Pattern Recognition Letters 34 (2013) 1094–1100

Contents lists available at SciVerse ScienceDirect

Pattern Recognition Letters

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

On improving robustness of LDA and SRDA by using tangent vectors

Mauricio Villegas *, Roberto Paredes

Institut Tecnològic d'Informàtica, Universitat Politècnica de València, Camí de Vera s/n, 46022 València, Spain

ARTICLE INFO

Article history:
Received 5 November 2012
Available online 15 March 2013

Communicated by J. Yang

Keywords:
Subspace learning
Dimensionality reduction
Tangent vectors
LDA
SRDA

ABSTRACT

In the area of pattern recognition, it is common for few training samples to be available with respect to the dimensionality of the representation space; this is known as the *curse of dimensionality*. This problem can be alleviated by using a dimensionality reduction approach, which overcomes the curse relatively well. Moreover, supervised dimensionality reduction techniques generally provide better recognition performance; however, several of these tend to suffer from the curse when applied directly to high-dimensional spaces. We propose to overcome this problem by incorporating additional information to supervised subspace learning techniques using what is known as *tangent vectors*. This additional information accounts for the possible differences that the sample data can suffer. In fact, this can be seen as a way to model the unseen data and make better use of the scarce training samples. In this paper, methods for incorporating tangent vector information are described for one classical technique (LDA) and one state-of-the-art technique (SRDA). Experimental results confirm that this additional information improves performance and robustness to known transformations.

1. Introduction

In the area of pattern recognition, it is common for few training samples to be available with respect to the dimensionality of the representation space; this is known as the *curse of dimensionality* (Bellman, 1961). To handle this problem, it has become popular to use dimensionality reduction (also known as subspace learning) as a preprocessing step. However, several dimensionality reduction techniques also struggle due to the lack of samples, or in other words, they are also affected by the curse. In these cases, a tandem strategy is often used by applying a more robust technique as an initial step. This strategy, though less useful from the discriminative point of view, reduces the dimensionality down to a more appropriate size for the subsequent *discriminative* dimensionality reduction. The most well-known tandem is PCA + LDA (Yang and Yang, 2003; Yang et al., 2005), i.e., where Principal Component Analysis is performed over the original representation space and afterwards Linear Discriminant Analysis (Fukunaga, 1990) is applied. Note that PCA is an unsupervised technique, whereas LDA is supervised, which is crucial since the use of a supervised technique generally helps to boost the recognition performance considerably.

The motivation for this paper was to improve supervised subspace learning techniques so that they are able to cope with scarce data in high-dimensional feature spaces. Even though a tandem strategy overcomes the curse of dimensionality for the less robust supervised subspace learning techniques, it would clearly be more desirable for these techniques to work well in high-dimensional spaces, up to the point of not necessarily requiring a previous dimensionality reduction. This goal is addressed in this paper by considering the known transformations that a sample can exhibit which do not modify the class membership. In fact, we can consider that these known transformations model the unseen samples (as if increasing the training set), thereby overcoming the curse of dimensionality. Consider for instance the rotations and displacements of facial images due to imperfect alignments. Even though these variations are expected to appear, it is known that they do not change the identity of the person appearing in the image. One method to account for the possible combinations of these base transformations is the *tangent distance* (Simard et al., 1993); however, it is only applicable to distance-based classifiers. In this work only the *tangent vectors* are used as a way to obtain more information from the training set, without imposing any restrictions on the classifier. Related to this paper, Schölkopf et al. (1997) and Mika et al. (1999) use the tangent vectors to improve Support Vector kernels and make them somewhat invariant to the tangent vector transformations.

The paper addresses two supervised techniques: the first is the classic LDA (including the PCA + LDA variant), and the second is the state-of-the-art Spectral Regression Discriminant Analysis (SRDA) (Cai et al., 2008; Chen et al., 2009). In the literature, there are many other methods that could be considered (see for instance Burges (2005) and van der Maaten and Postma (2009) for a review of some of them). Nevertheless, the techniques we have chosen are known

* Corresponding author. Tel.: +34 963877235; fax: +34 963877239.
E-mail addresses: mvillegas@iti.upv.es (M. Villegas), rparedes@iti.upv.es (R. Paredes).

0167-8655/$ - see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.patrec.2013.03.001

to perform well and illustrate an idea which could be applied to other methods in future works.

The contributions of the paper are the following. First, we reformulate LDA so that it is expressed in terms of the covariance matrix, which can be better estimated by using tangent vectors (see Section 3.1). This modification helps to overcome the singularity problems that LDA has when there are few training samples, improves recognition performance, and also increases the robustness of the learned subspace to known transformations. Second, we present a method to incorporate the tangent vector information in SRDA that keeps the characteristic of being solvable by systems of linear equations, thus continuing to be efficient for learning (see Section 3.2). Also, the recognition performance improves and the robustness of the learned subspaces to known transformations increases. Finally, in Section 4, we present empirical results that confirm the benefits when using the proposed modifications.

2. Preliminaries and overview of the tangent vectors

Suppose we have a point $\mathbf{x} \in \mathbb{R}^D$ generated from an underlying distribution, and that the possible transformations or manifold of $\mathbf{x}$ is given by $\hat{\mathbf{t}}(\mathbf{x},\boldsymbol{\alpha})$, a function which depends on a parameter vector $\boldsymbol{\alpha} \in \mathbb{R}^L$ with the characteristic that $\hat{\mathbf{t}}(\mathbf{x},\mathbf{0}) = \mathbf{x}$. The dimensionality of $\boldsymbol{\alpha}$ is essentially the degrees of freedom of possible variations that $\mathbf{x}$ can have. In real applications, the manifold $\hat{\mathbf{t}}(\mathbf{x},\boldsymbol{\alpha})$ is highly non-linear; however, for values close to $\boldsymbol{\alpha} = \mathbf{0}$, it can be reasonable to approximate it by a linear subspace. This can also be interpreted as representing the manifold by its Taylor series expansion evaluated at $\boldsymbol{\alpha} = \mathbf{0}$, and discarding the second and higher order terms (Simard et al., 1998), i.e.,

$$\hat{\mathbf{t}}(\mathbf{x},\boldsymbol{\alpha}) = \hat{\mathbf{t}}(\mathbf{x},\mathbf{0}) + \sum_{l=1}^{L}\alpha_l \frac{\partial \hat{\mathbf{t}}(\mathbf{x},\boldsymbol{\alpha})}{\partial \alpha_l} + \ldots \Bigg|_{\boldsymbol{\alpha}=\mathbf{0}} \tag{1}$$

$$\approx \mathbf{t}(\mathbf{x},\boldsymbol{\alpha}) = \mathbf{x} + \sum_{l=0}^{L}\alpha_l \mathbf{v}_l. \tag{2}$$

The partial derivatives $\mathbf{v}_l = \partial\hat{\mathbf{t}}/\partial\alpha_l$ are known as the *tangent vectors*, since they are tangent to the transformation manifold $\hat{\mathbf{t}}$ at point $\mathbf{x}$.

The concept of the tangent vector approximation is illustrated in Fig. 1 for a single direction of variability. As can be observed, the approximation can be quite good for small values of $\|\boldsymbol{\alpha}\|$; however, as the norm $\|\boldsymbol{\alpha}\|$ increases, the deviation from the true manifold $\hat{\mathbf{t}}$ is expected to increase.

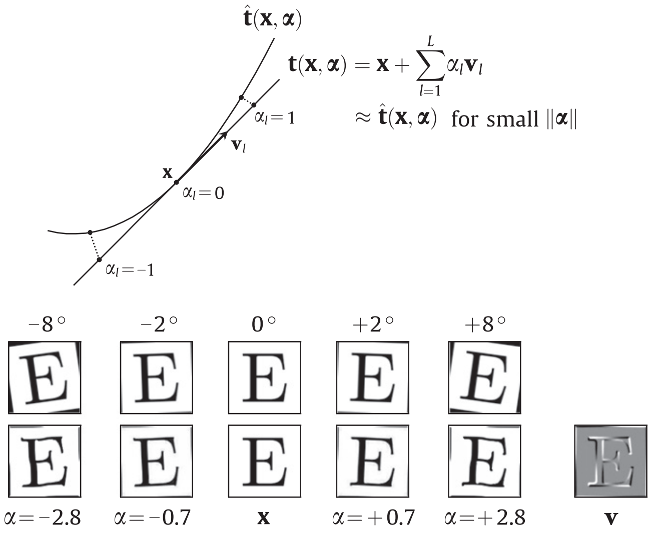
$\hat{\mathbf{t}}(\mathbf{x},\boldsymbol{\alpha})$
$\mathbf{t}(\mathbf{x},\boldsymbol{\alpha}) = \mathbf{x} + \sum_{l=1}^{L}\alpha_l\mathbf{v}_l$
$\approx \hat{\mathbf{t}}(\mathbf{x},\boldsymbol{\alpha})$ for small $\|\boldsymbol{\alpha}\|$
$\alpha_l = 1$
$\mathbf{v}_l$
$\mathbf{x}$
$\alpha_l = 0$
$\alpha_l = -1$
−8°
−2°
0°
+2°
+8°
$\alpha = -2.8$
$\alpha = -0.7$
$\mathbf{x}$
$\alpha = +0.7$
$\alpha = +2.8$
$\mathbf{v}$

Fig. 1. Top: An illustration of the linear approximation of transformations by means of tangent vectors. Bottom: An example of an image rotated at various angles and the corresponding rotation approximations using a tangent vector.

When comparing two points, as a similarity measure between them, it would be ideal to use the minimum distance between their respective transformation manifolds. As an approximation to this, one can use the minimum distance between the subspaces spanned by the tangent vectors (Simard et al., 1998), which is known as the *tangent distance* (TD). The *single-sided tangent distance* considers only one of the tangent subspaces and has the advantage of being more efficient to compute (Dahmen et al., 2001). From a classification perspective, the tangent subspace can either be for the reference (RTD) or the observation (OTD).

2.1. Estimation of tangent vectors

There are several methods to estimate the tangent vectors, although, unfortunately, there is no general way to estimate them for every task. The most intuitive method is to use the difference between the sample and its transformation as tangent vectors. However, this method can only be used if it is possible to generate a transformation of a sample. The most well-known method of estimating tangent vectors is the one proposed by Simard et al. (1998). This method is only applicable to image based problems, having been employed successfully to model the following: scaling, rotation, vertical and horizontal translation, parallel and diagonal hyperbolic transformations, and trace thickening.

There are other methods that try to estimate the tangent vectors from the training set, instead of adding some prior knowledge. One method of this type is presented in (Keysers et al., 2004), which is based on maximum likelihood estimation. Another method is to use the difference between a sample and its nearest neighbors from the same class as tangent vectors.

The methods of Simard and the nearest neighbors were used in the experiments. However, as discussed in Section 3 and empirically observed, the latter is less useful since it does not provide as much additional information and it does not help to overcome the singularity problems.

3. Tangent vectors in subspace learning

3.1. Tangent vectors in LDA

The objective of LDA is that the obtained subspace should discriminate the classes well. To this end, LDA simultaneously maximizes the distances between the class centers (between-class scatter matrix) and minimizes the distances within each class (within-class scatter matrix). It is straightforward to reformulate LDA so that it is stated in terms of the covariance matrix $\boldsymbol{\Sigma}_{\mathrm{x}}$ and a normalized between-class scatter matrix $\boldsymbol{\Sigma}_{\boldsymbol{\mu}}$. The objective function is then

$$\hat{\mathbf{B}} = \arg\max_{\mathbf{B}} \frac{\mathrm{Tr}(\mathbf{B}^{\mathsf{T}}\boldsymbol{\Sigma}_{\boldsymbol{\mu}}\mathbf{B})}{\mathrm{Tr}(\mathbf{B}^{\mathsf{T}}\boldsymbol{\Sigma}_{\mathrm{x}}\mathbf{B})}. \tag{3}$$

The solution of the LDA objective (3) is the following generalized eigenvalue decomposition

$$\boldsymbol{\Sigma}_{\boldsymbol{\mu}}\mathbf{B} = \boldsymbol{\Sigma}_{\mathrm{x}}\mathbf{B}\boldsymbol{\Lambda}, \tag{4}$$

with $\boldsymbol{\Lambda}$ being a diagonal matrix of generalized eigenvalues and the columns of $\mathbf{B}$ being the generalized eigenvectors.

By having a solution of LDA in terms of the covariance matrix, for a given dataset $\mathcal{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$, we are able to use a better empirical estimation for $\boldsymbol{\Sigma}_{\mathrm{x}}$ that considers tangent vectors (Keysers et al., 2004) given by