Neurocomputing 165 (2015) 46–53

Contents lists available at ScienceDirect

Neurocomputing

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

A pre-selecting base kernel method in multiple kernel learning

Peng Wu [a,b], Fuqing Duan [a,*], Ping Guo [a,*]

[a] Image Processing and Pattern Recognition Laboratory, Beijing Normal University, Beijing 100875, China
[b] Shandong Provincial Key Laboratory of Network based Intelligent Computing, University of Jinan, Jinan 250022, China

ARTICLE INFO

Article history:
Received 28 February 2014
Received in revised form
22 May 2014
Accepted 8 June 2014
Available online 16 April 2015

Keywords:
Multiple kernel learning
Kernel selection
Minimal redundancy maximal relevance
Kernel target alignment

ABSTRACT

The pre-defined base kernel greatly affects the performance of multiple kernel learning (MKL), but selecting the pre-defined base kernel still has no theoretical guidance. In practice, it is very difficult to select a set of appropriate base kernels without prior knowledge. In this paper, we propose a general strategy to pre-select a reasonable set of base kernels before the optimization process of MKL solvers. This strategy is based on the combination of minimal redundancy maximal relevance criteria and kernel target alignment (MRMRKA). First, we determine some candidate kernels while maintaining diversity of information; second, a set of base kernels with high discriminative ability and large diversity are selected using the MRMRKA method. These pre-selected base kernels will be used in the optimization process of the existing MKL solvers to generate better results. The experiments conducted on UCI and 15-scene datasets show that the performance of MKL is improved with the proposed pre-selected base kernel strategy.

1. Introduction

Multiple kernel learning (MKL) is a hot research topic in machine learning. It has been used in various studies and applications with great success, such as bioinformatics [1], computer vision [2], and natural language processing [3]. Compared to a single kernel, such as support vector machines (SVMs), MKL attempts to achieve better results by combining several base kernels instead of using only one specific kernel [4]. The base kernel combination coefficients and the parameter learning of the kernel play an important role in MKL. Earlier, Lanckriet et al. [5] addressed the MKL problem by formulating it as semi-definite programming. Bach et al. [6] reformulated it into a quadratically constrained quadratic programming problem. However, these two methods have a high computational cost. Sonnenburg et al. [7] treated it as a second-order cone programming problem that can be efficiently solved using interior point methods. Rakotomanonjy et al. [8] addressed it through a weighted 2-norm regularization formulation with an additional constraint on the weights, which encourages sparse kernel combination. These efforts speed up the optimization process, which makes MKL a potential solution to real-world problems.

The performance of multiple kernel learning depends on the pre-defined base kernels, but the problem of how to select the pre-defined base kernel has been traditionally left to the user. In practice, it is very difficult to select a set of appropriate base kernels without prior knowledge. One of the possible strategies is to choose as many candidate kernels as possible to alleviate the negative effects, e.g., a family of polynomial kernels of arbitrary degree or a family of Gaussian kernels with different variances restricted to a specific range, and use them directly as base kernels. The base kernels produced using this strategy may share a lot of redundant information, which will increase the computational cost in the optimization process of MKL solvers. For example, given a training dataset $D=\{\mathbf{x}_i, y_i\}_{i=1}^{N}$, where $\mathbf{x}_i$, y_i and N denote the feature vector of a sample, the corresponding class label and the sample size, respectively. A family of Gaussian kernels with different variances restricted in $[0.01, 100]$ are computed using

$$k(\mathbf{x}_i, \mathbf{x}_j) = \exp\left(-\frac{\|\mathbf{x}_i - \mathbf{x}_j\|^2}{\delta^2}\right). \quad (1)$$

Apparently, the differences among the adjacent kernels are small for a small step, e.g., 0.01, which shows that the sample feature vector contains redundant information. Most of the existing MKL solvers tend to pick out the important base kernels and obtain a sparse combination of them. Rather than using all of the base kernels directly, we can select some potential kernels before the optimization stage of MKL solvers, which works similarly to feature selection. A set of reasonable base kernels means that each kernel simultaneously has highly discriminative information and retains diversity. The strategies used for feature selection can be exploited to identify the important kernels from many candidate kernels.

* Corresponding authors.
E-mail addresses: pengwiseujn@gmail.com (P. Wu), fqduan@bnu.edu.cn (F. Duan), pguo@ieee.org (P. Guo).

http://dx.doi.org/10.1016/j.neucom.2014.06.094
0925-2312/

In this work, we focus on the problem of developing a general strategy to pre-select a reasonable set of base kernels according to the given task before the optimization process of MKL solvers. To achieve this goal, a method based on the combination of minimal redundancy maximal relevance criteria and kernel target alignment (MRMRKA) is proposed. The main contributions of our work are as follows: (1) we provide a feasible strategy for producing many potential kernels associated with a given task to users. (2) We show that useful kernels can be identified using the MRMRKA method. (3) We prove that the performance of two popular MKL solvers, i.e., simple multiple kernel learning (SimpleMKL) [8] and localized multiple kernel learning (LMKL) [9], are enhanced after using the pre-selected base kernels.

The rest of paper is organized as follows. The related works are given in Section 2. In Section 3, we introduce MKL, minimal redundancy maximal relevance criterion and kernel target alignment; our method is described in detail in Section 4. Experimental results on several benchmark datasets used for classification are presented and discussed in Section 5, and the conclusion and further work are presented in the last section.

2. The related work

After realizing the importance of choosing the most appropriate base kernels for learning, researchers proposed several learning kernel algorithms [10–14]. Rather than requesting users select a specific kernel, the learning kernel algorithms only require users to specify a family of kernels; this family of kernels can be used by a learning algorithm to form a combined kernel and derive an accurate predictor. Argyriou et al. [12] presented a framework that allows users to model richer families of kernels parameterized by a compact set and teach a suboptimal kernel using a well-designed greedy algorithm. Later, Argyriou et al. [13] found that discretizing a continuous parameter space was not necessary. Therefore, they extended the finite base kernels to infinite base ones. Both of the aforementioned methods belong to the one-stage method that consists of simultaneously learning the combination coefficients and the parameters embedded in the SVM solvers. Furthermore, Cortes et al. [10] proposed a two-stage technique and algorithm for learning kernels, which consists of learning a convex combination of base kernels in the first stage and using the learned kernel with a standard SVM solver in the second stage. Afkanpour et al. [14] found that the objective function used in Reference [13] was not jointly convex, which may lead to getting stuck in local optima. To overcome this problem, they proposed a forward stage-wise additive modeling procedure based on local search, which belongs to the group of two-stage kernel learning methods, the same as the one used in [10]. More recently, Sun et al. [15] proposed a method based on an ensemble learning strategy and MKL with an Lp-norm ($p \geq 2$) constraint to select a set of sub-kernels before MKL optimization and ultimately obtain a sparse combination of the pre-selected base kernels. The main advantage of learning kernel algorithms is to facilitate the use of MKL by specifying the parameters of a certain kernel family within an ideal range. The objective functions used in the optimization process require a linear combination of the learned kernels, which is not suitable for some non-linear cases and may lead to the over-fitting problem. Afkanpour et al. [14] observed some problems with their two-stage method; they found some kernels that produce better performance in the first-stage, but the overall performance of the method was worse than kernel combinations with worse first-stage performance. The method in Reference [15], a two-stage method, reduces the over-fitting risk by using a pre-selecting procedure in the first stage, but the linear combination of kernels in the second stage narrows the scope of application. Therefore, there is a need to set the regularization parameters carefully.

Inspired by the methods in References [10,14,15], we propose a two-stage method to tackle the problem of MKL. In the first stage, by utilizing MRMRKA, a set of base kernels that contains the most relevant information to the target and simultaneously retains large diversity is pre-selected from a set of candidate kernels. The difference between the proposed method and the ones in References [10,14] is that a set of kernels rather than a combination of kernels is employed. Using this strategy, we can avoid the over-fitting problem by dispersing the useful information around multiple kernels instead of keeping all of the information in one kernel. The set of base kernels have the flexibility to be chosen by users to combine them in the second stage, which is distinct from the linear combination form used in the literature.

3. The methodology

In this section, we will give a brief introduction of MKL, minimal redundancy maximal relevance criteria and kernel target alignment.

3.1. Multiple kernel learning

MKL can be derived from the canonical kernel method, i.e., SVM. Compared to SVM, MKL has a higher performance due to the combination of linear and nonlinear kernels instead of using only one specific kernel. Given a group of training samples $\{(\mathbf{x}_i, y_i)\}_{i=1}^{N}$, where $\mathbf{x}_i$ and y_i denote a sample with D-dimensional features; the corresponding class label has the value -1 or 1, respectively, and N is the sample size. In SVM, the solution to the problem is to find the linear discriminant with the maximum margin in Hilbert feature space, which can be formulated as follows:

$$f(\mathbf{x}) = \sum_{i=1}^{N} \alpha_i y_i k(\mathbf{x}_i, \mathbf{x}) + b, \tag{2}$$

where α_i and b are some coefficients to be learned from the training samples, while $k(\cdot,\cdot)$ is referred to as a kernel function, which specifies the inner product between all pairs of samples in a training dataset during the mapping process. The three common

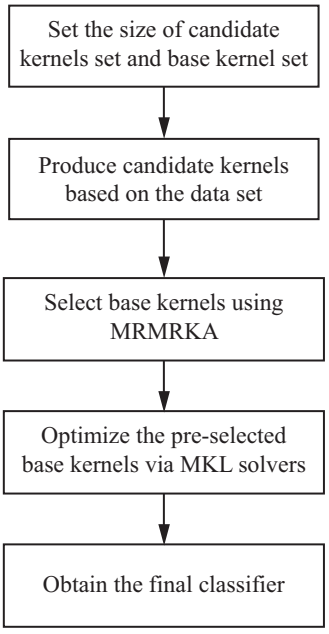

Fig. 1. The general scheme of the proposed method.