

Nonlinear Combination of Multiple Kernels for Support Vector Machines

Jinbo Li, Shiliang Sun

Department of Computer Science and Technology
East China Normal University
500 Dongchuan Road, Shanghai 200241, China
Email: slsun@cs.ecnu.edu.cn

Abstract—Support vector machines (SVMs) are effective kernel methods to solve pattern recognition problems. Traditionally, they adopt a single kernel chosen beforehand, which makes them lack flexibility. The recent multiple kernel learning (MKL) overcomes this issue by optimizing over a linear combination of kernels. Despite its success, MKL neglects useful information generated from the nonlinear interaction of different kernels. In this paper, we propose SVMs based on the nonlinear combination of multiple kernels (NCMK) which surmounts the drawback of previous MKL by the potential to exploit more information. We show that our method can be formulated as a semi-definite programming (SDP) problem then solved by interior-point algorithms. Empirical studies on several data sets indicate that the presented approach is very effective.

Keywords—Hadamard product; multiple kernel learning; semi-definite programming; support vector machine

I. INTRODUCTION

In recent years, support vector machines (SVMs) have been successfully applied to many pattern recognition problems [1]. The essence of SVMs is to use a kernel function k to implicitly map the original data \mathbf{x} into a feature space by a map Φ , and then seek linear relations in this space. As data distributions vary in different feature spaces, kernels play an important role in the performance of SVMs. Generally, a single kernel may not be sufficient to solve complex problems, e.g., those involving multiple heterogeneous data sources.

Some recent research on multiple kernel learning (MKL) has alleviated the above issue to a certain extent. While MKL can in principle be solved through cross-validation, researchers have developed more efficient methods. For example, Lanckriet et al. [2] considered conic combinations of kernel matrices and transformed it into a semi-definite programming (SDP) problem. Bach et al. [3] reformulated this problem and proposed an SMO algorithm for medium-scale problems. Other efficient MKL algorithms include those proposed by Sonnenburg et al. [4] and Rakotomamonjy et al. [5]. Recently, Gönen et al. [6] introduced a localized MKL approach using a gating model to select appropriate local kernel functions. A common feature of these MKL methods is that they consider linear combinations of multiple kernels or kernel matrices.

In this paper, we concentrate on the nonlinear combination of multiple kernels (NCMK) for SVMs. In our approach, the Hadamard product of any two kernel matrices generated by original kernels is firstly used to form several new kernel matrices. Then, a linear combination of these new kernel matrices and the original ones is considered. The optimal weight for each kernel matrix is then solved after we integrate this MKL with SVMs and transform the problem into a SDP problem. The nonlinearity of our kernel combination is mainly reflected by the Hadamard product approach to build new kernel matrices. Closure properties of kernels guarantee that the final combined kernel matrix is a reasonable symmetric positive semi-definite matrix of the original data.

In contrast to traditional MKL, our method can obtain more information with the same number of kernels. Traditional MKL can be seen as a special case of NCMK if the weights of all newly generated kernel matrices are fixed as zero. Experimental results show that SVMs with information learned from this interaction of different kernels by NCMK can achieve better performance.

II. BRIEF REVIEW OF SVMs AND MKL

A. SVMs

Given a training set $\{(x_i, y_i)\}_{i=1}^N$, $x_i \in R^m$ and $y_i \in \{\pm 1\}$, SVMs aim to find the optimal hyperplane that separates the two classes with a large margin.

1) *Hard Margin SVMs*: When the data are linearly separable in a feature space, the problem can be formulated as [1]:

$$\begin{aligned} \min_{\mathbf{w}, b} \quad & \frac{1}{2} \langle \mathbf{w}, \mathbf{w} \rangle \\ \text{s.t.} \quad & y_i (\langle \mathbf{w}, \Phi(x_i) \rangle + b) \geq 1, \quad i = 1, 2, \dots, N. \end{aligned} \quad (1)$$

The corresponding dual problem in terms of α_i (which represents \mathbf{w} through $\sum_{i=1}^N \alpha_i y_i \Phi(x_i)$) is

$$\begin{aligned} \max_{\alpha} \quad & \alpha' \mathbf{e} - \frac{1}{2} \alpha' G(K) \alpha \\ \text{s.t.} \quad & \alpha \geq 0, \quad \alpha' \mathbf{y} = 0, \end{aligned} \quad (2)$$

where vector $\alpha = [\alpha_1, \dots, \alpha_N]^T$, $G(K) = \text{diag}(\mathbf{y}) K \text{diag}(\mathbf{y})$ and \mathbf{e} is a vector whose entries are all ones.

2) *Soft Margin SVMs*: When the two classes are not linearly separable (e.g., due to noise), the misclassified patterns should be penalized. Thus, an extra term ξ^k is introduced to relax the condition for the optimal hyperplane and the problem becomes [7]:

$$\begin{aligned} \min_{\mathbf{w}, b} \langle \mathbf{w}, \mathbf{w} \rangle + C \sum_{i=1}^N \xi_i^k \\ \text{s.t. } y_i(\langle \mathbf{w}, \Phi(x_i) \rangle + b) \geq 1 - \xi_i, \\ \xi_i \geq 0, \end{aligned} \quad (3)$$

where C is a regularization parameter controlling the trade-off between maximizing the margin and minimizing the training error. For $k = 1$ and $k = 2$, it is called the 1-norm and 2-norm soft margin problem, respectively.

The dual problem for the 1-norm soft margin SVM is

$$\begin{aligned} \max_{\alpha} \alpha' \mathbf{e} - \frac{1}{2} \alpha' G(K) \alpha \\ \text{s.t. } 0 \leq \alpha \leq C \mathbf{e}, \alpha' \mathbf{y} = 0, \end{aligned} \quad (4)$$

and for the 2-norm margin SVM is

$$\begin{aligned} \max_{\alpha} \alpha' \mathbf{e} - \frac{1}{2} \alpha' (G(K) + \frac{1}{C} I) \alpha \\ \text{s.t. } \alpha \geq 0, \alpha' \mathbf{y} = 0. \end{aligned} \quad (5)$$

B. MKL

Recent studies show that using multiple different kernels instead of a single kernel improves the performance of SVMs. The strategy of MKL is to learn a weighted sum of different kernels. The combined kernel is defined as:

$$k(x_i, x_j) = \sum_{m=1}^M \sigma_m k_m(x_i, x_j), \quad (6)$$

where each kernel k_m is associated with a Hilbert space F_m , M is the total number of kernels, and $\{\sigma_m\}_{m=1}^M$ are coefficients to be learned under the convex combination constraints $\sigma_m \geq 0$ and $\sum_{m=1}^M \sigma_m = 1$.

As an instantiation, the objective function of MKL for 1-norm soft margin SVMs is formulated as follows:

$$\begin{aligned} \min_K \max_{\alpha} \alpha' \mathbf{e} - \frac{1}{2} \alpha' G(\sum_{m=1}^M \sigma_m K_{tr, m}) \alpha \\ \text{s.t. } 0 \leq \alpha \leq C \mathbf{e}, \alpha' \mathbf{y} = 0, \\ \sum_{m=1}^M \sigma_m K \succeq 0, \\ \text{trace}(K) = c, \end{aligned} \quad (7)$$

where $K_{tr, m}$ ($m = 1, \dots, M$) is obtained from training sets.

III. PROPOSED METHOD

In this section, we develop SVMs with NCMK. Both hard margin and soft margin SVMs can be integrated with NCMK and transformed into SDP problems.

A. SVMs with NCMK

NCMK is motivated by the justifiable assumption that the nonlinear combination of different kernels can yield important information for classification. In our approach, several new kernel matrices are given by the Hadamard product of any two kernel matrices computed from original kernels. The final kernel matrix used in SVMs is the weighted sum of these new kernel matrices and the original kernel matrices. The optimal weight for each kernel matrix will be calculated after we formulate the problem into an SDP problem.

The following closure properties of kernels, definition and theorem indicate that the Hadamard product is suitable for generating new kernel matrices.

Closure properties. Let k_1 and k_2 be kernels over $X \times X$, $X \subseteq \mathbb{R}^N$, $a \in \mathbb{R}^+$. Then the functions below are kernels [8]:

- 1) $k(x, z) = k_1(x, z) + k_2(x, z)$,
- 2) $k(x, z) = ak_1(x, z)$,
- 3) $k(x, z) = k_1(x, z)k_2(x, z)$.

Definition. For two matrices $A = [a_{ij}]$ and $B = [b_{ij}]$ of the same dimensions $m \times n$, we have the Hadamard product $A \odot B$ defined as $A \odot B = [a_{ij}b_{ij}]$ [9].

Theorem. If matrices A and B are positive semi-definite, then the Hadamard product $A \odot B$ is also positive semi-definite [9].

For kernel matrices K_i, K_j ($i, j = 1, \dots, M$), we define the newly formed kernel matrices \tilde{K}_t ($t = 1, \dots, M(M+1)/2$) by a nonlinear transformation as:

$$\tilde{K}_t = \begin{cases} K_i \odot K_j, & \text{for } i \neq j \\ K_i, & \text{for } i = j. \end{cases} \quad (8)$$

The definition and theorem for the Hadamard product indicate that the matrix \tilde{K}_t derived is positive semi-definite and the third closure property of kernels shows that \tilde{K}_t includes the inner product of the images of two original inputs in some feature spaces. In other words, \tilde{K}_t is a valid kernel matrix.

Therefore, the final kernel matrix \tilde{K} which includes both the new and original kernel matrices is denoted by:

$$\tilde{K} = \sum_{t=1}^{M(M+1)/2} \sigma_t \tilde{K}_t,$$

where σ_t ($t = 1, \dots, M(M+1)/2$) is the weight of each kernel matrix. According to the first and the second closure property, \tilde{K} is also a valid kernel matrix.

Below we apply NCMK to hard margin and soft margin SVMs.

1) *Hard Margin SVMs with NCMK*: We substitute \tilde{K} for K in equation (2) and add two constraints $\sigma \geq 0$ ($\sigma = [\sigma_1, \dots, \sigma_{M(M+1)/2}]'$) and $\mathbf{e}'\sigma = 1$ [10] to regularize the hypothesis space of weights.

Thus, hard margin SVMs based on NCMK can be formu-

lated as follows:

$$\begin{aligned} \min_{\sigma} \max_{\alpha} \quad & \alpha' \mathbf{e} - \frac{1}{2} \alpha' G(\tilde{K}) \alpha \\ \text{s.t.} \quad & \alpha \geq 0, \\ & \alpha' \mathbf{y} = 0, \\ & \sigma \geq 0, \\ & \mathbf{e}' \sigma = 1. \end{aligned} \quad (9)$$

2) *Soft Margin SVMs with NCMK*: We adopt the same strategy to get the objective function of 1-norm soft margin SVMs from equation (4):

$$\begin{aligned} \min_{\sigma} \max_{\alpha} \quad & \alpha' \mathbf{e} - \frac{1}{2} \alpha' G(\tilde{K}) \alpha \\ \text{s.t.} \quad & 0 \leq \alpha \leq C \mathbf{e}, \\ & \alpha' \mathbf{y} = 0, \\ & \sigma \geq 0, \\ & \mathbf{e}' \sigma = 1. \end{aligned} \quad (10)$$

Similarly, in accordance with equation (5), the objective function of 2-norm soft margin SVMs is:

$$\begin{aligned} \min_{\sigma} \max_{\alpha} \quad & \alpha' \mathbf{e} - \frac{1}{2} \alpha' G(\tilde{K} + \frac{1}{C} I) \alpha \\ \text{s.t.} \quad & \alpha \geq 0, \\ & \alpha' \mathbf{y} = 0, \\ & \sigma \geq 0, \\ & \mathbf{e}' \sigma = 1. \end{aligned} \quad (11)$$

B. SDP optimization

Semi-definite programming (SDP) deals with the optimization of convex functions over the convex cone of symmetric, positive semi-definite matrices or affine subsets of this cone [11]. It solves an optimization problem with a linear objective function, and linear matrix inequality (LMI) and affine equality constraints. Below we show that our approaches can be formulated by the SDP framework.

1) *SDP for Hard Margin SVMs with NCMK*: Both the objective function and constraints of equation (9) are convex. In terms of the standard SDP formulation, it can be rewritten as

$$\begin{aligned} \min_{u, \sigma} \quad & u \\ \text{s.t.} \quad & u \geq \max_{\alpha} \alpha' \mathbf{e} - \frac{1}{2} \alpha' G(\tilde{K}) \alpha, \\ & \alpha \geq 0, \\ & \alpha' \mathbf{y} = 0, \\ & \sigma \geq 0, \\ & \mathbf{e}' \sigma = 1. \end{aligned} \quad (12)$$

We now proceed to express the first constraint as an LMI using duality and Schur complement lemma. Its Lagrangian is:

$$L(\alpha, \beta, \gamma) = \alpha' \mathbf{e} - \frac{1}{2} \alpha' G(\tilde{K}) \alpha + \beta' \alpha + \gamma \alpha' \mathbf{y}, \quad (13)$$

where β and γ are Lagrange multipliers. Then by setting the derivative of $L(\alpha, \beta, \gamma)$ with respect to α to zero, we have

$$\alpha = (G(\tilde{K}))^{-1} (\mathbf{e} + \beta + \gamma \mathbf{y}), \quad (14)$$

Plugging equation (14) into equation (12) results in

$$u \geq (\mathbf{e} + \beta + \gamma \mathbf{y})' (G(\tilde{K}))^{-1} (\mathbf{e} + \beta + \gamma \mathbf{y}).$$

This indicates that for any $u > 0$ the above constraint holds if and only if $\beta \geq 0$ and γ exists. By Schur lemma, the inequality constraint in equation (12) is now equivalent to the following LMI:

$$\begin{bmatrix} G(\tilde{K}) & \mathbf{e} + \beta + \gamma \mathbf{y} \\ (\mathbf{e} + \beta + \gamma \mathbf{y})' & u \end{bmatrix} \succeq 0. \quad (15)$$

Substituting equation (15) into (12), the problem can be expressed as the following SDP:

$$\begin{aligned} \min_{u, \sigma} \quad & u \\ \text{s.t.} \quad & \begin{bmatrix} G(\tilde{K}) & \mathbf{e} + \beta + \gamma \mathbf{y} \\ (\mathbf{e} + \beta + \gamma \mathbf{y})' & u \end{bmatrix} \succeq 0, \\ & \sigma \geq 0, \mathbf{e}' \sigma = 1, \beta \geq 0. \end{aligned} \quad (16)$$

2) *SDP for Soft Margin SVMs with NCMK*: For the optimization problem shown in equation (10), we rewrite it in standard SDP as [10]:

$$\begin{aligned} \min_{u, \sigma} \quad & u \\ \text{s.t.} \quad & u \geq \max_{\alpha} \alpha' \mathbf{e} - \frac{1}{2} \alpha' G(\tilde{K}) \alpha, \\ & 0 \leq \alpha \leq C \mathbf{e}, \alpha' \mathbf{y} = 0, \sigma \geq 0, \mathbf{e}' \sigma = 1. \end{aligned} \quad (17)$$

The problem has a different Lagrangian for the inequality constraint:

$$\begin{aligned} L(\alpha, \beta, \gamma) = & \alpha' \mathbf{e} - \frac{1}{2} \alpha' G(\tilde{K}) \alpha + \beta' \alpha + \\ & \gamma \alpha' \mathbf{y} + \delta (C \mathbf{e} - \alpha), \end{aligned} \quad (18)$$

and

$$\alpha = (G(\tilde{K}))^{-1} (\mathbf{e} + \beta + \gamma \mathbf{y} - \delta). \quad (19)$$

Thereby, we get

$$u \geq (\mathbf{e} + \beta + \gamma \mathbf{y} - \delta)' (G(\tilde{K}))^{-1} (\mathbf{e} + \beta + \gamma \mathbf{y} - \delta) + 2C \delta' \mathbf{e}. \quad (20)$$

This implies that for any $u > 0$ the above constraint holds if and only if $\beta \geq 0$, $\delta \geq 0$ and γ exists. By Schur lemma, the inequality constraint in equation (17) is equivalent to the following LMI:

$$\begin{bmatrix} G(\tilde{K}) & \mathbf{e} + \beta + \gamma \mathbf{y} - \delta \\ (\mathbf{e} + \beta + \gamma \mathbf{y} - \delta)' & u - 2C \delta' \mathbf{e} \end{bmatrix} \succeq 0. \quad (21)$$

Table I
PERFORMANCE COMPARISON ON UCI DATASETS. IN NCMK, THE ORDER OF KERNELS IS $K_1, K_2, K_3, K_{12}, K_{13}$ AND K_{23}

		Breast cancer	Heart	Sonar
K_1	σ	1	1	1
	TA	0.964	0.783	0.62
K_2	σ	1	1	1
	TA	0.885	0.59	0.72
K_3	σ	1	1	1
	TA	0.874	0.843	0.604
MKL	σ	0.217/0.783 /0	0.003/0.867 /0.13	0/1/0
	TA	0.964	0.876	0.72
NCMK	σ	0/0/0 /0.918 /0.082/0	0.167/0.166 /0.166/0.167 /0.167/0.167	0/0/0 /0/1/0
	TA	0.978	0.895	0.753

Plugging equation (21) into (17) results in the following SDP:

$$\begin{aligned}
 & \min_{u, \sigma} u \\
 & \text{s.t.} \begin{pmatrix} G(\tilde{K}) & \mathbf{e} + \beta + \gamma \mathbf{y} - \delta \\ (\mathbf{e} + \beta + \gamma \mathbf{y} - \delta)' & u - 2C\delta' \mathbf{e} \end{pmatrix} \succeq 0, \\
 & \sigma \geq 0, \mathbf{e}' \sigma = 1, \\
 & \beta \geq 0, \delta \geq 0.
 \end{aligned} \tag{22}$$

The 2-norm soft margin SVMs have the same constraints with hard margin SVMs. Consequently, they will have the same equality constraints in standard SDP. The difference between the two objective functions of the two SVMs only leads to minor difference in the inequality constraints of SDP. The SDP formulation is:

$$\begin{aligned}
 & \min_{u, \sigma} u \\
 & \text{s.t.} \begin{pmatrix} G(\tilde{K}) + \frac{1}{C} I & \mathbf{e} + \beta + \gamma \mathbf{y} \\ (\mathbf{e} + \beta + \gamma \mathbf{y})' & u \end{pmatrix} \succeq 0, \\
 & \sigma \geq 0, \mathbf{e}' \sigma = 1, \\
 & \beta \geq 0.
 \end{aligned} \tag{23}$$

So far all the SVMs with NCMK have been formulated to SDP problems which can be solved efficiently with interior-point methods. Vector α is computed by equation (14) or (19) after solving the Lagrange multipliers δ , β and γ and the optimal weights σ .

IV. EXPERIMENT

For NCMK, we performed experiments with 1-norm soft margin SVMs and compared it with single kernel SVMs and traditional MKL on datasets from the UCI repository. The original kernels are respectively defined as $k_1(\mathbf{x}_1, \mathbf{x}_2) = (1 + \mathbf{x}_1' \mathbf{x}_2)^2$, $k_2(\mathbf{x}_1, \mathbf{x}_2) = \exp(-\varepsilon(\mathbf{x}_1 - \mathbf{x}_2)'(\mathbf{x}_1 - \mathbf{x}_2))$ and $k_3(\mathbf{x}_1, \mathbf{x}_2) = \mathbf{x}_1' \mathbf{x}_2$. For K_2 , ε is 1, 1 and 2.5 for

the three datasets, respectively. Each dataset was randomly partitioned into 80% training and 20% test data.

Test set accuracy (TA) and optimal weights σ for different kernels are given in Table I, where we see that the NCMK performs best.

V. CONCLUSION AND FUTURE WORK

We devised the NCMK for kernel combinations based on the Hadamard product for SVMs, which embodies the traditional MKL as a special case. Encouraging experimental results are observed with the new method.

Future work includes extending NCMK to support vector regression and one-class SVMs.

ACKNOWLEDGMENT

The authors would like to thank the National Natural Science Foundation of China and Shanghai Educational Development Foundation for funding respectively under Project 60703005 and 2007CG30.

REFERENCES

- [1] V. Vapnik. *Statistical Learning Theory*. John Wiley & Sons, New York, 1998.
- [2] G. Lanckriet, N. Cristianini, P. Bartlett, L. Ghaoui, M. Jordan. Learning the kernel matrix with semidefinite programming. *J. Mach. Learn. Res.*, 5:27-72, 2004.
- [3] F. Bach, G. Lanckriet, M. Jordan. Multiple kernel learning, conic duality, and the SMO algorithm. *ICML*, 2004.
- [4] S. Sonnenburg, G. Rätsch, C. Schäfer, B. Schölkopf. Large scale multiple kernel learning. *J. Mach. Learn. Res.*, 7:1531 - 1565, 2006.
- [5] A. Rakotomamonjy, F. Bach, S. Cnu, Y. Grandvalet. More efficiency in multiple kernel learning. *ICML*, 2007.
- [6] M. Gönen, E. Alpaydm. Localized multiple kernel learning. *ICML*, 2008.
- [7] B. Schölkopf, A. Smola. *Learning with Kernels*. MIT Press, 2001.
- [8] N. Cristianini, J. Shawe-Taylor. *Kernel Methods for Pattern Analysis*. Cambridge University Press, 2005.
- [9] X. Zhang. *Matrix Analysis and Applications*. Tsinghua University Press, Beijing, 2005.
- [10] C. Yeh, W. Su, S. Lee. Improving efficiency of multi-kernel learning for support vector machines. *ICMLC*, 2008.
- [11] Y. Nesterov, A. Nemirovsky. Interior point polynomial methods in convex programming: Theory and applications. *SIAM*, 1994.