Information Sciences 370–371 (2016) 50–62

Contents lists available at ScienceDirect

Information Sciences

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

Efficient learning of supervised kernels with a graph-based loss function

Binbin Pan[a,b], Wen-Sheng Chen[a,b,*], Bo Chen[a,b], Chen Xu[c]

[a] College of Mathematics and Statistics, Shenzhen University, Shenzhen 518060, China
[b] Shenzhen Key Laboratory of Media Security, Shenzhen University, Shenzhen 518060, China
[c] Institute of Intelligent Computing Science, Shenzhen University, Shenzhen 518060, China

ARTICLE INFO

Article history:
Received 3 September 2015
Revised 11 July 2016
Accepted 25 July 2016
Available online 27 July 2016

Keywords:
Kernel learning
Kernel methods
Side information
Loss function
Supervised learning

ABSTRACT

This paper presents our study of the problems associated with learning supervised kernels from a large amount of side information. We propose a new loss function derived from the Laplacian matrix of a special complete graph that is generated from the side information. We analyze the relationship between the proposed loss function and the kernel alignment. Our theoretical analysis shows that the proposed loss function has a close relationship with kernel alignment, that is, they both make use of side information that is fused in a matrix, in addition to a similar regularization strategy. Moreover, the proposed loss function has a linear form, and thus it is more efficient in learning side information than kernel alignment that has to be performed nonlinearly. The proposed loss function is used to generate new kernels as "low-cost" alternatives of kernels learned by certain state-of-the-art methods. The empirical results demonstrate the superiority of the proposed method over state-of-the-art methods in terms of classification accuracy and computational cost.

1. Introduction

In recent years, kernel methods have been widely applied to pattern recognition [37], image analysis [28], computer vision [4], and so on. Kernel methods are a class of algorithms that measure the similarities among data points via a kernel function (or simply a *kernel*), which is a similarity function over pairs of data points. The most popular example of kernel methods is the Support Vector Machine (SVM) [7,29,34]. Kernels play an important role in the performance of kernel methods. In contrast to manual selection of a kernel, an automatic method, given a specific data set and a specific application domain, is more valuable in practice, since it requires less prior knowledge from users. Lately, a few studies [5,8,11,12,14,18–23,27,32,35,36,38–41] focused on automatic kernel learning based on the side information of data. The kernels that encode the side information of data are referred to as supervised kernels. A large amount of side information exists in many training data sets collected for real-life applications. In face recognition, for example, a training data set often contains hundreds of persons, each of whom has several instances captured under various imaging conditions. The number of labels could be as high as thousands. One important problem related to learning supervised kernels is to determine how to make use of the side information effectively and efficiently. In general, there are two ways to utilize the side information. The first is to

* Corresponding author.
E-mail addresses: pbb@szu.edu.cn (B. Pan), chenws@szu.edu.cn (W.-S. Chen), chenbo@szu.edu.cn (B. Chen), xuchen@szu.edu.cn (C. Xu).

http://dx.doi.org/10.1016/j.ins.2016.07.064
0020-0255/

learn the side information by maximizing the kernel alignment, i.e., the alignment of the combined kernel with the available labels [8,19,39,40]. The other is to incorporate the side information by pairwise constraints [14,18,22].

Kernel alignment [8] aims to measure the similarity between the learned kernel and the target kernel derived from the labels. Kernel alignment can be viewed as a loss function that measures the interpretation cost of a kernel with respect to a given training set. Some algorithms were presented for learning supervised kernels based on kernel alignment [19,39,40]. All these algorithms maximize kernel alignment as an objective function along with additional constraints. The optimization problems are often formulated as Quadratically Constrained Quadratic Programming (QCQP) because of the quadratic form of the kernel alignment. Interior-point algorithms can be applied to solve the QCQP problem effectively. However, they have poor scalability, i.e., the computational time these algorithms require to process a large-scale or even medium-scale data set increases dramatically. Therefore, the nonlinear form of kernel alignment often complicates the optimization problem unnecessarily although the optimization problem involves linear constraints only. Moreover, kernel alignment is limited to class labels, i.e., it is not applicable to other types of side information such as pairwise similarity/dissimilarity. In contrast to kernel alignment, pairwise constraints incorporate side-information as linear constraints, which are able to process different types of side information [14,18,22]. The learning problems are often formulated as an optimization problem where the objective function is the Bregman matrix divergence and the constraints are linear constraints induced by pairwise constraints. The Bregman matrix divergence measures the discrepancy between the learned kernel matrix and an input kernel matrix. Such problems can be solved via the Bregman projection algorithm, which randomly chooses one constraint and solves the resulting problem exactly. Convergence requires every constraint to be visited several times. When a large amount of side information is available, it is quite possible that the number of constraints is of the same magnitude, or even much larger than the number of data points. In this case, it is computationally expensive to solve these problems.

In this paper, we propose a new loss function capable of exploiting side information efficiently for supervised kernel learning. The loss function is defined based on a complete and weighted graph where the weight of each edge is contributed by side information. We named this graph, which encodes the side information, the Supervised Complete Graph (SCG). The learned kernel is required to be smooth over the SCG. This means that a pair of data points is expected to be close in the kernel space if they belong to the same class. The resulted loss function is named *scg-loss*. We highlight the contributions of this paper as follows.

- Our scg-loss has two attractive properties. It is a linear function of the kernel matrix to be learned, and can be applied to various forms of side information such as class labels and pairwise similarity/dissimilarity. The linear form of an scg-loss allows us to develop efficient algorithms. To the best of our knowledge, none of the existing loss functions have these two properties simultaneously.
- We theoretically compare scg-loss with kernel alignment. It demonstrates that both scg-loss and kernel alignment aim at maximizing the inner product between a kernel matrix and the target matrix derived from side-information under a regularization scheme that controls the trace of the kernel matrix.
- The scg-loss is used to develop two new kernels as "low-cost" alternatives of kernels output by two existing kernel learning algorithms [14,40]. One resulting optimization problem is a Linear Programming (LP) problem that can be solved efficiently. The other problem leads to a closed-form solution and the core operation involved is matrix inversion.
- The experimental results indicate that algorithms using scg-loss are significantly more efficient than competing methods, while achieving comparable accuracy. More importantly, the computational complexity is nearly unchanged when more side information is utilized. This is important for dealing with a large amount of side information.

The remainder of the paper is organized as follows. Section 2 briefly reviews some related work for learning supervised kernels. Our loss function is presented in Section 3. A theoretical analysis of the proposed loss function is provided in Section 4. In Section 5, new kernels are developed using the proposed loss function. Experimental results are reported in Section 6 and conclusions are drawn in Section 7.

2. Related work

We firstly review the literature on supervised kernel learning. Then we briefly introduce the algorithms for learning supervised kernels using kernel alignment and pairwise constraints.

2.1. Supervised kernel learning

We are given a set of data points $X = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}$ and the associated side information. The side information is generally provided in two forms: class label and pairwise similarity/dissimilarity. For class label, each labeled data $\mathbf{x}_i$ is assigned to a label y_i. For pairwise similarity/dissimilarity, we collect two sets $\mathcal{S}$ and $\mathcal{D}$, where $\mathcal{S}$ contains the similar pairwise data and $\mathcal{D}$ includes the dissimilar pairwise data. Note that one can build pairwise similarity/dissimilarity from the class label, i.e., two data points are similar if they have the same label and dissimilar otherwise. Thus, pairwise similarity/dissimilarity is more general. The aim of supervised kernel learning is to learn a kernel matrix or a kernel function using the available data and side information.

Algorithms for learning supervised kernels can be roughly classified into two categories. The first category incorporates the side information by optimizing a loss function. The loss functions used in previous work are listed in Table 1. For each