Available online at www.sciencedirect.com
SCIENCE DIRECT®

Pattern Recognition 39 (2006) 164–170

PATTERN RECOGNITION
THE JOURNAL OF THE PATTERN RECOGNITION SOCIETY

www.elsevier.com/locate/patcog

Boosted discriminant projections for nearest neighbor classification

David Masip, Jordi Vitrià*

Computer Vision Center, Dept. Informàtica, Universitat Autònoma de Barcelona, Bellaterra, Spain

Abstract

In this paper we introduce a new embedding technique to find the linear projection that best projects labeled data samples into a new space where the performance of a Nearest Neighbor classifier is maximized. We consider a large set of one-dimensional projections and combine them into a projection matrix, which is not restricted to be orthogonal. The embedding is defined as a classifier selection task that makes use of the AdaBoost algorithm to find an optimal set of discriminant projections. The main advantage of the algorithm is that the final projection matrix does not make any global assumption on the data distribution, and the projection matrix is created by minimizing the classification error in the training data set. Also the resulting features can be ranked according to a set of coefficients computed during the algorithm. The performance of our embedding is tested in two different pattern recognition tasks, a gender recognition problem and the classification of manuscript digits.

Keywords: Feature extraction; Classifier selection; Linear discriminant analysis; Boosting; Prototype selection; Dimensionality reduction

1. Introduction

This paper deals with feature extraction applied to nearest neighbor classification. Feature extraction allows a compact representation of the input data, due to the dimensionality reduction achieved during the process, what increases the performance of the global scheme reducing the storage needs and the computational costs. In our case we have focused on discriminant analysis techniques, which take into account class membership of the input data, learning invariant characteristics that increase the classification ratios.

Maybe one of the first attempt to dimensionality reduction applied to classification is principal component analysis [1,2] where the goal is to find the linear projection matrix that preserves the maximum amount of input data variance. In discriminant analysis the labels are also considered in the linear feature extraction process, and the goal is to find the orthogonal set of basis that maximizes some separability criteria. The main problem of linear discriminant algorithms is their dependency on a set of assumptions that sometimes are not met in the data distribution [3].

Last years some nonlinear algorithms applied to feature extraction have appeared. Tenenbaum et al. [4] introduced the isomap algorithm, which tries to preserve the geodesic distances between points in the low-dimensional embedding. Roweis et al. [5] introduced a new nonlinear technique that preserves the local neighborhood of each point in the embedding process. The nonlinear nature of both techniques allows to represent the manifold that underlay the training samples, but there are some difficulties using both algorithms with new unseen input vectors. Also the features extracted using this nonlinear techniques cannot be ranked in order of importance for classification purposes.

What we purpose here is an embedding from high-dimensional space to a low-dimensional one, where the features are ranked according to coefficients computed within the algorithm. Also we have not made assumptions

[*] Corresponding author. Tel.: +34 93 581 1828.
E-mail addresses: davidm@cvc.uab.es (D. Masip), jordi@cvc.uab.es (J. Vitrià).

0031-3203/$30.00 © 2005 Pattern Recognition Society. Published by Elsevier Ltd. All rights reserved.
doi:10.1016/j.patcog.2005.06.004

on the data distributions, and we do not force our projection to be orthogonal [6]. Our embedding combines a set of simple 1D projections, which can complement each other to achieve better classification results. We have made use of AdaBoost algorithm as a natural way to select the feature extractors, and the coefficients that can rank the importance of each projection.

2. Feature extraction for classification

The main goal of this work is to find a mapping from a high-dimensional space to new one that optimizes a discriminability criteria on the input data that is suited for nearest neighbor classification. Discriminant analysis can be very useful for this task. In this section we will review the classic Fisher discriminant analysis (FLD), and an evolution of the algorithm introduced by Fukunaga and Mantock [7], the non parametric discriminant analysis (NDA), which improves the classification results by using the nearest neighbor classifier and also overcomes the two main drawbacks of FLD:

- Gaussian assumption over the class distribution of the data samples.
- Dimensionality of the subspaces obtained which is limited by the number of classes.

2.1. Discriminant analysis

2.1.1. Fisher discriminant analysis

The objective of discriminant analysis is to find the features that best separate the different classes. One of the most used criterions $\mathscr{J}$ to reach is to maximize

$$\mathscr{J} = \operatorname{tr}(\boldsymbol{S}^E \boldsymbol{S}^I), \tag{1}$$

where the matrices $\boldsymbol{S}^E$ and $\boldsymbol{S}^I$ generally represent the scatter of sample vectors between different classes and within a class respectively. It has been shown (see Refs. [8,9]) that the $M \times D$ linear transform that satisfies

$$\hat{\boldsymbol{W}} = \arg \max_{\boldsymbol{W}^{\mathrm{T}} \boldsymbol{S}^I \boldsymbol{W} = \boldsymbol{I}} \operatorname{tr}(\boldsymbol{W}^{\mathrm{T}} \boldsymbol{S}^E \boldsymbol{W}) \tag{2}$$

optimizes the separability measure $\mathscr{J}$. This problem has an analytical solution based on the eigenvectors of the scatter matrices. The algorithm presented in Table 1 obtains this solution [9].

The most widely spread approach for defining the within and between class scatter matrices is the one that makes use of only up to second-order statistics of the data. This was proposed in a classic paper by Fisher [3] and the technique is referred to as FLD. In FLD the within class scatter matrix is usually computed as a weighted sum of the class-conditional sample covariance matrices. If equiprobable priors are assumed for classes C_k, $k = 1, \ldots, K$, then

$$\boldsymbol{S}^I = \frac{1}{K} \sum_{k=1}^{K} \boldsymbol{\Sigma}_k, \tag{3}$$

where $\boldsymbol{\Sigma}_k$ is the class-conditional covariance matrix, estimated from the sample set. The between class-scatter matrix is defined by

$$\boldsymbol{S}^E = \frac{1}{K} \sum_{k=1}^{K} (\boldsymbol{\mu}_k - \boldsymbol{\mu}_0)(\boldsymbol{\mu}_k - \boldsymbol{\mu}_0)^{\mathrm{T}}, \tag{4}$$

where $\boldsymbol{\mu}_k$ is the class-conditional sample mean and $\boldsymbol{\mu}_0$ is the unconditional (global) sample mean.

Notice the rank of $\boldsymbol{S}^E$ is $K-1$, so the number of extracted features is, at most, one less than the number of classes. Also notice the parametric nature of the scatter matrix. The solution provided by FLD is blind beyond second-order statistics, so we cannot expect this method to accurately indicate which features should be extracted to preserve any complex classification structure.

Table 1
General algorithm for solving the discriminability optimization problem stated in Eq. (2)

(1)	Given $\boldsymbol{X}$ the matrix containing data samples placed as N D-dimensional columns, $\boldsymbol{S}^I$ the within class scatter matrix, and M maximum dimension of discriminant space.
(2)	Compute eigenvectors and eigenvalues for $\boldsymbol{S}^I$. Make $\boldsymbol{\Phi}$ the matrix with the eigenvectors placed as columns and $\boldsymbol{\Lambda}$ the diagonal matrix with only the nonzero eigenvalues in the diagonal. M^I is the number of non-zero eigenvalues.
(3)	Whiten the data with respect to $\boldsymbol{S}^I$, to obtain M^I dimensional whitened data, $\boldsymbol{Z} = \boldsymbol{\Lambda}^{-1/2} \boldsymbol{\Phi}^{\mathrm{T}} \boldsymbol{X}$.
(4)	Compute $\boldsymbol{S}^E$ on the whitened data.
(5)	Compute eigenvectors and eigenvalues for $\boldsymbol{S}^E$ and make $\boldsymbol{\Psi}$ the matrix with the eigenvectors placed as columns and sorted by decreasing eigenvalue value.
(6)	Preserve only the first $M^E = \min\{M^I, M, \operatorname{rank}(\boldsymbol{S}^E)\}$ columns, $\boldsymbol{\Psi}_M = \{\psi_1, \ldots, \psi_{M^E}\}$ (those corresponding to the M^E largest eigenvalues).
(7)	The resulting optimal transformation is $\hat{\boldsymbol{W}} = \boldsymbol{\Psi}_M^{\mathrm{T}} \boldsymbol{\Lambda}^{-1/2} \boldsymbol{\Phi}^{\mathrm{T}}$ and the projected data, $\boldsymbol{Y} = \hat{\boldsymbol{W}} \boldsymbol{X} = \boldsymbol{\Psi}_M^{\mathrm{T}} \boldsymbol{Z}$.