Contents lists available at ScienceDirect

Neurocomputing

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

Discriminant analysis via support vectors

Suicheng Gu [a,b], Ying Tan [a,b,*], Xingui He [a,b]

[a] Key Laboratory of Machine Perception (MOE), Peking University, Beijing 100871, PR China
[b] Department of Machine Intelligence, School of Electronics Engineering and Computer Science, Peking University, Beijing 100871, PR China

ARTICLE INFO

Available online 12 March 2010

Keywords:
Discriminant analysis via support vectors
Fisher's criteria
Linear discriminant analysis
Regularized discriminant analysis

ABSTRACT

In this paper, we show how support vector machine (SVM) can be employed as a powerful tool for *k*-nearest neighbor (kNN) classifier. A novel multi-class dimensionality reduction approach, discriminant analysis via support vectors (SVDA), is proposed. First, the SVM is employed to compute an optimal direction to discriminant each two classes. Then, the criteria of class separability is constructed. At last, the projection matrix is computed. The kernel mapping idea is used to derive the non-linear version, kernel discriminant via support vectors (SVKD). In SVDA, only support vectors are involved to compute the transformation matrix. Thus, the computational complexity can be greatly reduced for kernel based feature extraction. Experiments carried out on several standard databases show a clear improvement on LDA-based recognition.

1. Introduction

The *k*-nearest neighbors (kNN) [1] rule is one of the oldest and simplest methods for pattern classification. Feature extraction (dimensionality reduction) are often employed in helping kNN classifier to reduce computational complexity and improve classification accuracy.

The generic problem of linear dimensionality reduction is the following. Given a dataset $X=(x_1,x_2,\ldots,x_N)\in\mathcal{R}^{n\times N}$, find a transformation matrix $A=(a_1,\ldots,a_k)\in\mathcal{R}^{n\times k}$ that maps these N points to a set of points $Z=(z_1,z_2,\ldots,z_N)\in\mathcal{R}^{k\times N}$, such that z_i represents x_i, where $z_i=A^Tx_i$.

1.1. PCA and LDA

Principal component analysis (PCA) [2], also known as Karhunen–Loeve expansion, is a classical feature extraction and data representation technique widely used in the areas of pattern recognition and computer vision. Due to its simplicity and effectiveness, many variants of PCA were developed [3–5].

Linear discriminant analysis (LDA) [6], or called Fisher's linear discriminant (FLD), for feature extraction has been applied to a wide variety of problems such as face recognition. It often produces much better results than PCA. However, in practice, the LDA has three major problems: (1) It suffers from the small sample size (SSS) problem when dimensionality is greater than the sample size. (2) It creates subspaces that favor well separated classes over those that are not. (3) LDA assumes the data obey normal distribution. And it simply uses $\mu_a-\mu_c$ to discriminate two classes ω_a and ω_c. It fails to obtain the optimal direction to separate two classes.

Many algorithms tried to alleviate one or two of the problems in LDA. The regularized discriminant analysis (RDA) [7] added a multiple of identify matrix to the within-class matrix with regard to the small sample size problem. Another well-known approach is the Fisherface [8], in which LDA is employed after the PCA is used. Another technique, newLDA [9], first transforms the data into the null space of S_w. It then applies PCA to maximize the between-class scatter matrix in the transformed space.

1.2. Local learning

More recent years, many manifold (graph) based methods are implemented to preserve the local information and obtain a new subspace [10,11]. Some popular ones include: discriminant locally linear embedding (DLLE) [12], geometric mean for subspace selection (MGMD) [13], harmonic mean for subspace selection (MHMD) [14], discriminative locality alignment [15], transductive component analysis (TCA) [16], locality preserving projection (LPP) [17], marginal Fisher analysis (MFA) [18] and locality sensitive discriminant analysis (LSDA) [19], etc. To learn more about local learning methods, one can refer to [11].

1.3. Margin based discriminant

Large margin nearest neighbor (LMNN) [20] learns a Mahanalobis distance metric for kNN classification by semidefinite programming. Large margin component analysis (LMCA) [21]

[*] Corresponding author at: Key Laboratory of Machine Perception (MOE), Peking University, Beijing 100871, PR China.
E-mail address: ytan@pku.edu.cn (Y. Tan).

0925-2312/$ - see front matter © 2010 Elsevier B.V. All rights reserved.
doi:10.1016/j.neucom.2009.09.021

solves for a low-dimensional embedding of the data such that Euclidean distance in this space minimizes the large margin metric objective described in [20]. Yuan and Pang [22] iteratively selects a series of simple but effective 1D subspaces, and then combines the corresponding 1D projections by Adaboost.

Support vector machine (SVM) [23] is based on the statistical learning theory of Vapnik and quadratic programming learning theory. The superior classification performance of SVM has been justified in numerous experiments, particularly in high dimensionality and small sample size (SSS) problems. Bi et al. [24] described a methodology for performing variable ranking and selection using support vector machines (SVMs). Margin maximizing discriminant analysis (MMDA) [25] attempted to preserve as much discriminant information as possible by projecting the dataset onto margin maximizing directions (separating hyperplane normals) found by an SVM algorithm. The corresponding normal vectors of the hyperplanes are taken as new features and the data are projected onto them. The first MMDA feature is obtained by simply using the standard SVM. Then, after obtaining orthogonal MMDA features, the second feature is found by optimizing the SVM in the remaining feature subspace. It is intrinsically a two-class approach.

In this paper, we developed a supervised dimensionality reduction approach for multiple-class problems, by employing SVM. To make a contrast with LDA, we call this approach discriminant analysis via support vectors (SVDA). Both linear and nonlinear models, discriminant analysis via support vectors (SVDA) and kernel discriminant via support vectors (SVKD), are described.

The rest of this paper is organized as follows. In Section 2, the LDA and SVM are reviewed briefly. In Section 3, the proposed SVDA algorithm is introduced. We describe how to perform SVDA in reproducing kernel Hilbert space (RKHS) which gives rise to kernel SVDA in Section 4. The experimental results are presented in Section 5. Finally, a conclusion is given in Section 6.

Notation conventions used in this paper:

i,N	counter and number of training samples;
n	dimension of training samples;
X	training samples with size of $n \times N$;
φ	$\mathcal{R}^n \rightarrow \mathcal{F}$;
$\mathcal{K}$	$\mathcal{K}(x_i,x_j) = \langle \varphi(x_i),\varphi(x_j) \rangle$;
K	kernel matrix, $K_{i,j} = \mathcal{K}(x_i,x_j)$;
a,M	counter and number of classes;
μ_a	mean vector of class ω_a;
N_a	number of samples in class ω_a;
I_a	collection of sample indexes in class ω_a;

2. LDA and SVM

2.1. LDA

In LDA, within-class and between-class scatter matrices are used to formulate the criteria of class separability. A within-class scatter matrix characterizes the scatter of samples around their respective class mean vectors, and it is expressed by

$$S_w = \sum_{a=1}^{M} \sum_{i \in I_a} (x_i - \mu_a)(x_i - \mu_a)^T. \tag{1}$$

A between-class scatter matrix characterizes the scatter of the class means around the mixture mean μ. It is expressed by

$$S_b = \sum_{a=1}^{M} N_a (\mu_a - \mu)(\mu_a - \mu)^T. \tag{2}$$

Linear discriminant analysis (LDA) seeks directions that are efficient for discrimination. Fisher criterion is used to find the projection matrix and the objective function of LDA is

$$a_{opt} = \arg\max_a \frac{a^T S_b a}{a^T S_w a}. \tag{3}$$

One can solve the generalized eigenvalue problem:

$$S_b a = \lambda S_w a. \tag{4}$$

2.1.1. RDA

In practice, the small sample size (SSS) problem is often encountered, where S_w is singular. Therefore, the maximization problem can be difficult to solve. To address this issue, the term εI is added, where ε is a small positive number and I is the identity matrix of proper size. This results in maximizing

$$a_{opt} = \arg\max_a \frac{a^T S_b a}{a^T (S_w + \varepsilon I) a}. \tag{5}$$

This is a special case of Friedman regularized discriminant analysis with regard to the small sample size problem [7].

2.2. SVM

Generally, an SVM [23] solves a binary (two-class) classification problem, and multi-class classification is accomplished by combining multiple binary SVMs. An M-class problem can be decomposed into M binary problems with each separating one class from the others, or into $M(M-1)/2$ binary problems with each discriminating between a pair of classes. On a pattern x, the discriminant function of a binary SVM is given by

$$f(x) = \sum_{i=1}^{l} y_i \alpha_i \mathcal{K}(x,x_i) + b, \tag{6}$$

where l is the number of learning patterns, y_i is the target value of learning pattern x_i (+1 for the first class and -1 for the second class), b is a bias, and $\mathcal{K}(x,x_i)$ is a kernel function which implicitly defines an expanded feature space:

$$\mathcal{K}(x,x_i) = \varphi(x) \cdot \varphi(x_i), \tag{7}$$

where $\varphi(x)$ is the feature vector in the expanded feature space and may have infinite dimensionality. Several popular kernels are: linear kernel $K(x_i, x_j) = x_i^T x_j$; polynomial kernel $K(x_i, x_j) = (1 + x_T^i x_j)^p$ and RBF kernel $K(x_i,x_j) = \exp(-\|x_i - x_j\|^2/\sigma^2)$.

The discriminant function of Eq. (6) can be viewed as a generalized linear discriminant function with weight vector

$$w = \sum_{i=1}^{l} y_i \alpha_i \varphi(x_i). \tag{8}$$

The coefficients α_i ($i = 1,2,\ldots,l$) are determined according to the learning patterns by solving the following optimization problem:

$$\text{Minimize } \tau(w) = \frac{1}{2}\|w\|^2 + C \sum_{i=1}^{l} \zeta_i$$

subject to $y_i f(x_i) \geq 1 - \zeta_i$ and $\zeta_i \geq 0$, $i = 1,2,\ldots,l$.

This is a quadratic programming problem and can be converted into the following dual problem:

$$\begin{aligned} &\text{Minimize} \quad Q(\alpha) = \sum_{i=1}^{l} \alpha_i - \frac{1}{2} \sum_{i=1}^{l} \alpha_i \alpha_j y_i y_j \mathcal{K}(x_i,x_j) \\ &\text{subject to} \quad 0 \leq \alpha_i \leq C, \quad i = 1,2,\ldots,l, \\ &\qquad \text{and} \quad \sum_{i=1}^{l} \alpha_i y_i = 0, \end{aligned} \tag{9}$$

where C (default $C = 100$) is a parameter to control the tolerance of classification errors in learning.