Learning Kernels via Margin-and-Radius Ratios

Kun Gai **Guangyun Chen** **Changshui Zhang**
State Key Laboratory on Intelligent Technology and Systems
Tsinghua National Laboratory for Information Science and Technology (TNList)
Department of Automation, Tsinghua University, Beijing 100084, China
{gaik02, cgy08}@mails.thu.edu.cn, zcs@mail.thu.edu.cn

1 Introduction

Despite the great success of SVM, it is usually difficult for users to select suitable kernels for SVM classifiers. Kernel learning has been developed to jointly learn both a kernel and an SVM classifier [1]. Most existing kernel learning approaches, e.g., [2, 3, 4], employ the margin based formulation, equivalent to:

$$\min_{k,w,b,\xi_i} \quad \tfrac{1}{2}\|w\|^2 + C\textstyle\sum_i \xi_i, \quad \text{s.t.} \quad y_i\langle\phi(x_i;k), w\rangle + b + \xi_i \geq 1,\ \xi_i \geq 0, \tag{1}$$

where k is the learned kernel which implicitly defines a transformation $\phi(\cdot;k)$ to a feature space by $k(x_c, x_d) = \langle\phi(x_c;k), \phi(x_d;k)\rangle$, $(w,\ b)$ is an SVM classifier, and x_i, y_i and ξ_i are input instances, labels and hinge losses. To make the problem trackable, the learned kernel is usually restricted to a parametric form $k^{(\theta)}(\cdot,\cdot)$, where $\theta = [\theta_i]_i$ is the kernel parameter. The most common used form is a linear combination of multiple basis kernels, as

$$k^{(\theta)}(\cdot,\cdot) = \textstyle\sum_{j=1}^{m} \theta_j k_j(\cdot,\cdot), \ \ \theta_j \geq 0. \tag{2}$$

Let γ denote the margin of the SVM classifier with k. It is well known that $\gamma^{-2} = \|w\|^2$. Thus the term $\|w\|^2$ makes the margin based formulation (1) prefer the kernel that results in an SVM classifier with a larger margin. However, the margin itself can not well describe the goodness of a kernel. Any kernel, even one with a bad performance, can have arbitrarily large margin by enlarging the kernel's scaling, and may be selected to be the final solution [5]. Therefore the margin based kernel learning methods suffer from scaling problems. In linear combination cases, a remedy is to enforce a norm constraint on kernel parameters. Unfortunately, it is difficult to select suitable types of norm constraints, and with norm constraints the scaling problem also causes another initialization problem: different initial scalings of basis kernels lead to different final learned kernels (Examples can be found in [5]). In nonlinear combination cases, a norm constraint on kernel parameters even can not generally guarantees that the learned kernel's scaling keeps finite in the optimization process.

This paper reports our recently presented scaling-invariant principle and algorithm for kernel learning [5]. Motivated by the generalization bounds of kernel learning, we use the ratio between the margin of the SVM classifier with a kernel and the radius of the minimum enclosing ball (MEB) of data in the feature space endowed with the kernel as a measure of the goodness of the kernel, and propose a new kernel learning method. Our approach differs from the radius-based methods of Chapelle et al. [1] and Do et al. [6]. It has been shown that their methods are still sensitive to kernel scalings [5], causing the same problems as margin based methods. We prove that our formulation is invariant to scalings of learned kernels, and in linear combination cases it is also invariant both to initial scalings of basis kernels and to types of norm constraints on kernel parameters. Our proposed kernel learning problem can be transformed to a tri-level optimization problem. By establishing the differentiability of a general family of multilevel optimal functions, we give a simple and efficient gradient-based algorithm for kernel learning. Experiments show that our approach achieves higher accuracies both than SVM with the uniform combination of basis kernels and than other state-of-art kernel learning methods.

2 Measuring how good a kernel is

We now discuss how to measure the goodness of a kernel. For SVM with a kernel learned from a kernel family $\mathcal{K}$, if we restrict that the radius of the minimum enclosing ball in the feature space endowed with the learned kernel to be no larger than R, then the theoretical results of Srebro and Ben-David [7] say: for any fixed margin $\gamma > 0$ and any fixed radius $R > 0$, with probability at least $1 - \delta$ over a training set of size n, the gap between expected and empirical errors is no larger than $\sqrt{\frac{8}{n}(2 + d_\phi \log \frac{128en^3R^2}{\gamma^2 d_\phi} + 256\frac{R^2}{\gamma^2} \log \frac{en\gamma}{8R} \log \frac{128nR^2}{\gamma^2} - \log \delta)}$. Scalar d_ϕ denotes the pseudodimension [7] of the kernel family $\mathcal{K}$. The above results clearly state that the generalization error bounds of kernel learning in both linear and nonlinear cases depend on the ratio between the margin and the radius of the minimum enclosing ball of data. Therefore, we use the margin-and-radius ratio to measure how good a kernel is for kernel learning.

Given any kernel k, the radius of the minimum enclosing ball, denoted by $R(k)$, can be obtained by the following concave maximization problem [5].

$$R^2(k) = \max_{\beta_i} \textstyle\sum_i \beta_i k(x_i, x_i) - \sum_{i,j} \beta_i k(x_i, x_j)\beta_j, \quad \text{s.t. } \sum_i \beta_i = 1.\ \beta_i \geq 0. \tag{3}$$

3 Learning kernels with margin-and-radius ratio

By using the margin-and-radius ratio, we propose a new kernel learning formulation, as

$$\min_{k,w,b,\xi_i} \tfrac{1}{2}R^2(k)\|w\|^2 + C\textstyle\sum_i \xi_i, \quad \text{s.t. } y_i(\langle\phi(x_i;k), w\rangle + b) + \xi_i \geq 1,\ \xi_i \geq 0. \tag{4}$$

This optimization problem is called radius based kernel learning problem, referred to as *RKL*. It can be reformulated to

$$\min_k\ G(k), \tag{5}$$

$$\text{where } G(k) = \min_{w,b,\xi_i} \tfrac{1}{2}R^2(k)\|w\|^2 + C\textstyle\sum_i \xi_i,\ \text{ s.t. } y_i(\langle\phi(x_i;k), w\rangle + b) + \xi_i \geq 1,\ \xi_i \geq 0. \tag{6}$$

Functional $G(k)$ defines a measure of the goodness of kernel functions. This functional is scaling invariant: for any kernel k and any scalar $a > 0$, equation $G(ak) = G(k)$ holds [5].

Now consider the linear combination case in (2), and use $g_{\text{linear}}(\theta)$ to denote $G(k^{(\theta)})$ in such case. From scaling invariance we can get the following properties [5]. First, problems of minimizing $g_{\text{linear}}(\theta)$ under different types of norm constraints on θ are equivalent to each other, and also equivalent to the problem of minimizing $g_{\text{linear}}(\theta)$ without any norm constraint on θ. Second, the problems of minimizing $g_{\text{linear}}(\theta)$ with different initial scalings of basis kernels are also equivalent to each other. Therefore, our formulation completely addresses the problems in margin based methods.

3.1 Reformulation to a tri-level optimization problem

The remaining task is to optimize the RKL problem (5). Given a parametric kernel form $k^{(\theta)}$, for any parameter θ, to obtain the value of the objective function $g(\theta) \doteq G(k^{(\theta)})$ in (5), we need to solve the SVM-like problem in (6), which is a convex minimization problem and can be solved by its dual problem. Indeed, the whole RKL problem is transformed to a tri-level optimization problem:

$$\min_\theta\ g(\theta), \tag{7}$$

$$\text{where } g(\theta) = \Big\{\max_{\alpha_i} \textstyle\sum_i \alpha_i - \frac{1}{2r^2(\theta)} \sum_{i,j} \alpha_i\alpha_j y_i y_j K_{i,j}(\theta),\ \text{s.t.} \sum_i \alpha_i y_i = 0,\ 0 \leq \alpha_i \leq C\Big\}, \tag{8}$$

$$\text{where } r^2(\theta) = \Big\{\max_{\beta_i} \textstyle\sum_i \beta_i K_{i,j}(\theta) - \sum_{i,j} \beta_i K_{i,j}(\theta)\beta_j,\ \text{s.t.} \sum_i \beta_i = 1,\ \beta_i \geq 0\Big\}, \tag{9}$$

where $K(\theta)$ denotes the kernel matrix $[k^{(\theta)}(x_i, x_j)]_{i,j}$, and $r(\theta)$ denotes $R(k^\theta)$. If $g(\theta)$, which is the objective function in the top-level optimization, is differentiable and we can get its derivatives, then we can use gradient-based methods to solve the RKL problem. The Danskin's theory states the differentiability of single-level optimal functions [5], and has been successfully applied in many kernel learning methods, e.g., [1, 4]. Unfortunately, here our objective function $g(\theta)$ is a bi-level optimal function, and the Danskin's theory can not be directly applied, which makes the RKL problem much more challenging. Below we develop new results about multilevel optimal functions.

4 Differentiability of the multilevel optimization problem

Let Y be a metric space, and X, U and Z be normed spaces. Suppose: (1) The function $g_1(x, u, z)$, is continuous on $X \times U \times Z$. (2) For all $x \in X$ the function $g_1(x, \cdot, \cdot)$ is continuously differentiable. (3) The function $g_2(y, x, u)$ ($g_2 : Y \times X \times U \to Z$) is continuous on $Y \times X \times U$. (4) For all $y \in Y$ the function $g_2(y, \cdot, \cdot)$ is continuously differentiable. (5) Sets $\Phi_X \subseteq X$ and $\Phi_Y \subseteq Y$ are compact. By these notations, we propose the following theorem about bi-level optimal value functions [5].

Theorem 1. *Let us define a bi-level optimal value function as*

$$v_1(u) = \inf_{x \in \Phi_X} g_1(x, u, v_2(x, u)), \tag{10}$$

where $v_2(x, u)$ is another optimal value function as

$$v_2(x, u) = \inf_{y \in \Phi_Y} g_2(y, x, u). \tag{11}$$

If for any x and u, $g_2(\cdot, x, u)$ has a unique minimizer $y^*(x, u)$ over Φ_Y, then $y^*(x, u)$ are continuous on $X \times U$, and $v_1(u)$ is directionally differentiable. Furthermore, if for any u, the $g_1(\cdot, u, v_2(\cdot, u))$ has also a unique minimizer $x^*(u)$ over Φ_X, then

1. *the minimizer $x^*(u)$ are continuous on U,*
2. *$v_1(u)$ is continuously differentiable, and its derivative is equal to*

$$\frac{dv_1(u)}{du} = \left(\frac{\partial g_1(x^*, u, v_2)}{\partial u} + \frac{\partial v_2(x^*, u)}{\partial u} \frac{\partial g_1(x^*, u, v_2)}{\partial v_2} \right) \Bigg|_{v_2 = v_2(x^*, u)}, \quad \text{where } \frac{\partial v_2(x^*, u)}{\partial u} = \frac{\partial g_2(y^*, x^*, u)}{\partial u}. \tag{12}$$

To apply this theorem to the objective function $g(\theta)$ in the RKL problem (7), we shall make sure the following two conditions are satisfied. First, both the MEB problem (9) and the SVM dual problem (8) must have unique optimal solutions. This can be guaranteed by that the kernel matrix $K(\theta)$ is strictly positive definite. Second, the kernel matrix $K(\theta)$ shall be continuously differentiable to θ. Both conditions can be met in the linear combination case when each basis kernel matrix is strictly positive definite, and can also be easily satisfied in nonlinear cases. If these two conditions are met, then $g(\theta)$ is continuously differentiable and

$$\frac{dg(\theta)}{d\theta} = -\frac{1}{2r^2(\theta)} \sum_{i,j} \alpha_i^* \alpha_j^* y_i y_j \frac{dK_{i,j}(\theta)}{d\theta} + \frac{1}{2r^4(\theta)} \sum_{i,j} \alpha_i^* \alpha_j^* y_i y_j K_{i,j}(\theta) \frac{dr^2(\theta)}{d\theta}, \tag{13}$$

where α_i^* is the optimal solution of the SVM dual problem (8), and

$$\frac{dr^2(\theta)}{d\theta} = \sum_i \beta_i^* \frac{dK_{i,i}(\theta)}{d\theta} - \sum_{i,j} \beta_i^* \frac{dK_{i,j}(\theta)}{d\theta} \beta_j^*, \tag{14}$$

where β_i^* is the optimal solution of the MEB dual problem (9). In above equations, the value of $\frac{dK_{i,j}(\theta)}{d\theta}$ is needed, and its deriving is easy. For example, for the linear combination kernel $K_{i,j}(\theta) = \sum_m \theta_m K_{i,j}^m$, we have $\frac{\partial K_{i,j}(\theta)}{\partial \theta_m} = K_{i,j}^m$. For the Gaussian kernel $K_{i,j}(\theta) = e^{-\theta \|x_i - x_j\|^2}$, we have $\frac{dK_{i,j}(\theta)}{d\theta} = -K_{i,j}(\theta) \|x_i - x_j\|^2$.

5 Algorithm

With the derivative of $g(\theta)$, we use the standard gradient projection approach with the Armijo rule for selecting step sizes to address the RKL problem [5]. To compare with the most popular kernel learning algorithm, simpleMKL [4], in experiments we employ the linear combination kernel form, as defined in (2). In addition, we also consider three types of norm constraints on kernel parameters: L_1, L_2 and no norm constraint. The L_1 and L_2 norm constraints are as $\sum_j \theta_j = 1$ and $\sum_j \theta_j^2 = 1$, respectively. The calculation of the objective function $g(\theta)$ and its gradient needs solving an MEB problem (9) and an SVM problem (8), and both of them can be efficiently solved by SMO algorithms. In experiments our approach usually achieves approximate convergence within one or two dozens of invocations of SVM and MEB solvers. More analyses about the algorithm and its output solutions can be found in [5].

6 Experiments

In this section, we illustrate the performances of our presented RKL approach, in comparison with SVM with the uniform combination of basis kernels (Unif), the margin based MKL method using formulation (1) (MKL), and the radius-based principle by Chapelle et al. [1] (KL-C), under different types of norm constraints. The evaluation is made on eleven public available data sets from UCI repository and LIBSVM Data (For details see [5]). The used basis kernels are the same as in SimpleMKL [4]: 10 Gaussian kernels with bandwidths $\gamma_G \in \{0.5, 1, 2, 5, 7, 10, 12, 15, 17, 20\}$ and 10 polynomial kernels of degree 1 to 10. All basis kernel matrices have been normalized to unit trace, as in [4, 8]. The initial kernel parameter θ is set to be $\frac{1}{20}e$, where e is a unit vector. The trade-off coefficients C in SVM, MKL, KL-C and RKL are automatically selected from $\{0.01, 0.1, 1, 10, 100\}$ by 3-fold cross-validations on training sets. For each data set, we split it to five parts, and each time we use four parts as the training set and the remaining one as the test set. The average accuracies with standard deviations and average numbers of selected basis kernels are reported in Table 1.

Table 1: The testing accuracies (Acc.) with standard deviations (in parentheses), and the average numbers of selected basis kernels (Nk). We set the numbers of our method to be bold if our method outperforms both Unif and other two kernel learning approaches under the same norm constraint.

Index	1		2		3		4		5		6		7		8	
	Unif		MKL		KL-C		Ours		MKL		KL-C		Ours		Ours	
Constraint			L_1		L_1		L_1		L_2		L_2		L_2		No	
Data set	Acc.	Nk	Acc.	Nk	Acc.	Nk	Acc.	Nk	Acc.	Nk	Acc.	Nk	Acc.	Nk	Acc.	Nk
Ionosphere	94.0(1.4)	20	92.9(1.6)	3.8	86.0(1.9)	4.0	**95.7**(0.9)	2.8	94.3(1.5)	20	84.4(1.6)	18	**95.7**(0.9)	3.0	95.7(0.9)	3.0
Splice	51.7(0.1)	20	79.5(1.9)	1.0	80.5(1.9)	2.8	**86.5**(2.4)	3.2	82.0(2.2)	20	74.0(2.6)	14	**86.5**(2.4)	2.2	86.3(2.5)	3.2
Liver	58.0(0.0)	20	59.1(1.4)	4.2	62.9(3.5)	4.0	**64.1**(4.2)	3.6	67.0(3.8)	20	64.1(3.9)	11	64.1(4.2)	8.0	64.3(4.3)	6.6
Fourclass	81.2(1.9)	20	97.7(1.2)	7.0	94.0(1.2)	2.0	**100** (0.0)	1.0	97.3(1.6)	20	94.0(1.3)	17	**100** (0.0)	1.0	100 (0.0)	1.6
Heart	83.7(6.1)	20	84.1(5.7)	7.4	83.3(5.9)	1.8	84.1(5.7)	5.2	83.7(5.8)	20	83.3(5.1)	19	**84.4**(5.9)	5.4	84.8(5.0)	5.8
Germannum	70.0(0.0)	20	70.0(0.0)	7.2	71.9(1.8)	9.8	**73.7**(1.6)	4.8	71.5(0.8)	20	71.6(2.1)	13	**73.9**(1.2)	6.0	73.9(1.8)	5.8
Musk1	61.4(2.9)	20	85.5(2.9)	1.6	73.9(2.9)	2.0	**93.3**(2.3)	4.0	87.4(3.0)	20	61.9(3.1)	19	**93.5**(2.2)	3.8	93.3(2.3)	3.8
Wdbc	94.4(1.8)	20	97.0(1.8)	1.2	97.4(2.3)	4.6	97.4(1.6)	6.2	96.8(1.6)	20	97.4(2.0)	11	**97.6**(1.9)	5.8	97.6(1.9)	5.8
Wpbc	76.5(2.9)	20	76.5(2.9)	7.2	52.2(5.9)	9.6	76.5(2.9)	17	75.9(1.8)	20	51.0(6.6)	17	**76.5**(2.9)	15	76.5(2.9)	15
Sonar	76.5(1.8)	20	82.3(5.6)	2.6	80.8(5.8)	7.4	**86.0**(2.6)	2.6	85.2(2.9)	20	80.2(5.9)	11	**86.0**(2.6)	2.6	86.0(3.3)	3.0
Coloncancer	67.2(11)	20	82.6(8.5)	13	74.5(4.4)	11	**84.2**(4.2)	7.2	76.5(9.0)	20	76.0(3.6)	15	**84.2**(4.2)	5.6	84.2(4.2)	7.6

The results in Table 1 can be summarized as follows. (a) RKL gives the best results on most sets. Under L_1 norm constraints, RKL (Index 4) outperforms all other methods (Index 1, 2, 3) on 8 out of 11 sets, and also gives results equal to the best ones of other methods on the remaining 3 sets. In particular, RKL gains 5 or more percents of accuracies on Splice, Liver and Musk1 over MKL, and gains more than 9 percents on four sets over KL-C. Under L_2 norm constraints, RKL (Index 7) outperforms other methods (Index 5, 6) on 10 out of 11 sets, with only 1 inverse result. (b) Both MKL and KL-C are sensitive to the types of norm constraints (Compare Index 2 and 5, as well as 3 and 6). For MKL and KL-C, different types of norm constraints fit different data sets. (c) RKL is invariant to the types of norm constraints. (d) An interesting thing is that, our presented RKL gives sparse solutions on most sets, whatever types of norm constraints are used. Compared to MKL and KL-C under L_2 norm constraints, RKL provides not only higher performances but also more sparsity, which benefits both interpretability and computational efficiency in prediction.

7 Conclusion

By using the margin-and-radius ratio as a measure of the goodness of a kernel, we propose a scaling invariant principle and a simple algorithm for kernel learning. The experiments validate that our approach outperforms other state-of-art kernel learning methods.

References

[1] O. Chapelle, V. Vapnik, O. Bousquet, and S. Mukherjee. Choosing multiple parameters for support vector machines. *Machine Learning*, 46(1):131–159, 2002.

[2] G.R.G. Lanckriet, N. Cristianini, P. Bartlett, L.E. Ghaoui, and M.I. Jordan. Learning the kernel matrix with semidefinite programming. *The Journal of Machine Learning Research*, 5:27–72, 2004.

[3] S. Sonnenburg, G. Rätsch, and C. Schäfer. A general and efficient multiple kernel learning algorithm. In *Adv. Neural. Inform. Process Syst. (NIPS 2005)*, 2006.

[4] A. Rakotomamonjy, F. Bach, S. Canu, and Y. Grandvalet. SimpleMKL. *Journal of Machine Learning Research*, 9:2491–2521, 2008.

[5] Kun Gai, Guangyun Chen, and Changshui Zhang. Learning kernels with radiuses of minimum enclosing balls. In *Adv. Neural. Inform. Process Syst. (NIPS 2010)*, 2010.

[6] H. Do, A. Kalousis, A. Woznica, and M. Hilario. Margin and Radius Based Multiple Kernel Learning. In *Proceedings of the European Conference on Machine Learning (ECML 2009)*, 2009.

[7] N. Srebro and S. Ben-David. Learning bounds for support vector machines with learned kernels. In *Proceedings of the International Conference on Learning Theory (COLT 2006)*, pages 169–183. Springer, 2006.

[8] M. Kloft, U. Brefeld, S. Sonnenburg, P. Laskov, K. Müller, and A. Zien. Efficient and Accurate lp-Norm Multiple Kernel Learning. In *Adv. Neural. Inform. Process Syst. (NIPS 2009)*, 2009.