Neurocomputing 257 (2017) 16–23

Contents lists available at ScienceDirect

Neurocomputing

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

A multi-scale kernel learning method and its application in image classification

Jian Bao [a], Yangyang Chen [a,*], Li Yu [b], Chunwei Chen [a]

[a] Institute of Intelligent and Software Technology, Hangzhou Dianzi University, Hangzhou, China
[b] Information and communication engineering institute, Wuyi University, Jiangmen, China

ARTICLE INFO

Article history:
Received 12 July 2016
Revised 18 November 2016
Accepted 19 November 2016
Available online 6 February 2017

Keywords:
Multi-scale kernel
Multiple kernel learning
Kernel polarization
Image classification

ABSTRACT

The success of support vector machine depends on the kernel function, which directly affects the performance of SVM. Therefore, to improve the generalization of SVM, we will study the selection of kernel function. The multi-scale kernel method is one particular type of multiple kernel method which combines multi-scale kernels through a multi-kernel learning framework. It has the capability of generalizing not only the scattered region of a training set very well but also generalizing the dense region of data sets very well. Inspired by the advantages of the multi-scale kernel learning method, we applied kernel centered polarization to construct an optimization problem which was used to learn the multi scale kernel function and select the optimal parameters. A thorough analysis and proofs are provided. Experimental results show that the proposed kernel learning method and algorithm are reasonable and effective and have very good generalization performance.

1. Introduction

The kernel method [1,2,6] is a learning method based on kernel functions; it is widely used in various fields of machine learning. Support vector machines (SVMs) are the most successful application of kernel methods. The kernel function can easily extend a linear SVM to nonlinear, because the kernel function of complex inner product computation of high-dimensional space is converted into low dimensional input space kernel function computation, eliminating the need to design feature space [3], and cleverly solving calculations in high dimensional space of problems such as "dimension disaster". Cortes and Vapnik [4] proposed the SVM method, and because of its inherent advantages, the SVM has become a hot spot in the machine learning field since that time. The kernel method maps the input space to the feature space. Most of the time, it leads to good generalization effects. But if the selected kernel function is improper, the generalization performance will not be as good. So, the success of kernel methods depends on the selection of kernels.

In general, the methods of cross validation or "leaving one out" are used to choose kernel functions. The algorithms of the two methods are simple, but their complexity is high, and they require a large amount of calculation. For the purpose of overcoming the disadvantages of these approaches, many methods have been proposed to minimize the upper boundary of errors. The RM (radius margin) [5] boundary is one of the most common error boundaries. For the sake of further improving calculation efficiency, many kernel measurement methods select a proper kernel function by measuring the distribution of samples in feature space. For example, Cristianini [7] proposed the kernel target alignment (KTA) for the first time. This method can be used to measure the quality of a kernel matrix. Additionally, it is easy to implement the algorithm with low complexity, and is widely used in the selection of kernel function. Subsequently, Baram [8] proposed kernel polarization which can be regarded as non-normalized kernel alignment; however, the method described above is single kernel learning method. In addition, each kernel function has a different characteristic, so in different scenarios, there is a large difference in performance of the kernel function. To ameliorate the above problems, multi-kernel learning methods [10–16] appeared. Although multi-kernel methods have been successfully applied, they are only generating a new kernel function according to the Mercer condition and using the linear combination of simple single-kernel functions. There is no perfect theory for the selection of kernel functions. These methods cannot solve the problem of the uneven distribution of samples, limiting the expression ability of the decision function. In this case, the multi-scale multiple kernel learning has arisen. For example Kingsbury [17] used several multi scale kernels to classify step by step. This method has the capability to seek out a suitable kernel scale for the input space for each local area. This kind of method is flexible and practical.

* Corresponding author
E-mail address: cyy3900@foxmail.com (Y. Chen).

http://dx.doi.org/10.1016/j.neucom.2016.11.069

For multi-scale kernel learning, it is critical to determine the multi-scale kernel coefficients. There are many ways to determine the coefficients of the kernel function. Some use the idea of averaging the effect of the kernel [17,31–33], so that different kernel functions have the same effect on the decision function. In addition, some use intelligent optimization methods [31] to obtain the objective function of the parameter values. However, the large numbers of iterative steps in this kind of optimization method greatly increase the learning time for SVMs. These methods have different kernel synthesis coefficients, but they are still empirical methods. With the increase of the number of kernel functions, the dimension of the optimization problem will increase greatly. However, kernel polarization makes use of the information from the complete training set and can be computed efficiently. Furthermore, it is independent of the actual learning machine used. We applied kernel polarization to construct an optimal problem which was used to learn the multi scale kernel function and select the optimal parameters. A thorough analysis and proofs are provided. In summary, the contributions of this paper are:

(1) A multi-scale kernel learning method is proposed.
(2) It is proven that this method can determine the optimal multi-scale kernel function with low algorithm complexity.
(3) Comprehensive experiments were conducted to empirically analyze our method and the algorithm on six image databases. The experimental results demonstrate that our algorithm outperforms other methods including SVM, TSVM [34], LapSVM [35], and k-nearest neighbor.

The following sections will be organized as follows: Section 2 introduce kernel evaluation measures and multi-scale kernel learning methods. We give a detailed description and analysis of multi-scale multiple kernel learning in Section 3.The detailed description of the proposed method and algorithm are presented in Section 4. Section 5 presents the experimental results. The paper concludes in Section 6.

2. Related works

2.1. Kernel evaluation measures

The model selection problem involves selecting the kernel and optimization. The kernel evaluation measure is a model selection method. It [18–22] is a good measure of model selection, which utilizes the distribution of the samples, and is an efficient method. In addition, it is independent of any particular learning method. Cristianini [7] first proposed kernel target alignment (KTA). It has been widely used in kernel function selection. Baram put forward kernel polarization [8], which can be seen as non-normalized KTA;

$$\mathrm{P}(K,Y) = \langle K,Y\rangle_F = \sum_{i=1}^{n}\sum_{j=1}^{n} y_i y_j k\left(x_i, x_j\right), \tag{1}$$

K is the kernel matrix, $K = k(x_i, x_j)$, $y = (y_1, \ldots, y_n)^T$, $\langle K,\ yy^T\rangle_F$ is the Frobenius inner product of K and yy^T, $Y{=}yy^T$ is the target matrix; the kernel polarization criterion represents the similarity of the kernel matrix K and the target matrix. The greater the kernel polarization criterion value is achieved, the better the kernel function will be. So, when selecting a kernel function, it will be better to choose a kernel that allows reaching the highest kernel polarization criterion value.

2.2. Multiple kernel learning

Multiple kernels learning [12–16] is a flexible learning based on the kernel function. This kind of method is better than single kernel learning. The simplest and the most commonly used multiple kernel learning method is to linearly combine the basic kernel functions together, which can be described as follows:

$$\begin{cases} K = \sum\limits_{i=1}^{m} \alpha_i K_i \\ \sum\limits_{i=1}^{m} \alpha_i = 1 \\ \alpha_i \geq 0, \end{cases}$$

K_i represents the basic kernel, m stands for the sum of the basic kernels, α_i represents the weighted coefficient of K_i. Multiple kernel learning can be converted into selecting the basic kernel function and selecting the appropriate weight coefficient. The feature space of the samples is a combination of several feature spaces. Due to the use of a combination of various basic kernel feature mapping capabilities, it solves the problem of selecting the kernel function and related variables very well. Multiple kernel learning greatly improves the recognition rate and generalization ability. The most important issue is how to learn to obtain the weights. To solve this problem, more effective multiple kernel learning theories and methods have been proposed in recent years. In the early stage, the multiple kernel was learned by boosting methods [39], semi-definite programming [12], quadratically constrained quadratic program [13], semi-infinite linear program [14,25], and by Hyper kernels [11]. Subsequently, Simple MKL [28,29] was proposed. By combining the multiple kernel learning and SVM method, multiple kernel learning has been applied in many fields.

2.3. Multi-scale kernel learning

The multi-scale kernel learning method is one kind of specialized kernel method. This method fuses several different scale kernels together, and is very flexible and effective. This method is currently performing well in application. For example, Kingsbury [17] used two different scale kernels to perform step by step classification. Zheng [30] and Yang [31] proposed multi-scale support vector regression which was used to estimate functions and forecast the time series. In addition, multi-scale kernel and SVMs can be combined together and applied to image compression [32], hot spot detection and modeling [33] and measuring time-series similarity [38].Our paper uses polarization as the objective function to construct an optimal problem which can be used to learn the proposed multi-scale kernel. Our algorithm is simple and effective.

3. Multi-scale multiple kernel method

The multi-scale kernel learning method is one kind of specialized kernel method which is flexible and effective. This method fuses kernels together. A series of multi-scale kernel functions should be found as the base kernel in the first step. Then the multi-scale multiple kernel function should be constructed on this foundation. A Gaussian kernel can be described as follows:

$$k(x,z) = \exp\left(-\frac{\|x-z\|^2}{2\sigma^2}\right). \tag{3}$$

A Gaussian kernel is one kind of multi-scale kernel function which is capable of universal expression. We will use it in our paper and each function will be assigned a different bandwidth.

$$\exp\left(-\frac{\|x-z\|^2}{2{\sigma_1}^2}\right), \ldots, \exp\left(-\frac{\|x-z\|^2}{2{\sigma_m}^2}\right), \tag{4}$$

$\sigma_1 < \cdots < \sigma_\mathrm{m}$. For the Gaussian kernel, the higher bandwidth we use, the flatter the function will be. Based on this property, we know that the functions with low bandwidths will be a better