European Journal of Operational Research 203 (2010) 692–697

Contents lists available at ScienceDirect

European Journal of Operational Research

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

Computational Intelligence and Information Management

Learning the optimal kernel for Fisher discriminant analysis via second order cone programming

Reshma Khemchandani [a], Jayadeva [b], Suresh Chandra [a,*]

[a] Department of Mathematics, Indian Institute of Technology, Hauz Khas, New Delhi 110016, India
[b] IBM India Research Lab, Block-C, Institutional Area Vasant Kunj, New Delhi 110070, India

ARTICLE INFO

Article history:
Received 16 March 2007
Accepted 15 September 2009
Available online 19 September 2009

Keywords:
Fisher discriminant analysis
Kernel methods
Machine learning
Kernel optimization
Support vector machines
Convex optimization
Second order cone programming
Semidefinite programming

ABSTRACT

Kernel Fisher discriminant analysis (KFDA) is a popular classification technique which requires the user to predefine an appropriate kernel. Since the performance of KFDA depends on the choice of the kernel, the problem of kernel selection becomes very important. In this paper we treat the kernel selection problem as an optimization problem over the convex set of finitely many basic kernels, and formulate it as a second order cone programming (SOCP) problem. This formulation seems to be promising because the resulting SOCP can be efficiently solved by employing interior point methods. The efficacy of the optimal kernel, selected from a given convex set of basic kernels, is demonstrated on UCI machine learning benchmark datasets.

1. Introduction

Kernel based methods have proved to be a powerful tool for solving classification problems. Recently, kernel Fisher discriminant analysis (KFDA) has received much attention in the literature (Mika et al., 1999, 2001; Yang et al., 2005). Kernel Fisher discriminant analysis requires the factorization of the Gram matrix into a within-class and a between-class scatter matrices, that are computed by using given training samples. The KFDA is computationally very simple, but its classification performance depends very much on the choice of the kernel function.

Generally, kernels are chosen by predefining a kernel model (Gaussian, polynomial, etc.), and adjusting the kernel parameters by a tuning procedure. The classifier's performance on a subset of the training data, commonly referred to as the validation set, is the main criterion for selection of the kernel. This kernel selection procedure can be computationally very expensive.

Support Vector Machines (SVMs) are powerful tools for classification problems. They have emerged from a research in statistical learning theory on how to regulate generalization in learning, and choose a tradeoff between structural complexity and empirical risk. SVMs classify points by assigning them to one of two disjoint half spaces, either in the pattern space or in a higher-dimensional feature space.

One of the most popular SVM classifiers is the "maximum margin" one, which aims at minimizing an upper bound on the generalization error through maximizing the margin between two parallel planes (Burges, 1998; Cortes et al., 1995; Vapnik, 1995) which are at a unit distance from the maximum margin classifier. Determining the classifier requires the minimization of a quadratic function subject to linear inequality constraints, which is a convex programming task. Working on the lines of SVMs, Mika et al. (2001) gave an alternative formulation of KFDA which also minimizes a convex quadratic function subject to linear inequality constraints.

In recent years, several authors (Bach et al., 2004; Bennett et al., 2002; Hamers et al., 2003; Lanckriet et al., 2004; Yang et al., 2005) have proposed the use of a non-negative linear combination of kernels formed by a family of different kernel functions and parameters. Here, the aim is to find an "optimal" linear combination of kernel functions chosen from the kernel family. Using this approach, the final kernel is constructed according to the specific classification problem to be solved, without sacrificing generalization performance, thereby avoiding the need to predefine a kernel. Thus, by combining kernels optimally and with appropriate regularization, the prediction accuracy is improved, which is the ultimate goal of classification.

* Corresponding author. Tel.: +91 11 26591479; fax: +91 11 26581005.
E-mail addresses: reshmaiitd@gmail.com (R. Khemchandani), jayadeva@ee.iitd.ac.in (Jayadeva), chandras@maths.iitd.ac.in (S. Chandra).

0377-2217/$ - see front matter © 2009 Elsevier B.V. All rights reserved.
doi:10.1016/j.ejor.2009.09.020

In recent work on optimal kernel selection, Fung et al. (2004) have developed an iterative method based on a quadratic programming formulation of the Fisher discriminant analysis given by Mika et al. (2001). Kim et al. (2006) have shown that the optimal kernel selection problem in KFDA can be reformulated as a semidefinite programming problem (SDP), which can be solved using interior point methods.

In this paper, we consider the problem of finding the data-dependent "optimal" kernel function via second order cone programming (SOCP). Taking motivation from Lanckriet et al. (2004), we define a generalized performance measure for the quadratic programming formulation of KFDA. We then formulate the kernel selection problem as a convex optimization problem over the cone of positive semidefinite matrices. The main contribution of the present work is to show that the latter problem can also be reformulated as a second order cone programming problem, which can be solved efficiently by employing interior point algorithms.

The paper is organized as follows: In Section 2 we summarize linear discriminant analysis and its kernel version. Section 3 briefly discusses some recent developments in the field of optimal kernel selection for KFDA. In Section 4, by introducing the generalized performance measure for KFDA, we propose a SOCP formulation for finding an optimal kernel over the set of positive semidefinite matrices. Experimental results are given in Section 5, while Section 6 is devoted to concluding remarks.

2. Linear discriminant analysis

Linear Fisher discriminant analysis (LFD) is a classification method that projects n-dimensional data onto a line, and performs classification in this one dimensional space. The projection is chosen so as to maximize the between-class mean, and minimize the within-class variance.

Let the patterns to be classified be denoted by a set of k vectors $X = [x_1, x_2, \ldots, x_k]$ in the n-dimensional real space $\mathbf{R}^n$, and let $y_i \in \{1, -1\}$ denote the class to which the ith pattern belongs. Let $X_1 = [x_1^1, x_2^1, \ldots, x_{k_1}^1]$ and $X_2 = [x_1^2, x_2^2, \ldots, x_{k_2}^2], (k = k_1 + k_2)$, be samples from two different classes with class labels +1 and -1. Then the Fisher's linear discriminant is given by the vector $z \in \mathbf{R}^n$ that maximizes the Fisher discriminant ratio (FDR)

$$J(z) = \frac{z^T S_B z}{z^T S_W z}, \tag{1}$$

where S_W is the within-class scatter matrix and S_B is the between-class scatter matrix. The matrices S_W and S_B are given by

$$S_B = (m_1 - m_2)(m_1 - m_2)^T, \tag{2}$$

$$S_W = \sum_{i=1}^{2} \sum_{x \in X_i} (x - m_i)(x - m_i)^T, \tag{3}$$

$$\text{where} \quad m_i = \frac{1}{k_i} \sum_{l=1}^{k_i} x_l^i, \quad i = 1, 2. \tag{4}$$

The intuition behind LFD is to find the direction that maximizes the projected class mean, while minimizing the class variance in this direction. Thus, the classifier is less prone to overfitting, and is therefore preferred over other complex classifiers.

To obtain the nonlinear version of the Fisher discriminant ratio, a nonlinear mapping $\phi : \mathbf{R}^n \to \mathbf{H}$ is considered, where $\mathbf{H}$ is some high dimensional feature space. Then the kernel Fisher discriminant u in $\mathbf{H}$ is obtained by maximizing

$$J(u) = \frac{u^T S_B^\phi u}{u^T S_W^\phi u}, \tag{5}$$

where the within-class scatter matrix S_W^ϕ and the between-class scatter matrix S_B^ϕ are given by

$$S_B^\phi = (m_1^\phi - m_2^\phi)(m_1^\phi - m_2^\phi)^T, \tag{6}$$

$$S_W^\phi = \sum_{i=1}^{2} \sum_{x \in X_i} (\phi(x) - m_i^\phi)(\phi(x) - m_i^\phi)^T, \tag{7}$$

$$\text{with} \quad m_i^\phi = \frac{1}{k_i} \sum_{l=1}^{k_i} \phi(x_l)^i, \quad i = 1, 2. \tag{8}$$

Mika et al. (2001) have recently shown that the above problem can be reformulated as the following quadratic programming problem:

$$\begin{aligned} (\text{KFSVM}) \quad & \min_{u,b,q} && \frac{C}{2}\|q\|^2 + \frac{1}{2}(u^T u + b^2) \\ & \text{subject to} && K(X, X^T)u - be + q = y, \end{aligned} \tag{9}$$

where $q \in \mathbb{R}^k$; $u \in \mathbb{R}^k; b \in \mathbb{R}$; e is a vector of ones of appropriate dimension; $C \geqslant 0$ is a parameter; $y = [y_1, y_2, \ldots, y_k]$ where, $y_i = 1 \text{if} x_i \in X_1$ and $y_i = -1 \text{if} x_i \in X_2$, and $K(X, X^T) = (k_{ij})$ is called the kernel matrix with $k_{ij} = \phi(x_i)^T \phi(x_j)$.

Once the optimal u and b are obtained, the class label of a new pattern $x \in \mathbf{R}^k$ (Vapnik, 1995) is decided as follows:

$$x \in \begin{cases} \text{class } 1; & \text{if } \left(K(x, X^T)u - b\right) > 0 \\ \text{class } -1; & \text{if } \left(K(x, X^T)u - b\right) \leqslant 0. \end{cases} \tag{10}$$

3. Some recent work on optimal kernel selection for KFDA

In this section, we give a very brief description of some of the recent work on optimal kernel selection for KFDA. In the literature, KFDA has been discussed with respect to a given kernel only, e.g. a Gaussian, a polynomial, etc. In Fung et al. (2004), it has been proposed that the kernel matrix $\mathbb{K}_F$ be taken as a non-negative linear combination of p given kernel matrices K_j, $(j = 1, \ldots, p)$, i.e., $\mathbb{K}_F = \sum_{j=1}^{p} \beta_j K_j$, where $\beta_j \geqslant 0, (j = 1, 2, \ldots, p)$. We have used the notation $\mathbb{K}_F$ to denote the optimal kernel obtained by using Fung et al.'s procedure.

The set $\{K_1, K_2, \ldots, K_p\}$ of given p kernel matrices could be viewed as a predefined set of p initial guesses of kernel matrices. As pointed out in Lanckriet et al. (2004), these kernels can be linear, Gaussian, polynomial, or all Gaussians with different hyper parameters.

Fung et al. (2004) developed an iterative method, termed as A-KFD, for selection of the optimal combination of kernels. The formulation of Fung et al. (2004) involves solving the following two optimization problems iteratively

$$\min_{u,b} \quad \frac{C}{2} \left\| y - \left(\sum_{j=1}^{p} \beta_j K_j u + be \right) \right\|^2 + \frac{1}{2}(u^T u + b^2) \tag{11}$$

and

$$\min_{\beta \geqslant 0} \quad \frac{C}{2} \left\| y - \left(\sum_{j=1}^{p} \beta_j K_j u + be \right) \right\|^2 + \frac{1}{2}(\beta^T \beta), \tag{12}$$

where, C, e and other symbols are as defined in the previous section. This method alternates between the problem of optimizing the decision vector (u, b) and the coefficients $\beta_j \geqslant 0, (j = 1, 2, \ldots, p)$. The first optimization problem deals with the optimization of u and b for a fixed β and the other deals with the optimization of β for fixed u and b as obtained from the first problem. Problems (11) and (12) are solved iteratively up to some predefined maximal number of iterations, or when there is a sufficiently small change in