Pattern Recognition Letters 54 (2015) 43–49

Contents lists available at ScienceDirect

Pattern Recognition Letters

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

Kernel optimization using nonparametric Fisher criterion in the subspace ☆

Xingyu Wu [a], Xia Mao [a], Lijiang Chen [a,*], Yuli Xue [a], Alberto Rovetta [b]

[a] School of Electronic and Information Engineering, Beihang University, Mailbox 206 Xueyuan Road, Beijing 100191, PR China
[b] Department of Mechanics, Polytechnic University of Milan, Via La Masa 34, Milan 20156, Italy

ARTICLE INFO

Article history:
Received 28 May 2014
Available online 23 December 2014

Keywords:
Dimensionality reduction
Kernel optimization
Fisher criterion

ABSTRACT

Kernel optimization plays an important role in kernel-based dimensionality reduction algorithms, such as kernel principal components analysis (KPCA) and kernel discriminant analysis (KDA). In this paper, a nonparametric Fisher criterion is proposed as the objective function to find the optimized kernel parameters. Unlike other criterions that rooted in the kernel feature space, the proposed criterion works in the low-dimensional subspace to measure the separability of different patterns. Experiments on 13 different benchmark datasets show the effectiveness of the proposed method, in comparison with other criterions and the kernel space methods.

1. Introduction

Kernel methods have been successfully applied in various classification tasks. These methods work by employing a nonlinear mapping $\phi(\cdot)$ from the original input space to an arbitrarily large or infinite dimensional kernel feature space F, and the linear algorithms can result in better performances in F compared to the original input space. As the mapping $\phi(\cdot)$ is often implicit, a kernel function is introduced to replace the dot product of two samples in F. Thus operations in the high-dimensional kernel feature space can be replaced by the kernel functions in the original input space. This is the essence of kernel methods [1].

The performance of kernel methods depends on the kernel function, which needs to satisfy the Mercer's condition [2]. Generally used kernel functions include the RBF kernel, sigmoid kernel, polynomial kernel, etc. When the kernel type is determined, the remaining problem becomes how to find the optimized kernel parameters, i.e., kernel optimization. If inappropriate kernel parameters are selected, the performance of kernel methods can be even worse than that of their linear counterparts.

As to the kernel optimization problem, various methods have been proposed. Currently the most employed technique is the k-fold cross validation (CV) [3], in which a large percentage of the data is used to train the kernel algorithm, and the remaining (smaller) percentage is employed to test how the classification accuracy varies when different kernel parameters are used. The parameters yielding the highest accuracy are kept. This method can be extended to optimize multiple parameters [4]. But CV only selects kernel parameters from a set of discrete values defined empirically, and brings large computation amount. Furthermore, it can only be performed when sufficient training samples are available. Thus, the CV method may fail to be applied on the Small Sample Size (SSS) problem [5,6].

Some other works provides alternatives to CV. Generalized cross validation (GCV) is an approximation to the leave-one-out CV but much faster [7]. The Bootstrap method [8] draws a series of subset from the training set to validate different parameter combinations. The Bootstrap method allows overlapping between the training set and the test set, so it is better suited to small set at the price of possible over fitting. Just like CV, these methods have to hold aside a portion of samples to validate the selected parameters iteratively. Some other works use in-sample methods that skip the validation process, and the test error is estimated by the training error analytically to select the optimized parameters. These methods evaluate the training model complexity and built the bias model to estimate the test error, and the estimation process varies depending on the loss function. Generally used in-sample methods include the squared error loss, Akaike information criterion (AIC) and Bayesian information criterion (BIC) [9].

Another kind of kernel optimization technique is the criterion-based methods, in which different criterions are used as object functions to measure the class separability, and the goal of optimization is then changed to find the kernel parameters that can maximize the object function. The Fisher criterion [10] is perhaps the most famous measure for pattern separability, which aims to maximize the ratio of

☆ This paper has been recommended for acceptance by G. Moser.

* Corresponding author. Tel.: +86 10 8231 6739; fax: +86 10 8233 9508.
E-mail address: moukyou@buaa.edu.cn (L. Chen).

http://dx.doi.org/10.1016/j.patrec.2014.11.016
0167-8655/

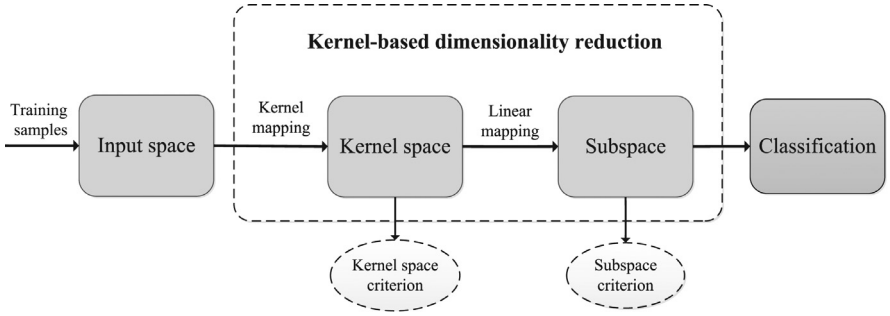

Fig. 1. An illustration of the kernel space criterion and subspace criterion in the kernel-based classification framework.

the between-class scatter matrix and the within-class scatter matrix. The Fisher criterion is used by Wang et al. [11], Kim et al. [12], Huang et al. [13] and Wang [14] in the kernel space to find the best kernel parameters. You et al. [15] modified the Fisher criterion by making the class Normal distributions in the kernel space most homoscedastic while maximizing class separability. Liu et al. [16] took the uniformity of class-pair separability into consideration along with the class separability. Wang et al. [17] focused on optimizing the Gaussian kernel, in which a decomposition of the Fisher criterion is used to derive the explicit expression of the objective function, and two different forms of the Fisher formulation are compared.

Current kernel optimization methods are rooted in the kernel feature space, because it is widely believed that an optimized kernel should maximize the pattern discriminant when input samples are projected to the kernel feature space [18]. However, when it comes to the classification task, the situation is slightly different. As shown in Fig. 1, the classification is actually performed in a low-dimensional subspace, which can be regarded as a linear projection of the kernel space on its principal axes (as to the principal components analysis based methods) or most discriminant axes (as to the discriminant analysis based methods). Compared with the kernel space, we are more interested in the spatial relationship between different patterns in the subspace, which is more relevant to the final classification accuracy. Meanwhile, the optimized kernel parameters depend on the subsequent dimensionality reduction algorithm, so it is suboptimal to search the kernel parameter without considering the subsequent processing. Furthermore, some non-kernel parameters may be involved in the dimensionality reduction algorithms. However, these parameters cannot be optimized by the kernel space criterions. In current researches [12,13,15–17,19], although the kernel parameters are optimized, the other parameters are simply set as empirical values. This deviates from the original purpose of optimization. Due to the above analysis, in this research we propose a criterion-based method rooted in the subspace, aiming to find the best parameters.

Another important problem is about how to construct a criterion to suit the subspace. The classical Fisher criterion only works when different patterns are linearly separable, and has the underlying constraint that all patterns need to satisfy the Gaussian condition. To alleviate these constraints, we propose a nonparametric Fisher criterion. The idea is motivated by the recent nonparametric discriminant analysis (NDA) researches [20–22]. The difference is that these works aim to get the best separated subspace, while our goal is to evaluate the separability of a given subspace. Therefore, we use a new strategy to deemphasize the overlapped pattern parts in the criterion.

The main contributions of the proposed method are two folds. Firstly, we propose the concept of subspace criterions to optimize the kernel in the dimensionality reduction process, and a detailed comparison is conducted with conventional and state-of-the-art methods to demonstrate its performance. Secondly, we propose a nonparametric Fisher criterion, which is tailored for separability measure in the subspace, and the effectiveness is showed both theoretically and by experiment.

The rest of the paper is organized as follows: In Section 2, firstly the Fisher criterion is briefly reviewed, and then the proposed nonparametric Fisher criterion is introduced in detail. In Section 3, the complexity of the proposed method is analyzed and compared with conventional approaches. In Section 4, experiments on thirteen different data sets are conducted to verify the proposed method. Finally, a conclusion is drawn in Section 5.

2. Main idea

2.1. The Fisher criterion for separability measure

Let $\mathbf{D} = \{\mathbf{D}_i\}_{i=1}^{C}$ be a training set of C classes, in which each class $\mathbf{D}_i = \{\mathbf{x}_{ij}\}_{j=1}^{N_i}$ consists of N_i samples, and $N = \sum_{i=1}^{C} N_i$ is the total sample number. Each sample $\mathbf{x}_{ij}$ belongs to a H-dimensional input space $\Re^H$. For the kernel-based methods, $\mathbf{x}_{ij}$ is then mapped to the high-dimensional kernel space F by $\phi(\cdot) : \Re^H \to F$. The *between-class* scatter matrix S_{B}^{ϕ} and *within-class* scatter matrix S_{W}^{ϕ} in the kernel space are defined as

$$S_{\mathrm{B}}^{\phi} = \sum_{i=1}^{C} N_i (m_i^{\phi} - m^{\phi})(m_i^{\phi} - m^{\phi})^{\mathrm{T}} \tag{1}$$

and

$$S_{\mathrm{W}}^{\phi} = \sum_{i=1}^{C} \sum_{j=1}^{N_i} (\phi(\mathbf{x}_{ij}) - m_i^{\phi})(\phi(\mathbf{x}_{ij}) - m_i^{\phi})^{\mathrm{T}} \tag{2}$$

where m_i^{ϕ} denotes the mean of the training samples from class i, and m^{ϕ} is the mean of all the training samples. As the mapping $\phi(\cdot)$ is implicit, it is difficult to calculate S_{B}^{ϕ} and S_{W}^{ϕ} directly, so *kernel trick* is used here to solve this problem. A kernel function $k(\mathbf{x}_1, \mathbf{x}_2)$ satisfies the condition $\phi(\mathbf{x}_1)^{\mathrm{T}}\phi(\mathbf{x}_2) = k(\mathbf{x}_1, \mathbf{x}_2)$. We define a Gram matrix $\mathbf{K}_{\mathrm{A,B}}$ whose element K_{ij} is $k(\mathbf{x}_i, \mathbf{x}_j)$, with the constraint that $\mathbf{x}_i \in A$ and $\mathbf{x}_j \in B$. The operator $sum(\cdot)$ denote the summation of all elements in a matrix. Thus the trace of S_{W}^{ϕ} and S_{B}^{ϕ} can be expressed as the explicit form just using the input space $\Re^H$ [14]:

$$\begin{aligned} \mathrm{tr}(S_{\mathrm{B}}^{\phi}) &= \sum_{i=1}^{C} N_i \| m_i^{\phi} - m^{\phi} \|^2 \\ &= \sum_{i=1}^{C} \frac{\mathrm{sum}(\mathbf{K}_{\mathbf{D}_i,\mathbf{D}_i})}{N_i} - \frac{\mathrm{sum}(\mathbf{K}_{\mathbf{D},\mathbf{D}})}{N} \end{aligned} \tag{3}$$

and

$$\begin{aligned} \mathrm{tr}(S_{\mathrm{W}}^{\phi}) &= \sum_{i=1}^{C} \sum_{j=1}^{N_i} \| \phi(\mathbf{x}_{ij}) - m_i^{\phi} \|^2 \\ &= \mathrm{tr}(\mathbf{K}_{\mathbf{D},\mathbf{D}}) - \sum_{i=1}^{C} \frac{\mathrm{sum}(\mathbf{K}_{\mathbf{D}_i,\mathbf{D}_i})}{N_i} \end{aligned} \tag{4}$$

The separability of different patterns in F is measured by anticipating that $\mathrm{tr}(S_{\mathrm{B}}^{\phi})$ is maximized while $\mathrm{tr}(S_{\mathrm{W}}^{\phi})$ is minimized. Based on this