mathbb{R}^d$ and $c_i \in \{1, \ldots, C\}$ are an input and the corresponding class, respectively. The between-class scatter matrix S_B and the within-class scatter matrix S_W are defined, respectively, as:

$$\begin{aligned} S_B &= \sum_{c=1}^{C} N_c (m_c - m)(m_c - m)^T, \\ S_W &= \sum_{i=1}^{N} (x_i - m_{c_i})(x_i - m_{c_i})^T, \end{aligned} \tag{1}$$

where N_c is the number of samples belonging to class c, and $m \triangleq \frac{1}{N}\sum_{i=1}^{N} x_i$ and $m_c \triangleq \frac{1}{N_c}\sum_{i \in \{j | c_j = c\}} x_i$ are the total mean and the class mean of the input data.

The LDA is formulated to find M projection vectors $\{w_i\}_{i=1}^M$ that maximize Fisher's criterion, as follows:

$$W_{LDA} = \underset{W}{argmax} \frac{|W^T S_B W|}{|W^T S_W W|}. \tag{2}$$

Here, the ith column of W corresponds to w_i. Maximizing the above Fisher's criterion is equivalent to solving the following eigenvalue decomposition problem:

$$S_B w_i = \lambda_i S_W w_i \quad \lambda_1 \geqslant \lambda_2 \geqslant \cdots \geqslant \lambda_m. \tag{3}$$

Then, the linear projections $\{w_i\}_{i=1}^M$ can be obtained. However, conventional LDA is very sensitive to the presence of outliers, because both S_B and S_W in (1) are dominated by a set of outliers with large norms. To alleviate this problem, we propose a novel method that utilize the L_p-norm instead of the L_2-norm in the subsequent subsection.

2.2. Algorithm: LDA-L_p

It is well known that an algorithm based on the L_p-norm is less sensitive to the samples with large norms compared to the corresponding algorithm based on the L_2-norm.

Therefore, we define a new maximization problem for the design of an L_p-norm-based LDA. Consider the following L_p-norm maximization problem with the constraint $||w||_2 = 1$.

$$F_p(w) = \frac{\sum_{c=1}^{C} N_c |w^T (m_c - m)|^p}{\sum_{i=1}^{N} |w^T (x_i - m_{c_i})|^p}. \tag{4}$$

This can be solved by taking the gradient of $F_p(w)$ with respect to w. An important point to note here is that because of the absolute value operator in (4), the gradient of $F_p(w)$ is not well defined on some singular points. To avoid this technical difficulty, a sign function below is introduced.

$$sgn(a) = \begin{cases} 1 & \text{if } a > 0, \\ 0 & \text{if } a = 0, \\ -1 & \text{if } a < 0. \end{cases} \tag{5}$$

With the help of this sign function, (4) can be rewritten as follows:

$$F_p(w) = \frac{\sum_{c=1}^{C} N_c [\mathrm{sgn}(w^T (m_c - m)) w^T (m_c - m)]^p}{\sum_{i=1}^{N} [\mathrm{sgn}(w^T (x_i - m_{c_i})) w^T (x_i - m_{c_i})]^p}. \tag{6}$$

Now, in order to get an optimal w which maximizes (6), we can take a gradient of $F_p(w)$ in (6) with respect to w as follows:

$$\begin{aligned} &\nabla_w = \frac{dF_p(w)}{dw} = \frac{A \times B}{E} - \frac{C \times D}{E}, \\ &\text{where} \quad A = p \sum_{c=1}^{C} N_c \mathrm{sgn}(w^T (m_c - m)) |w^T (m_c - m)|^{p-1} (m_c - m), \\ &B = \sum_{i=1}^{N} [\mathrm{sgn}(w^T (x_i - m_{c_i})) w^T (x_i - m_{c_i})]^p, \\ &C = \sum_{c=1}^{C} N_c [\mathrm{sgn}(w^T (m_c - m)) w^T (m_c - m)]^p, \\ &D = p \sum_{i=1}^{N} \mathrm{sgn}(w^T (x_i - m_{c_i})) |w^T (x_i - m_{c_i})|^{p-1} (x_i - m_{c_i}), \\ &E = \Big(\sum_{i=1}^{N} [\mathrm{sgn}(w^T (x_i - m_{c_i})) w^T (x_i - m_{c_i})]^p\Big)^2. \end{aligned} \tag{7}$$

The above gradient is well defined when $w^T(m_c - m) \neq 0$ and $w^T(x_i - m_{c_i}) \neq 0$ for all x_i. Furthermore, it is also well defined if $p > 1$ on singular points where $w^T(m_c - m) = 0$ or $w^T(x_i - m_{c_i}) = 0$ for some x_i's. On the other hand, if $p = 1$ the term A or D in (7) is not well defined on the singular points because 0^0 is hard to define, and if $p < 1$, A or D diverges at the singular points. To avoid this problem, we add a singularity check step before computing the gradient.

The optimal solution to this problem can be obtained using the steepest-gradient method as follows:

i. Initialization
   - $t \leftarrow 0$. Set $w(0)$ such that $||w(0)||_2 = 1$.
ii. Singularity check (applies only when $p \leqslant 1$)
   - If $w(t)^T(m_c - m) = 0$ or $w(t)^T(x_i - m_{c_i}) = 0, w(t) \leftarrow \frac{(w(t)+\delta)}{||w(t)+\delta||_2}$ where δ is a small random vector.
iii. Computation of ∇_w in (7)
iv. Gradient search
   - $w(t+1) \leftarrow w(t) + \alpha \nabla_w$ where α is a learning rate.