Information Sciences 343–344 (2016) 79–93

Contents lists available at ScienceDirect

Information Sciences

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

Fisher-regularized support vector machine

Li Zhang[a,b,*], Wei-Da Zhou[c]

[a] School of Computer Science and Technology & Provincial Key Laboratory for Computer Information Processing Technology, Soochow University, Suzhou 215006, China
[b] Collaborative Innovation Center of Novel Software Technology and Industrialization, Soochow University, Suzhou 215123, Jiangsu, China
[c] AI Speech Ltd., Suzhou 215123, Jiangsu, China

ARTICLE INFO

Article history:
Received 1 December 2014
Revised 16 January 2016
Accepted 23 January 2016
Available online 30 January 2016

Keywords:
Fisher regularization
Support vector machine
Kernel trick
Maximum margin regularization
Fisher discriminant analysis

ABSTRACT

Support vector machine (SVM) and Fisher discriminant analysis (FDA) are two commonly used methods in machine learning and pattern recognition. A combined method of the linear SVM and FDA, called SVM/LDA (linear discriminant analysis), has been proposed only for the linear case. This paper generalizes this combined method to the nonlinear case from the view of regularization. A Fisher regularization is defined and incorporated into SVM to obtain a Fisher regularized support vector machine (FisherSVM). In FisherSVM, there are two regularization terms , the maximum margin regularization and Fisher regularization, which allow FisherSVM to maximize the classification margin and minimize the within-class scatter. Roughly speaking, FisherSVM can approximatively fulfill the Fisher criterion and obtain good statistical separability. This paper also discusses the connections of FisherSVM to graph-based regularization methods and the mathematical programming method to Kernel Fisher Discriminant Analysis (KFDA). It shows that FisherSVM can be thought of as a graph-based supervised learning method or a robust KFDA. Experimental results on artificial and real-world data show that FisherSVM has a promising generalization performance.

1. Introduction

Both support vector machine (SVM) and Fisher discriminant analysis (FDA) are commonly used, and each of them represents a great development in machine learning and pattern recognition. Due to their powerful abilities of generalization and nonlinear processing, SVMs have attracted substantial attention from researchers and practitioners over ten the last years [6,39,45]. Many methods related to SVM have been proposed [9,33,48]. SVM and its variants have been widely applied to pattern recognition [6,51], regression estimation [49,50], time series forecasting [24], credit rating analysis [20] and other uses. SVMs have successful applications due to the implementation of the structural risk minimization principle and the introduction of the kernel trick. SVMs can achieve good performance by using kernels satisfying Mercer's condition, such as Gaussian kernels and wavelet kernels [49,50]. So far, almost all linear learning methods can be generalized to corresponding nonlinear ones using the kernel trick [19]. The kernel trick also has other advantages. Using the representer theorem [38], the decision function of kernel learning methods can be concisely represented as a combination of kernels associated with samples. In addition, kernel learning can usually be cast into optimization problems. Typically, SVMs can be formulated as

[*] Corresponding author at: School of Computer Science and Technology & Provincial Key Laboratory for Computer Information Processing Technology, Soochow University, Suzhou 215006, China. Tel.: +86 15295659028.
E-mail address: zhangliml@suda.edu.cn, lizhang.ml@gmail.com (L. Zhang).

http://dx.doi.org/10.1016/j.ins.2016.01.053
0020-0255/© 2016 Elsevier Inc. All rights reserved.

convex quadratic programming problems. At present, there are a number of decomposition methods that are available to find solutions efficiently [21,34,53].

As a method of feature extraction, FDA, also known as linear discriminant analysis (LDA), tries to find an optimal subspace spanned by discriminant vectors such that the subspace projections of samples have a maximum separability [14]. In fact, FDA seeks one or more discriminant vectors by maximizing

$$\mathbf{w}^* = \arg\max_{\mathbf{w}} J(\mathbf{w}) = \frac{trace(\mathbf{w}^T\mathbf{S}_b\mathbf{w})}{trace(\mathbf{w}^T\mathbf{S}_w\mathbf{w})} \tag{1}$$

where $\mathbf{w}$ are discriminant vectors or projection vectors, $J(\mathbf{w})$ is also known as the Fisher criterion, $trace(\cdot)$ is the trace of a matrix, $\mathbf{S}_b$ and $\mathbf{S}_w$ denote the between-class and within-class scatter matrices, respectively. It is known that FDA is optimal under Bayesian decision theory for Gaussian distributed classes with identical covariance matrices. Actually, Cai et al. showed that FDA can be interpreted in the framework of a linear extension of graph embedding [7]. In other words, FDA can also be considered as a graph-based learning method. Other known graph-based methods include mincuts [5], Gaussian random fields and harmonic functions [54], local and global consistency [52], spectral graph transducers [22], Laplacian regularized Gaussian mixture model [18], the Laplacian probabilistic latent semantic indexing [8], Laplacian SVM [4,26] and spectral regression (SR) [7].

Recent developments on FDA focus on two issues. One is small sample size problems [15], that is, where the dimension of the data is larger than the number of samples. In the case of a small sample size problem, $\mathbf{S}_w$ is singular, and maximizing $J(\mathbf{w})$ is an ill-posed problem. A number of methods are proposed to find discriminant vectors for small sample size problems, such as the PCA+FDA method [43] and the null space method [11]. On the other hand, FDA can only extract linear discriminant vectors. To remedy this linear limitation, the kernel FDA (KFDA) [3,27,29] was developed by applying the kernel trick to FDA. KFDA can extract nonlinear discriminant features. Generally, KFDA has a better discriminant performance than FDA, as supported by many empirical results [3,13,25,29,30]. because high-dimensional kernel mapping is employed in KFDA, small sample size problems always occur. Thus, some methods for solving small sample size problems for FDA were also generalized for KFDA, such as CFKDA (KPCA+FDA) [47] and regularized kernel discriminant analysis [25].

As commonly used methods, performance comparisons of SVM and FDA are widely available [17,28,46]. SVM and FDA are very competitive. Some recent studies exploited the relationship between FDA and SVM [27,28], especially least squares SVM (LS-SVM) [44]. Mika suggested an “improved” equivalent version to KFDA in [27], which is exactly equivalent to LS-SVM. In [27], the term containing the within-class scatter in KFDA is changed to a square loss under the Gaussian distribution assumption. In addition, the balanced relative margin machine allows for a smooth transition from SVM to FDA [23].

Intuitively, SVM tries to maximize the separating margin under a small empirical risk, whilst FDA seeks to minimize the within-class scatter under a fixed between-class scatter, which can also be viewed as the average between-class distance. Both of them try to find a good balance. However, we also notice that SVM pays more attention to the geometric classification margin, while FDA focuses on the statistical properties of data, i.e., the between-class and within-class variances. From this point of view, we expect to combine SVM and FDA to achieve a better generalization ability. One direct way to combine them is to utilize FDA as a feature extractor and then SVM as a classifier, i.e., FDA + SVM. This combination is considered in [17], which reports some promising experimental results. FDA + kernel SVM and KFDA + linear SVM both outperform SVM with raw features on face data [17]. In this direct combination, the features extracted by FDA might not be appropriate to SVM, and the number of the features needs to be decided a priori, because FDA and SVM are separately implemented. A more promising combination method was proposed in [46], called SVM/LDA. A new term for minimizing the within-class scatter is added to the optimization problem of the linear SVM. The experimental results show that SVM/LDA is also superior to SVM [46]. However, their approach uses only the linear kernel instead of other nonlinear kernels, e.g., the radial basis function (RBF) kernel.

To remedy this, Fisher-regularized SVM (FisherSVM) is presented by generalizing SVM/LDA to the reproducing kernel Hilbert space (RKHS) from the view of regularization. We put the emphasis on the Fisher criterion as a regularizer and discuss the connections of FisherSVM to other regularization methods. Similar to SVM, FisherSVM can also be cast into a quadratic convex optimization problem. Thus, the solution of FisherSVM is globally optimal. To summarize, our work has two main contributions:

1. We define a Fisher regularization in RKHS, which is represented by the within-class scatter. By comparison with the Laplacian regularization, we find that the Fisher regularization can also be viewed as a graph-based regularization.
2. We propose FisherSVM for supervised learning by introducing the Fisher regularization into SVM, which is a nonlinear extension of the SVM/LDA method.

The rest of this paper is organized as follows. In Section 2, we briefly review the related work on SVM/LDA. Section 3 describes the Fisher regularization, presents FisherSVM and discusses the connections of FisherSVM to other related methods. We apply our method to classification problems on synthetic and real-world data in Section 4 and conclude this paper in Section 5.

2. SVM/LDA

SVM and FDA (or LDA) are widely used in machine learning and pattern recognition. In the following, we briefly review SVM and SVM/LDA.