

PAPER • OPEN ACCESS

Inner Product Optimization for Effective Multiple Kernel Learning

To cite this article: Guo Niu and Zhikui Duan 2019 *IOP Conf. Ser.: Earth Environ. Sci.* **234** 012063

View the [article online](#) for updates and enhancements.

Inner Product Optimization for Effective Multiple Kernel Learning

Guo Niu and Zhikui Duan

the School of Electronics and Information Engineering, Foshan University, Foshan, China
niuguo@mail2sysu.edu.cn

Abstract. This paper proposes a novel and effective method for multiple kernel learning (MKL) that differs from standard MKL approaches. The new algorithm, named effective multiple kernel learning (EMKL), proposes a learn function space generated by multiple kernels with a group of parameters, as well as constructs a new inner-product with an optimized symmetric positive-definite matrix for the solution space of learning problem. We also describe the proposed EMKL algorithm within a general-purpose image classification framework. The EMKL algorithm has a closed-form solution, thus enabling optimization without the need for the additional iterative operations that are common in standard MKL algorithms. Hence, our proposed EMKL has a lower computational cost than that of conventional MKL techniques. The results of experiments on two benchmark datasets show the effectiveness of the proposed algorithm.

1. Introduction

Kernel method plays a key role in machine learning and has been successfully used in data classification tasks[1][2]. The proposed kernel methods can be roughly divided into two categories. The first term maps the image data from the original feature space to a high-dimensional data space, and performs kernel trick to derive the relationship between data in the high-dimensional feature space. This group includes KPCA [3] and kernel Fisher discriminant analysis [4]. The second type of kernel method [6]-[11] uses kernel function to obtain a label function that solves the learning problem by reproducing kernel Hilbert spaces (RKHSs) [5][13]. RKHSs is the function space that can be searched to find the optimal solution to the learning problem.

In these kernel methods, using multiple kernels rather than a single kernel is an important progression. Among the applications for multiple kernels, multiple kernel learning (MKL) [6] is a successful method that enhances the interpretability of the classifier with a combination of base kernels, further improving the performance of kernel methods. MKL algorithms have been widely researched [7]-[11], and reviewed [7][8]. MKL gives a feasible scheme for combining multiple kernels. However, MKL is computationally expensive, owing to the limitations of optimization procedures. Therefore, one common research direction for MKL focuses on effectively solving the MKL problem. For instance, Lanckriet et al. [9] proposed a semi-definite programming method for solving kernel weights, called SimpleMKL [10]. SimpleMKL is a state-of-the-art algorithm that can be used to solve the MKL problem with a useful subgradient descent method. Xu et al. [11] derived a variation of the equivalence between MKL and the so-called group lasso to obtain a closed-form solution for the kernel weights. However, MKL tasks remain challenging, with the reason that the optimal combination of kernels must be calculated and the optimal classifier must be determined during each iteration.



In this paper, a novel multiple kernel learning method via the inner product optimization is proposed, named effective multiple kernel learning (EMKL). EMKL uses a group of kernels and the corresponding parameters to construct the solution space for a learning problem. EMKL offers a closed-form solution, which simplifies the optimization procedure of the solution of learning problem, as well as searches an optimal label function for the learning problem in the new defined space spanned by multiple kernels. To analyze the solution space of the learning problem, we describe the manner in which the RKHS is generated by a single kernel. By analyzing the similarities and differences among single kernel methods, MKL and EMKL can be easily understood.

The paper is organized as follows: we introduce MKL in Section 2. The RKHS is described in Section 3. In Section 4, the proposed EMKL algorithm is described in detail. In Section 5, the experimental results on two real-world data sets are presented, demonstrating the benefits of EMKL. Finally, Section 6 gives some conclusions.

2. Multiple Kernel Learning (MKL)

Consider a set of data $X = (x_1, \dots, x_n) \in R^{D \times n}$ in a D-dimensional feature space, the corresponding labels are the set $y = (y_1, \dots, y_n) \in \{-1, +1\}^n$. Let $k(x, v) = \sum_{j=1}^m \beta_j k_j(x, v)$ be a combined kernel function, where $\{k_j(x, v), x, v \in X\}$ denote the m base kernels and $\{\beta_i\}_i^m$ denote the weights, and $K = \sum_{j=1}^m \beta_j K_j$ is the multiple kernel matrix. Generally, the learning problem of MKL is

$$f^* = \underset{f \in H}{\operatorname{argmin}} \sum_{i=1}^n V(y_i, f(x_i)) + \gamma \|f\|_H^2 \quad (1)$$

where f is the label function, $V(\bullet)$ is the loss function and H is an RKHS endowed with kernel function $k(x, v)$. MKL method learns the decision function

$$f(x) = \sum_{i=1}^n \alpha_i k(x, x_i) \quad (2)$$

Obviously, MKL algorithms must identify two sets of parameters, as well as obtain the solution of the problem (2). Equation (2) constitutes the bulk of the MKL procedure.

3. Reproducing Kernel Hilbert Spaces (RKHS)

RKHS [5] theory provides an unified framework for machine learning. The RKHS can be generated by a symmetric positive-definite kernel function. To do so, a kernel function $k(x, v)$ is first used to build a linear space with the data X , i.e. $S(X) = \operatorname{span}\{k_v(x) | k_v(x) = k(x, v), v \in X\}$. Second, defining an inner product $\langle \bullet, \bullet \rangle$ on $S(X)$, i.e., for all $f \in S(X)$, since $f(x) = \sum_i \alpha_i k(x, v_i)$, the inner product of f can be defined as:

$$\langle f, f \rangle = \sum_i \sum_j \alpha_i \alpha_j k(v_i, v_j) \quad (3)$$

It can easily be demonstrated that the definition of $\langle \bullet, \bullet \rangle$ meets the requirements of inner product. Therefore, $H = (S(X), \langle \bullet, \bullet \rangle)$ is an inner space. The crucial component to this kernel method is applied to RKHS, which describe the relationship between data or functions by the inner product.

4. Effective Multiple Kernel Learning via Inner Product Optimization

In this paper, we propose a effective MKL method with inner product optimization. The label

function in our algorithm is set with multiple kernels:

$$f(x) = \sum_{i=1}^n \alpha_i k_i(x, x_i) \quad (4)$$

According to Eq. 4, the number of multiple kernels $\{k_i(x, v), i=1 \cdots n\}$ is fixed. Therefore, the classification function f is determined by each data sample. But this classification function cannot be directly used to the framework of the MKL algorithm (Eq.1). To solve this problem, we learn a suitable and effective solution space for the new classification function. Inspired by the generating of RKHS, a new inner-product space can be defined through the inner product optimization.

We first built a linear space $S_T(X) = \text{span}\{k_i(x, x_i) | i=1 \cdots n, x_i \in X\}$. Then, base on the definition of the inner product, $\langle f, f \rangle = \sum_i \sum_j \alpha_i \alpha_j \langle k_i(\bullet, x_i) k_j(\bullet, x_j) \rangle$ can be obtained. Because $k_i(x_j, x_i) \neq k_j(x_i, x_j)$ (here, $k_i(x, v)$ and $k_j(x, v)$ are different kernel functions when $i \neq j$), i.e., the corresponding kernel matrix

$$K_S = \begin{bmatrix} k_1(x_1, x_1) & \cdots & k_n(x_1, x_n) \\ \vdots & \ddots & \vdots \\ k_1(x_n, x_1) & \cdots & k_n(x_n, x_n) \end{bmatrix}_{n \times n}$$

is not symmetric. To solve this problem, we further construct a learned inner product for describe the relationship between data. Before that, a new symmetric and positive-definite matrix with a regular term is defined, i.e., $K_t = (K_S + K_S^T) / 2 + \rho I$, where ρ is the regular parameter. The new define matrix can be optimized by the Bregman divergence[12]. For K_t , the Bregman divergence is $D_d(K_t, K_S) = \text{tr}(K_t K_S^{-1}) - \log \det(K_t K_S^{-1}) - \varepsilon$. With a slack variable δ for avoiding no feasible solution, the learning problem of K_t becomes

$$\min D_d(K_t, K_S) + \delta \|\xi\|_2^2$$

where $\xi_{ij} = K_t(x_i, x_i) + K_t(x_j, x_j) - K_t(x_i, x_j) - d(x_i, x_j)$.

With the symmetric positive-definite matrix K_t , the inner product on $S_T(X)$ can be obtained, i.e., for any $f \in S_T(X)$, $\langle f, f \rangle = \vec{\alpha}^T K_t \vec{\alpha}$. It is easy to prove that the defined $\langle f, f \rangle$ meets the requirements of the inner product and that, consequently, $H_t = (S_T(X), \langle \bullet, \bullet \rangle)$ is an inner product space. Because $S_T(X)$ is finite-dimensional, H_t is a Hilbert space. The EMKL estimates the new classification function by minimizing

$$f^* = \underset{f \in H_t}{\operatorname{argmin}} \sum_{i=1}^n V(y_i, f(x_i)) + \gamma \|f\|_{H_t}^2 \quad (5)$$

when the squared loss function is selected for the label function, the problem of EMKL then is reformulated as $\min_{\vec{\alpha}} (y - K_S \vec{\alpha})^T (y - K_S \vec{\alpha}) + \gamma \vec{\alpha}^T K_t \vec{\alpha}$. Obviously, because $K_S = \sum_{i=1}^n \alpha_i K_i$, EMKL only need to solve one set of parameters $\vec{\alpha}$ and has a closed-form solution.

5. Experiments

In our experiments, the general single-kernel method and MKL algorithm are used to compared our EMKL. We use the USPS handwritten digit dataset [14] and the object dataset ETH-80 for experiments.

The USPS handwritten digit set contains 10 handwritten digits (form ‘0’ to ‘9’), each class has 1100 grayscale images. The ETH-80 set has eight class of images, each class contains 10 objects, and each object has 41 images from different perspectives.

The two parameters δ and γ are selected from set $\{100, 10, 5, 1, 0.1, 0.01, 0.001\}$. The Gaussian kernel $k(u, v) = \exp \frac{\|u-v\|^2}{t^2}$ is selected as the base kernel. In the single-kernel algorithm, $t = \frac{1}{n} \sum_{i,j}^{n} \text{dis}(x_i, x_j)$, where $\text{dis}(x_i, x_j)$ denotes the distance between x_i and x_j , and n is the number of data samples. For each class, the number of data sample is set to $u = 10, 20, 30, 50$.

The experimental results on two real-world data set are shown in Table 1. As can be seen, in the two-class experiments, MKL has the lowest error rate when $u = 50$, and our EMKL has a little improvement over the single-kernel method. And for the multi-class experiments, EMKL has the best results. The reason is that for MKL the optimized kernel learned from the multiple kernels can successfully describe the data distribution, for the single-kernel method it only use an empirical kernel. Therefore, one suitable kernel can improve classification result effectively. For multi-class data samples, the data distribution becomes diverse. Therefore, our EMKL is achieves superior performance. For the ETH-80 set, the data distribution of each class is entirely different. The experimental results on ETH-80 are shown in Table 2. EMKL offered considerably improved performance over the single-kernel method, as well as generally outperformed MKL. We also record the computation time of each algorithm, as shown in Fig. 1. It is clear that the computational cost on MKL increases exponentially with the number of data samples. By contrast, the run-time for EMKL has a little variety when the number of data is given.

Table 1. Classification error rates (%) of experiments on USPS Set

	Single-kernel	MKL	EMKL
Two-class			
u=10	3.6052	3.3755	4.5781
u=20	2.4272	2.1892	2.5791
u=30	1.9013	1.6576	1.9203
u=50	1.3785	1.2857	1.3519
Multi-class			
u=10	18.1441	16.7352	15.9085
u=20	10.0634	8.9668	8.7451

Table 2. Classification error rates (%) of experiments on ETH-80 Set

	Single-kernel	MKL	EMKL
Two-class			
u=10	6.0714	5.9056	6.1578
u=20	4.5232	5.3374	5.4716
u=30	3.8273	3.5682	3.5115
u=50	3.4286	2.9813	2.5443
Multi-class			
u=10	23.8465	21.2316	17.8945
u=20	18.1856	15.9473	14.1571

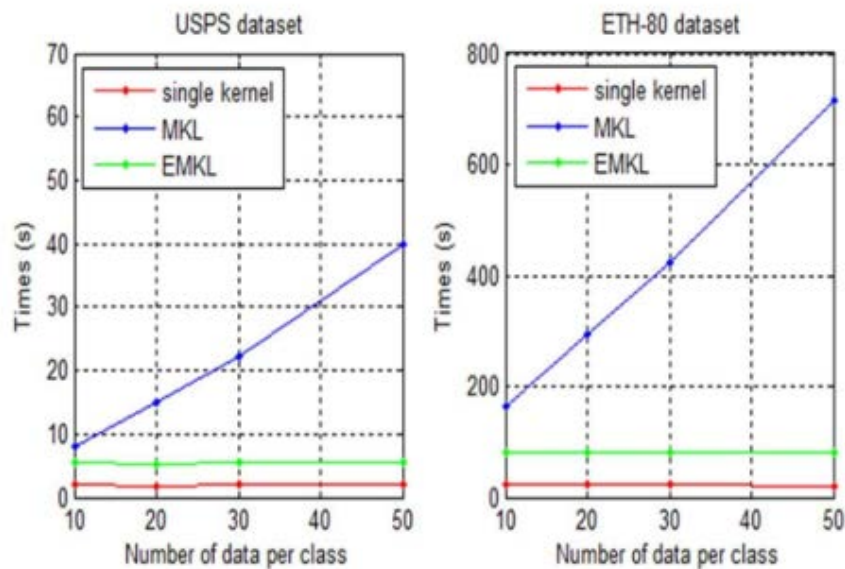


Figure 1. The computation time for three algorithm (single-kernel, MKL and EMKL)

6. Conclusion

Kernel method play a key role in machine learning. In this paper, we proposed a novel and effective MKL method for classification. Compared to a single-kernel approach, our algorithm introduces multiple better kernels to learning problem with diverse data. By comparison with general MKL approach, our proposed EMKL has less computation cost, because the parameters of multiple kernels are not introduced to the learning problem. The results of experiments on two benchmark data sets show the effectiveness of the proposed algorithm.

7. Acknowledgments.

This work was supported by Dr. start-up Foundation (gg07033) of Foshan University.

8. References

- [1] Shawe-Taylor, J. and Cristianini, N., Kernel Methods for Pattern Analysis, Cambridge University Press, (2004)
- [2] Scholkopf, B. and Smola, A. J., Learning with Kernels, MIT press, (2001)
- [3] Scholkopf, B. and Somla, A. J and Müller, K. R., Nonlinear component analysis as a kernel eigenvalue problem, Neural Comput., pp. 1299-1319, (1998)
- [4] Mika, S. and Ratsch, G. and Weston, J. and Scholkopf, J. B. and Müller, K. R., Fisher discriminant analysis with kernels, IEEE Signal Process. Soc. Workshop Neural Netw. Signal Process., 1999, pp. 41-48, (1999)
- [5] Aronszajn, N., The theory of reproducing kernels and their applications, Cambridge Phil. Soc. Proc., pp. 133-153, (1943)
- [6] Lanckriet, G. and De Bie, T. and Cristianini, N. and Jordan, M. and Noble, W., A statistical framework for genomic data fusion, Bioinformatics, pp. 2626-2635, (2004)
- [7] Tobar, F. A. and Kung, S.-Y. and Mandic, D. P., Multikernel least mean square algorithm, IEEE Trans. Neural Netw. Learn. Syst., pp. 265-277, (2014)
- [8] Bucak, S. and Jin, R. and Jain, A., Multiple kernel learning for visual object: a review, IEEE Trans. Pattern Anal. Mach. Intell. pp. 1354 - 1369, (2013)
- [9] Gnen, M. and Alpaydin, E., Multiple kernel learning algorithms, J. Mach. Learn. Res., 2011, pp. 2211-2268, (2011)

- [10] Rakotomamonjy, A. and Bach, F. and Canu, S. and Grandvalet, Y., SimpleMKL, J. Mach. Learn. Res. , pp. 2491-2521, (2008)
- [11] Xu, Z. and Jin, R. and Yang, H. and King, I. and Lyu, M. R., Simple and Efficient Multiple Kernel Learning by Group Lasso, ICML, (2010)
- [12] K. Q. Weinberger and L. K. Saul., Fast solvers and efficient implementations for distance metric learning, ICML, (2008).
- [13] Niu, G, Ma, Z. M., and Liu, S. Y., A Multikernel-Like Learning Algorithm Based on Data Probability Distribution, Math. Problems in Engineering, pp.1-18,(2017)
- [14] <http://www.cs.toronto.edu/roweis/data.html>
- [15] <http://people.csail.mit.edu/jjl/libpmk/samples/eth.html>