Pattern Recognition Letters 49 (2014) 85–91

Contents lists available at ScienceDirect

Pattern Recognition Letters

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

# Kernel Reference Discriminant Analysis ☆

Alexandros Iosifidis *, Anastastios Tefas, Ioannis Pitas

*Department of Informatics, Aristotle University of Thessaloniki, Thessaloniki 54124, Greece*

ARTICLE INFO

*Article history:*
Received 20 December 2013
Available online 9 July 2014

*Keywords:*
Kernel Discriminant Analysis
Kernel Spectral Regression
Optimized class representation

ABSTRACT

Linear Discriminant Analysis (LDA) and its nonlinear version Kernel Discriminant Analysis (KDA) are well-known and widely used techniques for supervised feature extraction and dimensionality reduction. They determine an optimal discriminant space for (non)linear data projection based on certain assumptions, e.g. on using normal distributions (either on the input or in the kernel space) for each class and employing class representation by the corresponding class mean vectors. However, there might be other vectors that can be used for classes representation, in order to increase class discrimination in the resulted feature space. In this paper, we propose an optimization scheme aiming at the optimal class representation, in terms of Fisher ratio maximization, for nonlinear data projection. Compared to the standard approach, the proposed optimization scheme increases class discrimination in the reduced-dimensionality feature space and achieves higher classification rates in publicly available data sets.



## 1. Introduction

Linear Discriminant Analysis (LDA) is a well-known algorithm for supervised feature extraction and dimensionality reduction. It aims at the determination of an optimal subspace for linear data projection, in which the classes are better discriminated. Non-linear extensions [12,25,24,3,22,16,5] exploit data representations in arbitrary-dimensional feature spaces (determined by applying a non-linear data mapping process). After the determination of the data representation in the arbitrary-dimensional feature space, a linear projection is calculated, which corresponds to a non-linear projection of the original data. In both cases, the adopted criterion is the ratio of the between-class scatter to the within-class scatter in the reduced-dimensionality feature space, which is usually referred to as Fisher ratio.

LDA optimality is based on the assumptions of: (a) normal class distributions with the same covariance structure and (b) class representation by the corresponding class mean vector. Under these assumptions, the maximization of the Fisher ratio leads to maximal class discrimination in the reduced-dimensionality feature space. Although relying on rather strong assumptions, both LDA and its kernel extensions have proven very powerful and they have been widely used in many applications, including face recognition/verification [28,27,7,14,30], human action recognition [8,9], person identification [21,10,29] and speech recognition [4].

By observing that the between-class and within-class scatter matrices employed for the determination of the optimal data projection matrix in LDA can be considered to be functions of the class representation, it has been recently shown that, when the two aforementioned assumptions are not met, the adoption of class representations different from the class mean vectors leads to increased class discrimination in the reduced-dimensionality feature space [11]. In addition, it has been shown that, given a data projection matrix determined by maximizing the criterion adopted in LDA, the optimal class representations can be analytically calculated. In order to determine both the optimal data projection matrix and the optimal class representations, an iterative optimization scheme has been proposed [11].

In this paper, we extend the method in [11] in order to operate in arbitrary-dimensional feature spaces for non-linear supervised feature extraction and data projection. We formulate an optimization problem that exploits a non-linear data mapping process to an arbitrary-dimensional feature space, in which optimized class representations are determined. By employing such optimized class representations, a linear data projection from the arbitrary-dimensional feature space to a reduced-dimensionality discriminant feature space is subsequently calculated. We show that, the determination of the optimal class representation in the arbitrary-dimensional feature space has a closed form solution, similar to the linear case. For the determination of the optimal data projection exploiting the optimal class representations, we introduce the proposed criterion to the Spectral Regression framework [3] and we describe an efficient algorithm to this end. Finally, we combine the two aforementioned optimization processes and propose an

☆ This paper has been recommended for acceptance by J. Yang.

* Corresponding author. Tel./fax: +30 2310996304.
*E-mail addresses:* aiosif@aiia.csd.auth.gr, iosifidis.alekos@gmail.com (A. Iosifidis).

http://dx.doi.org/10.1016/j.patrec.2014.06.013
0167-8655/

iterative optimization scheme for the determination of both the optimal class representation and the optimal (non-linear) data projection. The proposed criterion is evaluated on standard classification problems, as well as on human action and face recognition problems. It is shown that, by exploiting optimized class representations, increased class discrimination can be achieved in the decision space leading to enhanced classification performance.

The rest of the paper is structured as follows. We briefly describe the non-linear version of LDA, i.e. the Kernel Discriminant Analysis (KDA), in Section 2. The proposed Kernel Reference Discriminant Analysis (KRDA) algorithm is described in detail in Section 3. Experimental results comparing its performance with the standard approach are provided in Section 4. Finally, conclusions are drawn in Section 5.

## 2. Kernel Discriminant Analysis

Let us denote by $\mathbf{x}_{ij} \in \mathbb{R}^D,\ i = 1,\ldots,C,\ j = 1,\ldots,N_i$ a set of $D$-dimensional data, each belonging to one of $C$ classes. The number of samples belonging to class $i$ is equal to $N_i$. In order to determine a nonlinear data projection, the input space $\mathbb{R}^D$ is mapped to an arbitrary-dimensional feature space $\mathcal{F}$ (usually having the properties of Hilbert spaces) [19,2,1] by employing a function $\phi(\cdot) : \mathbf{x}_{ij} \in \mathbb{R}^D \rightarrow \phi(\mathbf{x}_{ij}) \in \mathcal{F}$ determining a nonlinear mapping from the input space $\mathbb{R}^D$ to the arbitrary-dimensional feature space $\mathcal{F}$. $\phi(\cdot)$ can either be chosen based on the properties of the problem at hand, e.g. for histogram-based data representations the RBF-$\chi^2$ kernel has been proven to be the state-of-the-art choice [31], or can be determined by applying kernel selection methods. In the second case, a linear combination of a priori chosen kernel functions is usually learned based on optimization, e.g. as in [13]. In $\mathcal{F}$, we would like to determine a data projection matrix $\mathbf{P}$ that can be used to map a given sample $\phi(\mathbf{x}_{ij})$ to a low-dimensional feature space $\mathbb{R}^d$ of increased class discrimination power:

$$\mathbf{y}_{ij} = \mathbf{P}^T\phi(\mathbf{x}_{ij}), \quad \mathbf{y}_{ij} \in \mathbb{R}^d. \tag{1}$$

This can be achieved by maximizing the following criterion:

$$\mathcal{J}_{KDA}(\mathbf{P}) = \frac{trace(\mathbf{P}^T\mathbf{S}_b\mathbf{P})}{trace(\mathbf{P}^T\mathbf{S}_w\mathbf{P})}, \tag{2}$$

where the matrices $\mathbf{S}_b,\ \mathbf{S}_w$ are given by:

$$\mathbf{S}_b = \sum_{i=1}^{C} N_i(\phi(\mathbf{m}_i) - \phi(\mathbf{m}))(\phi(\mathbf{m}_i) - \phi(\mathbf{m}))^T, \tag{3}$$

$$\mathbf{S}_w = \sum_{i=1}^{C}\sum_{j=1}^{N_i}\big(\phi(\mathbf{x}_{ij}) - \phi(\mathbf{m}_i)\big)\big(\phi(\mathbf{x}_{ij}) - \phi(\mathbf{m}_i)\big)^T. \tag{4}$$

In (3) and (4), $\phi(\mathbf{m}_i)$ is the mean vector of class $i$ in $\mathcal{F}$, i.e. $\phi(\mathbf{m}_i) = \frac{1}{N_i}\sum_{j=1}^{N_i}\phi(\mathbf{x}_{ij})$. $\phi(\mathbf{m})$ is the mean vector of the entire set in $\mathcal{F}$, i.e. $\phi(\mathbf{m}) = \frac{1}{N}\sum_{i=1}^{C}\sum_{j=1}^{N_i}\phi(\mathbf{x}_{ij})$, where $N = \sum_{i=1}^{C}N_{ij}$. The direct maximization of (2) is intractable, since $\mathbf{S}_b,\ \mathbf{S}_w$ are matrices with arbitrary (possibly infinite) dimensions. In practice we overcome this problem by exploiting the so-called kernel trick [19,2,1]. That is, the maximization of (2), as well as the multiplication in (1), are inherently computed by using dot-products in $\mathcal{F}$.

The maximization of (2) with respect to $\mathbf{P}$ leads to the determination of a data projection that can be used to map the original data to a reduced-dimensionality feature space where the data dispersion from the corresponding class mean vectors is minimized and the dispersion of class mean vectors from the total mean is maximized. In the cases where the classes (when represented in $\mathcal{F}$) follow normal distributions with the same covariance structure, by maximizing (2) maximal class discrimination can be achieved. However, this is a strong assumption which may not be met in many real problems. As has been shown in [11], the determination of optimized class representations enhances class discrimination in the projection space in the cases where the assumptions of LDA are not met. In the following, we describe an iterative optimization scheme that can be exploited in order to determine both the optimal class representations in $\mathcal{F}$ and the optimal projection for non-linear data mapping exploiting such optimized representations.

## 3. Kernel Reference Discriminant Analysis

In this Section we describe in detail the proposed Kernel Reference Discriminant Analysis algorithm. Let us denote by $\mathbf{\Phi}_i$ a matrix containing the samples belonging to class $i$ (represented in $\mathcal{F}$), i.e. $\mathbf{\Phi}_i = [\phi(\mathbf{x}_{i1}),\ldots,\phi(\mathbf{x}_{iN_i})]$. By using $\mathbf{\Phi}_i,\ i = 1,\ldots,C$ we can construct the matrix $\mathbf{\Phi} = [\mathbf{\Phi}_1,\ldots,\mathbf{\Phi}_C]$ containing the representations of the entire data set in $\mathcal{F}$. The so-called kernel matrix $\mathbf{K} \in \mathbb{R}^{N\times N}$ is given by $\mathbf{K} = \mathbf{\Phi}^T\mathbf{\Phi}$. Let us denote by $\mathbf{K}_i \in \mathbb{R}^{N\times N_i}$ a matrix containing the columns of $\mathbf{K}$ corresponding to the samples belonging to class $i$. That is, $\mathbf{K} = [\mathbf{K}_1,\ldots,\mathbf{K}_C]$, where $\mathbf{K}_i = \mathbf{\Phi}^T\mathbf{\Phi}_i$.

In KRDA, each class $i$ is represented by a vector $\phi(\boldsymbol{\mu}_i)$. $\phi(\boldsymbol{\mu}_i)$ is not restricted to be the class mean in $\mathcal{F}$, but can be any vector enhancing class discrimination in the projection space $\mathbb{R}^d$. In order to determine both the optimal data projection matrix $\mathbf{P}$ and the optimal class representations $\phi(\boldsymbol{\mu}_i)$, we propose to maximize the following criterion with respect to both $\mathbf{P}$ and $\boldsymbol{\mu}_i$:

$$\mathcal{J}_{KRDA}(\mathbf{P},\boldsymbol{\mu}_i) = \frac{trace(\mathbf{P}^T\tilde{\mathbf{S}}_b(\boldsymbol{\mu}_i)\mathbf{P})}{trace(\mathbf{P}^T\tilde{\mathbf{S}}_w(\boldsymbol{\mu}_i)\mathbf{P})}, \tag{5}$$

where the matrices $\tilde{\mathbf{S}}_b(\boldsymbol{\mu}_i),\ \tilde{\mathbf{S}}_b(\boldsymbol{\mu}_i)$ are given by:

$$\tilde{\mathbf{S}}_b(\boldsymbol{\mu}_i) = \sum_{i=1}^{C} N_i\big(\phi(\boldsymbol{\mu}_i) - \phi(\mathbf{m})\big)\big(\phi(\boldsymbol{\mu}_i) - \phi(\mathbf{m})\big)^T, \tag{6}$$

$$\tilde{\mathbf{S}}_w(\boldsymbol{\mu}_i) = \sum_{i=1}^{C}\sum_{j=1}^{N_i}\big(\phi(\mathbf{x}_{ij}) - \phi(\boldsymbol{\mu}_i)\big)\big(\phi(\mathbf{x}_{ij}) - \phi(\boldsymbol{\mu}_i)\big)^T. \tag{7}$$

$\tilde{\mathbf{S}}_w$ describes the class dispersion with respect to $\phi(\boldsymbol{\mu}_i)$ in $\mathcal{F}$. That is, the maximization of (5) leads to the determination of a data projection that can be used to map the original data to a reduced-dimensionality feature space $\mathbb{R}^d$, where the data dispersion from the corresponding class reference vector $\tilde{\boldsymbol{\mu}}_i = \mathbf{P}^T\phi(\boldsymbol{\mu}_i)$ is minimized, while the dispersion of the class reference vectors from the total mean is maximized. In the following, we assume that the data set is centered in $\mathcal{F}$.[1]

### 3.1. Calculation of $\mathbf{P}$

In order to determine the optimal data projection matrix $\mathbf{P}$ we work as follows. Let us denote by $\mathbf{p}$ an eigenvector of the problem $\tilde{\mathbf{S}}_b(\boldsymbol{\mu}_i)\mathbf{p} = \lambda\tilde{\mathbf{S}}_w(\boldsymbol{\mu}_i)\mathbf{p}$ with eigenvalue $\lambda$. $\mathbf{p}$ can be expressed as a linear combination of the data (representated in $\mathcal{F}$) [19,2,1], i.e. $\mathbf{p} = \sum_{i=1}^{C}\sum_{j=1}^{N_i}a_{ij}\phi(\mathbf{x}_{ij}) = \mathbf{\Phi}\mathbf{a}$, where $\mathbf{a} \in \mathbb{R}^N$. In addition, we can express $\phi(\boldsymbol{\mu}_i)$ as a linear combination of the samples belonging to class $i$, i.e. $\phi(\boldsymbol{\mu}_i) = \sum_{j=1}^{N_i}b_{ij}\phi(\mathbf{x}_{ij}) = \mathbf{\Phi}_i\mathbf{b}_i$, where $\mathbf{b}_i \in \mathbb{R}^{N_i}$. As it will be described in Appendix A, by setting $\mathbf{Ka} = \mathbf{u}$, the aforementioned eigenproblem can be transformed to the following equivalent eigenproblem:

$$\mathbf{B}(\mathbf{b}_i)\mathbf{u} = \lambda\mathbf{W}(\mathbf{b}_i)\mathbf{u}. \tag{8}$$

[1] This can always be done by using $\tilde{\phi}(\mathbf{x}_{ij}) = \phi(\mathbf{x}_{ij}) - \phi(\mathbf{m})$, leading to a centered version of the kernel matrix given by $\tilde{\mathbf{K}} = \frac{1}{N}\mathbf{K1} - \frac{1}{N}\mathbf{1K} + \frac{1}{N^2}\mathbf{1K1}$, where $\mathbf{1} \in \mathbb{R}^{N\times N}$ is a matrix of ones.