Available online at www.sciencedirect.com
ScienceDirect
Neural Networks 21 (2008) 214–221
Neural Networks
www.elsevier.com/locate/neunet

2008 Special Issue

Learning representations for object classification using multi-stage optimal component analysis☆

Yiming Wu[a,*], Xiuwen Liu[a], Washington Mio[b]

[a] Department of Computer Science, Florida State University, Tallahassee, FL 32306, USA
[b] Department of Mathematics, Florida State University, Tallahassee, FL 32306, USA

Received 8 August 2007; received in revised form 2 December 2007; accepted 11 December 2007

Abstract

Learning data representations is a fundamental challenge in modeling neural processes and plays an important role in applications such as object recognition. Optimal component analysis (OCA) formulates the problem in the framework of optimization on a Grassmann manifold and a stochastic gradient method is used to estimate the optimal basis. OCA has been successfully applied to image classification problems arising in a variety of contexts. However, as the search space is typically very high dimensional, OCA optimization often requires expensive computational cost. In multi-stage OCA, we first hierarchically project the data onto several low-dimensional subspaces using standard techniques, then OCA learning is performed hierarchically from the lowest to the highest levels to learn about a subspace that is optimal for data discrimination based on the K-nearest neighbor classifier. One of the main advantages of multi-stage OCA lies in the fact that it greatly improves the computational efficiency of the OCA learning algorithm without sacrificing the recognition performance, thus enhancing its applicability to practical problems. In addition to the nearest neighbor classifier, we illustrate the effectiveness of the learned representations on object classification used in conjunction with classifiers such as neural networks and support vector machines.

Keywords: Dimension reduction; Linear data representations; Feature selection; Object recognition; Object classification; Optimal component analysis

1. Introduction

Learning algorithms for neural network models have been a focal point (Bishop, 1995; Geman & Bienenstock, 1992). Bishop (1995) stated that the choice of pre-processing and feature extraction techniques is "one of the most significant factors in determining the performance of the final system". In the past decades, linear subspace representation methods, such as Principal Component Analysis (PCA) (Jolliffe, 1986; Turk & Pentland, 1991), Independent Component Analysis (ICA) (Comon, 1994; Hyvarinen, Karhunen, & Oja, 2001), Canonical Correlation Analysis (CCA) (Anderson, 2003; Reiter, Donner, Langs, & Bischof, 2006) and Linear Discriminant Analysis (LDA) (Duda, Hart, & Stock, 2000; Zhao, Chellappa, & Phillips, 1994), have been widely used for learning representations suitable for neural networks. For example, Zhu and Yu (1994) implemented a system for face recognition with eigenfaces and a backpropagation neural network. Eleyan and Demirel (2005) proposed a face recognition method in which features are first extracted using PCA and faces are classified using feed-forward neural networks. ICA-based recognition methods, (e.g. Bartlett, Movellen, and Sejnowski (2002) and Kwak and Pedrycz (2007)), tend to give better recognition performance than PCA-based methods as they take high-order statistics of data into account. LDA-based methods, on the other hand, use class information and try to find an optimal basis that maximize the between-class scatter while minimizing the within-class scatter, and are also frequently employed in face and object recognition (Etemad & Chellappa, 1997).

These classical linear representation methods, in general, are not optimal for classification or recognition. For example, PCA and ICA are optimized for data reconstruction and statistical independence, not for the selection of discriminative features. CCA is another multivariate statistical method which extracts

☆ An abbreviated version of some portions of this article appeared in Wu, Liu, and Mio (2007) as part of the IJCNN 2007 Conference Proceedings, published under IEE copyright.

* Corresponding author. Tel.: +1 850 645 2257; fax: +1 850 644 0058.
E-mail addresses: ywu@cs.fsu.edu (Y. Wu), liux@cs.fsu.edu (X. Liu), mio@math.fsu.edu (W. Mio).

0893-6080/$ - see front matter © 2007 Elsevier Ltd. All rights reserved.
doi:10.1016/j.neunet.2007.12.023

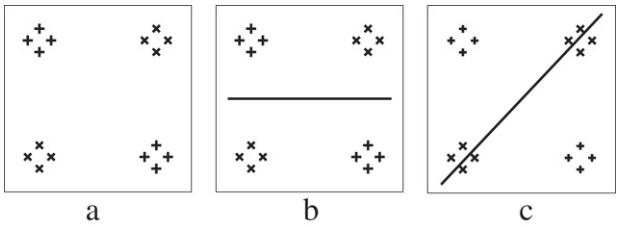

Fig. 1. A synthetic dataset consisting of two classes, each with two clusters of four points: (a) the data set of two classes ('+' and '×') with eight points each in $\Re^2$; (b) the one-dimensional subspace obtained from PCA, ICA, and LDA; (c) a one-dimensional optimal subspaces representation obtained using OCA.

the most coherent features among two data channels. LDA assumes that the conditional probability distribution of each class is Gaussian with the same variance. As the distributions of real images are typically non-Gaussian (e.g. Srivastava, Liu, and Grenander (2002)), in recognition tasks, there is no theoretic guarantee of optimality of LDA basis. This is also evident in comparative studies reported in the literature (e.g. Belhumeour, Hespanha, and Kriegman (1997) and Martinez and Kak (2001)). In fact, one can construct examples in which all the common choices of learning algorithms give the worst possible performance. Such an example is shown in Fig. 1, which consists of two classes ('+' and '×') with eight points each, and the points are presented in clusters of four. It can be shown that the one-dimensional subspace resulting from PCA, ICA, and LDA coincides with either the horizontal or the vertical axis. If we use the nearest neighbor classifier and let a point from each cluster be used for training, the one-dimensional basis obtained from PCA, ICA, and LDA gives the worst performance.

It is thus apparent that, in the context of object recognition, a more relevant question is that of finding a linear representation that optimally selects discriminating features. Unlike the classical methods, the recently proposed Optimal Component Analysis (OCA) (Liu, Srivastava, & Gallivan, 2004; Srivastava & Liu, 2005) provides a general optimality criterion. The search for optimal linear representations, or an optimal subspace, is based on a stochastic optimization process which maximizes a pre-specified performance function over all subspaces of a particular dimension and is estimated using a Markov Chain Monte Carlo (MCMC) type algorithm. OCA exhibits good performance on face and object recognition. Fig. 1(c) shows an optimal subspace representation obtained by the OCA method.

The stochastic search techniques employed in OCA typically result in heavy computational costs, which limits the applicability of OCA to practical problems that involve feature extraction and object recognition. As an example, consider a facial recognition experiment based on the ORL data set (Samaria & Harter, 1994). OCA learning takes approximately one day to run 1000 iterations to estimate an optimal subspace. Obviously, this is not practical for most object recognition applications. In our previous work, a two-stage strategy was proposed to address this problem (Wu, Liu, Mio, & Gallivan, in press). In this approach, the input data is first reduced to a lower dimension using methods such as PCA or LDA; then, the OCA search is performed in the reduced space. This strategy leads to significant computational gains. However, it is generally difficult to determine a good choice for the reduced subspace. In this paper, a multi-stage strategy is proposed to address this problem. The idea of multi-stage OCA (M-OCA) was presented in a previous short paper (Wu et al., 2007): the data is first hierarchically reduced into several levels using shrinkage matrices; then, the OCA search is performed hierarchically from the lowest to the highest levels. The basis is expanded progressively from the optimal basis obtained in the previous level. As the learning process of each level starts with a good initial selection from the previous level, M-OCA achieves good recognition performance. Also, since the dimensions of the Grassmann manifolds at the lower levels are much smaller than that of the Grassmannian in the original space, M-OCA reduces the computational costs associated with the original algorithm significantly, thus making OCA learning feasible in applications.

The rest of the paper is organized as follows: Section 2 gives a brief review of OCA and the proposed M-OCA method is presented in Section 3; A comprehensive study of the performance of the M-OCA algorithm is presented in Section 4; Section 5 concludes the paper with a summary and a discussion of future work.

2. Optimal component analysis

Optimal Component Analysis is a dimension reduction technique designed to find an optimal subspace (of a prescribed dimension) of feature space that optimizes the ability of the nearest neighbor classifier to index and classify images or other data. The measurement of optimality is based on training data and the algorithm yields an orthonormal basis of the estimated optimal subspace. More specifically, let $U \in \Re^{n\times d}$ be a matrix whose columns form an orthonormal basis of a d-dimensional subspace of $\Re^n$, where n is the size of the input image and d is the dimension of the desired subspace (generally $n \gg d$). For an image I, viewed as an n-vector, the vector of coefficients is given by $\alpha(I, U) = U^T I \in \Re^d$ and represents the orthogonal projection of I onto the subspace S_U spanned by the columns of U. Suppose the training data consists of representatives of C classes of images, with each class represented by k_{train} images denoted by $I_{c,1}, \ldots, I_{c,k_{\text{train}}}$, where $c = 1, \ldots, C$. Let

$$\rho(I_{c,i}, U) = \frac{\min_{c'\neq c, j} D(I_{c,i}, I_{c',j}; U)}{\min_{j\neq i} D(I_{c,i}, I_{c,j}; U) + \epsilon}. \tag{1}$$

The numerator is the distance from $I_{c,i}$ to the closest training image not in its class and the denominator is the distance from $I_{c,i}$ to the closest training image in the same class. Here, D denotes Euclidean distance; that is,

$$D(I_1, I_2; U) = \|\alpha(I_1, U) - \alpha(I_2, U)\|, \tag{2}$$

where $\|\cdot\|$ is the usual 2-norm. In Eq. (1), $\epsilon > 0$ is a small number introduced to avoid division by zero. Note that large values of ρ are desirable, since this means that $I_{c,i}$ will be closer to its class than to other classes after projection onto the subspace S_U. A performance function F is defined to