Kernel-Based Machine Learning with Multiple Sources of Information

Kernbasiertes Maschinelles Lernen mit mehreren Informationsquellen

Marius Kloft*, Courant Institute of Mathematical Sciences and Memorial Sloan-Kettering Cancer Center, New York, USA

* Correspondence author: mkloft@cs.nyu.edu

Summary We present a new methodology for fusing information from multiple data sources – or kernels – in machine learning. Previous approaches promoted sparse combinations of kernels, which, however, may discard important information. We present a flexible approach based on ℓ_p-norm regularization, allowing for non-sparse solutions. In a theoretical analysis we show lower and upper generalization bounds of order up to $O(M/n)$, overcoming the best previously known upper bounds for the problem, which achieved $O(\sqrt{M/n})$. The computational experiments indicate that the novel algorithms are up to two orders of magnitude faster than previous approaches. Applications to computational biology and computer vision show accuracies that go beyond the state-of-the-art. ►►► **Zusammenfassung** Diese Arbeit gibt zunächst eine grundlegende Einführung in Theorie und Praxis des Maschinellen Lernens mit multiplen Kernen und skizziert den Stand der Forschung. Weiter entwickelt die Arbeit eine neue Methodologie des Lernens mit mehreren Kernen und beweist deren Effizienz und Effektivität. Sie entwickelt Algorithmen zur Optimierung des assoziierten mathematischen Programmes, die im Vergleich zu vorherigen Ansätzen um bis zu zwei Größenordnungen schneller sind. Unsere theoretische Analyse des Generalisierungsfehlers zeigt dabei Konvergenzraten mit Ordnungen von maximal $O(M/n)$, frühere Analysen präzisierend, die bisher nur $O(\sqrt{M/n})$ erreichten. In Anwendungen auf zentrale Fragestellungen der Bioinformatik und des Maschinellen Sehens werden Vorhersagegenauigkeiten erreicht, die den bisherigen Stand der Forschung signifikant übertreffen, wodurch eine Grundlage zur Erschließung neuer Anwendungsfelder geschaffen wird.

Keywords ACM CCS, computing methodologies, kernel methods, multiple kernel learning ►►►
Schlagwörter Maschinelles Lernen, multiple Kernmethoden

1 Introduction

In machine learning, we aim at learning the unknown relation between two random variables X and Y from data $\mathcal{D} = (x_1, y_1), \ldots, (x_n, y_n)$, so that, when observing a new x (called *pattern*), we may make accurate predictions of the corresponding y (called *label*). In the information age, machine learning is becoming an increasingly important tool. For example, consider content-based information retrieval; say $\{x_1, \ldots, x_n\}$ is a set of images and the label $y_i \in \{0, 1\}$ indicates whether an image x_i contains some object, say, for example, a cat. Machine learning then

The dissertation has been recommended to the GI Dissertation Award 2011 by the Technische Universität Berlin. The examiners were Prof. Dr. Klaus-Robert Müller, Technische Universität Berlin, Prof. Peter Bartlett PhD, UC Berkeley, and Prof. Dr. Gilles Blanchard, Universität Potsdam.

computes, from a set of labeled images, a classifier that can predict the presence or absence of a cat in a new, previously unseen image.

A very modern and elegant, yet simple approach consists in the paradigm of *kernel-based machine learning*, which states that we may convert, in a principled manner, any learning algorithm that is solely based on Euclidean scalar products into a more powerful, non-linear one, by simple substituting the scalar products $\langle x_i, x_j \rangle$ by the so-called *kernel* $k(x_i, x_j)$.[1] A prominent example of a kernel-based learning machine is the *support vector machine* (SVM), which performs prediction based on the rule

$$y := \operatorname{sign}\left(\sum_{i=1}^{n} \alpha_i y_i k(x_i, x) \right), \tag{1}$$

which predicts, for any pattern x, a corresponding label y. The α_is are hereby computed by solving a certain cleverly chosen mathematical program, that is,[2]

$$\max_{\substack{\alpha_1,\ldots,\alpha_n \geq 0: \\ \sum_{i=1}^{n} \alpha_i y_i = 0}} \quad \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j k(x_i, x_j)\,. \tag{2}$$

In many modern applications, the data is characterized by multiple sources or representations of information so that multiple views of the data are available. For example, in image retrieval, an image can be represented by its color distribution over spatial tilings, but also by shape and local gradient information. Each view gives rise to a kernel k_m, $m = 1, \ldots, M$. A sophisticated way to "fuse" the information contained in the various kernels is to form a new, weighted kernel by linear combination of the many kernels, that is,

$$k = \sum_{i=1}^{M} \theta_m k_m\,, \tag{3}$$

where $\theta_m \geq 0$ specifies the weight of the mth kernel. Previous approaches to this so-called *multiple kernel learning* (MKL) require the weight vector $\theta := (\theta_1, \ldots, \theta_M)$ being sparse, that is, many of the weights θ_m are put to zero [2; 3].

Sparse kernel combinations can be easily interpreted and analyzed, but, unfortunately, often achieve suboptimal accuracies [5]. The author believes that much of the enthusiasm of scientists for sparse models in multiple kernel learning stems from a general preference or trend for sparse models in computer science and statistics. But sparse models may discard relevant and possibly complementary information, if the underlying ground truth is dense. In the discussed dissertation [13], we present and analyze a novel methodology for non-sparse information fusion in kernel-based machine learning. The three cornerstones of the thesis can be characterized as follows:

1. *Theoretical foundations*: we prove upper bounds on the statistical generalization performance of multiple kernel learning, achieving rates as fast as $O(M/n)$, considerably pushing forward the best previously known bounds of [4], who achieved $O(\sqrt{M/n})$. We also prove a lower bound, which shows that our result is tight.
2. *Algorithms*: we develop and implement new algorithms for solving the mathematical program associated with multiple kernel learning that are up to two orders of magnitude faster than the best previous algorithms for the problem.
3. *Applications*: we apply the novel methodology to challenging problems from computer vision and bioinformatics, significantly advancing the state-of-the-art.

2 Methodology

An intuitive way of extending the prediction rule (1) to multiple kernels is to replace the occurring kernel by the "weighted kernel" as given in (3). This leads to the prediction rule

$$y := \underbrace{\operatorname{sign}\left(\sum_{i=1}^{n} \alpha_i y_i \sum_{m=1}^{M} \theta_m k_m(x_i, x) \right)}_{=:f(x,\theta)}, \tag{4}$$

First, we observe that we may rescale the weight vector $\theta = (\theta_1, \ldots, \theta_M)$ by any positive factor $\mu > 0$ without changing the prediction rule. To see this, denote the right hand side of (4) by $f(x, \theta)$ and note that, for any $\mu > 0$, it holds $f(x, \theta) = f(x, \mu\theta)$.

It is thus natural to require the weights θ_m being on a meaningful scale. In the past, this problem has been addressed by requiring $\sum_{m=1}^{M} \theta_m = 1$, which, however, may be a too restrictive assumption if the optimal kernel weight is non-sparse. We propose to relax the requirement to $\|\theta\|_p = 1$, having the flexibility of a parameter p; note here that the ℓ_p-norm is defined as $\|\theta\|_p = \sqrt[p]{\sum_{m=1}^{M} \theta_m^p}$.

But how do we find good values for the weights $\theta_1 \ldots, \theta_M$? A very elegant way is to include the θ_ms into the set of optimization variables in (2). This leads to the following optimization problem:

$$\min_{\substack{\theta_1,\ldots,\theta_n \geq 0: \\ \|(\theta_1,\ldots,\theta_n)\|_p = 1}} \quad \max_{\substack{\alpha_1,\ldots,\alpha_n \geq 0: \\ \sum_{i=1}^{n} \alpha_i y_i = 0}} \quad \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j \sum_{m=1}^{M} \theta_m k_m(x_i, x_j)\,. \tag{5}$$

Solving the above problem outputs optimal variables α_i and θ_m, which are used to perform prediction based on (4). But how to solve the above optimization problem? This is discussed in the next section.

[1] A proven choice of a kernel consists in the *Gaussian kernel* $k(x_i, x_j) = \exp(-\lambda \|x_i - x_j\|)$ for some suitable choice of $\lambda > 0$.

[2] For details, see the excellent introduction to kernel-based learning by [1].

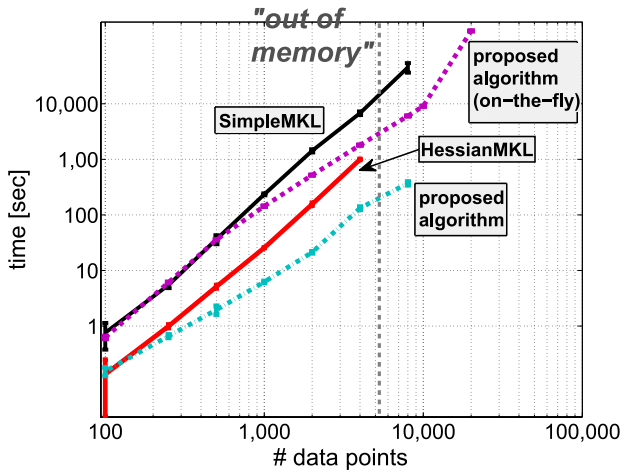

Figure 1 Runtime of proposed algorithms.

3 Algorithms

In the dissertation [13], we present three efficient algorithms for solving the optimization problem (5):

- a block coordinate descend method [5]
- a cutting plane algorithm with sequential quadratically constrained programming [6]
- a Newton descend method [7].

The most simple of the three algorithms is the block coordinate descend one, which works as follows:

Algorithm 1.

1: **initialization:**
initialize $\theta_m = \sqrt[p]{1/M}$ for all $m = 1,\ldots,M$

2: **repeat**

3: α**-step:** solve (5) with respect to variables $\alpha_1,\ldots,\alpha_n$, keeping the θ_ms fixed.

4: θ**-step:** solve the primal of (5) with respect to $\theta_1,\ldots,\theta_M$, keeping the α_is fixed

5: **until** converged

An advantage of the proposed algorithm is that it provably converges, as established by the following theorem:

Theorem 1. *If the kernels $k_1,\ldots,k_M$ are strictly positive definite, Algorithm 1 converges to a globally optimal point.*

All algorithms are implemented in C++ into the SHOGUN machine learning toolbox [8] and equipped with interfaces to MATLAB, Octave, Python und R. From the empirical analysis shown in Fig. 1, we observe that our algorithms – for the first time – facilitate to effectively employ thousands of kernels and ten thousands of data points at the same time. We observe them to be up to two orders of magnitude faster than the state-of-the-art, namely, SimpleMKL [3] and HessianMKL [9]. While the latter went out of memory for some 10 000 data points und 1000 kernels, our algorithms can deal with large-scale data by an efficient on-the-fly implementation.

4 Theoretical Analysis

The proposed methodology enjoys favorable theoretical guarantees: we show the following upper bound on the local Rademacher complexity of ℓ_p-norm multiple kernel learning [10; 11].

Theorem 2 (Rademacher bound). *The local Rademacher complexity of ℓ_p-norm multiple kernel learning is bounded by*

$$R_r(H_p) \leq \min_{t\in[q,2]} \sqrt{\frac{16}{n}\|\eta\|_{\frac{t^*}{2}}} + \frac{\sqrt{Be}CM^{\frac{1}{t^*}}t^*}{n},$$

where $\eta_m = \sum_{j=1}^{\infty} \min\left(rM^{1-\frac{2}{t^*}}, ceC^2t^{*2}\lambda_j^{(m)}\right)$, $\eta = (\eta_1,\ldots,\eta_M)$, and $\lambda_j^{(m)}$ denotes the jth eigenvalue of the mth kernel (sorted in descending order). Furthermore, we denote $B^2 := \sup_x k(x,x)$, $q = 2p/(p+1)$, and $t^* := \frac{t}{t-1}$ for the conjugated exponent of t.

We also prove a matching upper bound,

$$R_r(H_p) \geq \sqrt{\frac{c}{n}\sum_{j=1}^{\infty} \min\left(rM, D^2M^{2/q^*}\lambda_j^{(1)}\right)},$$

so that we may conclude that our result is tight. It follows the following generalization bound for learning with multiple kernels:

Theorem 3 (Generalization bound). *Suppose $\|k\|_\infty \leq B$ and $\exists d > 0$, $\alpha > 1$, so that $\forall m : \lambda_j^{(m)} \leq d_{max}j^{-\alpha}$. Then the following holds: The loss of ℓ_p-norm multiple kernel learning is, for any $p \in [1,\ldots,2]$ and $z > 0$, with probability greater equal than $1-e^{-z}$, bounded by*

$$P(l\hat{f} - l_{f^*}) \leq \min_{t\in[q,2]} 186 \cdot$$

$$\cdot \sqrt{\frac{3-\alpha_m}{1-\alpha_m}}\left(d_{\max}D^2L^2t^{*2}\right)^{\frac{1}{1+\alpha}} F^{\frac{\alpha-1}{\alpha+1}} M^{1+\frac{2}{1+\alpha}\left(\frac{1}{t^*}-1\right)} n^{-\frac{\alpha}{1+\alpha}}$$

$$+ \frac{47\sqrt{B}DLM^{\frac{1}{t^*}}t^*}{n} + \frac{(22BDLM^{\frac{1}{t^*}} + 27F)z}{n}.$$

We observe that the above bound leads to convergence rates of order up to $O(M/n)$, which considerably improves the tightest previous result, that is, the bound of order $O(\sqrt{M/n})$ proved by [4]. Note that, typically, the number of kernels, M, is much smaller than the number of data points, n. For instance, when $M = 10$ and $n = 100\,000$, the bound of [4] contains a factor of $\sqrt{M/n} = 1/100$, while our bound achieves a factor of $M/n = 1/10\,000$ – an improvement of two orders of magnitude.

5 Applications

In the application domains of computational biology and computer vision, we often encounter a multitude of complementary information sources/kernels, which renders the use of multiple kernel learning very attractive. Previous analyses – with the notable exception of the analysis of [12] on subcellular localization of proteins – failed to prove the effectiveness of multiple kernel learning. In

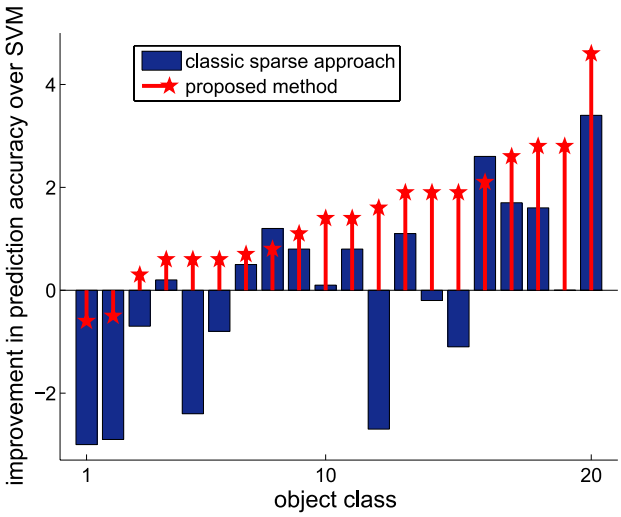

Figure 2 Accuracy of proposed methodology in an object recognition experiment.

contrast, we show that the proposed methodology can significantly raise the bar; we focus on computer vision in this presentation and refer the interested reader to the dissertation [13] for further details on applications in computational biology.

5.1 Visual Object Recognition

This area of computer vision concerns the recognition of objects in images – a difficult task because objects can be rotated, displaced, illuminated, and partially obstructed from view. Furthermore, some features may be crucial for the detection of certain object classes, but almost unnecessary for another class. For instance, color information can be very helpful to detect stop signs, but is ineffective to detect cars or balloons. Although this cries for the use of methods that incorporate multiple information sources or kernels, previous analyses did not show any advantage of those methods over a plain SVM.

We experiment on the official dataset of the PASCAL VOC Challenge 2008, which consists of 8780 images associated with up to 20 object classes. We employ multiple kernels based on color histograms of oriented gradients, visual words, and pixel colors over two color channels. In total this results in 12 kernels. We evaluate the algorithms based on the official error measure of the challenge, that is, the average prediction precision of an image averaged over all recall values. The results are shown in Fig. 2; vertical bars indicate the difference in average precision with respect to a plain SVM (using a simple kernel average). We observe that the new methodology is advantageous in 18 out of the 20 object classes, while the classic, sparse approach does not lead to consistent improvements. While the considered experiment is a test case, it paves the way for real-world deployment. For instance, the proposed methodology is an integral part of our winning submission at the recent ImageCLEF 2011 photo annotation challenge [14].

6 Conclusion

We have developed a methodology to non-sparse information fusion in kernel-based machine learning, which enjoys favorable theoretical guarantees. Our empirical analysis on challenging problems from the domains of computer vision and computation biology showed that accuracies can be achieved that go beyond the state-of-the-art. The proposed optimization algorithms were shown to be up to two orders of magnitude faster than existing ones. The method is undermined by deep foundations of statistical learning theory: we show upper and lower bounds on the generalization error of order $O(M/n)$, while previous bounds achieved $O(\sqrt{M/n})$.

Finally, we would like to remark that it might be worthwhile to rethink the current strong preference for *sparse* methods in machine learning – or in the scientific community in general. The present work clearly demonstrates that sparse models may improve over dense ones quite impressively. In fact such rethinking seems to already taking place: for instance in the social sciences, Gelman [15] claims that even in causal models "There are (almost) no true zeros"; in contrast, Gelman suggests that already weak connectivity in a causal graphical model may be sufficient for *all* variables to be required for optimal predictions (i. e., to have non-zero coefficients). The present work serves as a foundation for non-sparse information fusion in machine learning and may serve as a good starting point for further applications in science and technology.

References

[1] K.-R. Müller, S. Mika, G. Rätsch, K. Tsuda, and B. Schölkopf. An introduction to kernel-based learning algorithms. In: *IEEE Neural Networks*, 12(2):181–201, 2001.

[2] G. Lanckriet, N. Cristianini, L. E. Ghaoui, P. Bartlett und M. I. Jordan. Learning the kernel with semi-definite programming. In: *Journal of Machine Learning Research*, 5:27–72, 2004.

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

[4] C. Cortes, M. Mohri und A. Rostamizadeh. Generalization Bounds for Learning Kernels. In: *Proc. of the 27th Int'l Conf. on Machine Learning*, pp. 247–254, 2010.

[5] M. Kloft, U. Brefeld, S. Sonnenburg und A. Zien.ℓ_p-norm Multiple Kernel Learning. In: *Journal of Machine Learning Research*, 12:953–997, 2011.

[6] M. Kloft, U. Brefeld, P. Laskov und S. Sonnenburg. Non-Sparse Multiple Kernel Learning. In: *Proc. of the NIPS Workshop on Kernel Learning*, 2008.

[7] M. Kloft, U. Brefeld, S. Sonnenburg, P. Laskov, K.-R. Müller und A. Zien. Efficient and Accurate Lp-Norm Multiple Kernel Learning. In: *Advances in Neural Information Processing Systems 22*, pp. 997–1005. MIT Press, 2009.

[8] S. Sonnenburg, G. Rätsch, S. Henschel, C. Widmer, J. Behr, A. Zien, F. de Bona, A. Binder, C. Gehl und V. Franc. The SHOGUN Machine Learning Toolbox. In: *Journal of Machine Learning Research*, 11:1799–1802, 2010.

[9] O. Chapelle und A. Rakotomamonjy. Second Order Optimization of Kernel Parameters. In: *Proc. of the NIPS Workshop on Kernel Learning*, 2008.

[10] M. Kloft und G. Blanchard. The Local Rademacher Complexity of Lp-Norm Multiple Kernel Learning. In: *Advances in Neural Information Processing Systems 24*, pp. 2438–2446, 2011.
[11] M. Kloft und G. Blanchard. On the convergence rate of multiple kernel learning. In: *Journal of Machine Learning Research*, 13:2465–2502, 2012.
[12] A. Zien und C. S. Ong. Multiclass multiple kernel learning. In: *Proc. of the 24th Int'l Conf. on Machine Learning*, pp. 1191–1198, 2007.
[13] M. Kloft. *ℓ_p-Norm Multiple Kernel Learning*. Dissertation, Technische Universität Berlin, Oct 2011.
[14] A. Binder, W. Samek, M. Kloft, C. Müller, K.-R. Müller, and M. Kawanabe. The joint submission of the TU Berlin and Fraunhofer FIRST (TUBFI) to the ImageCLEF2011 Photo Annotation Task. In: *V. Petras, P. Forner, and P. D. Clough, editors, CLEF (Notebook Papers/Labs/Workshop)*, 2011. ISBN 978-88-904810-1-7.
[15] A. Gelman. Causality and Statistical Learning. In: *American Journal of Sociology*, 117(3):955–966, 2011.

Received: January 21, 2013

Dr. Marius Kloft is a postdoctoral research fellow at the Courant Institute of Mathematical Sciences and the Memorial Sloan-Kettering Cancer Center, working on problems from the domains of machine learning and cancer genetics. He is mentored by Mehryar Mohri and Gunnar Rätsch, respectively. From 2007 to 2011, he was a doctoral student at the machine learning program of Technische Universität Berlin and advised by Klaus-Robert Müller. He was co-advised by the learning theoreticians Gilles Blanchard and Peter L. Bartlett. During his doctoral studies, he spent a year abroad at the University of California, Berkeley, visiting Peter L. Bartlett's learning theory group. In 2006, he received a diploma (MSc equivalent) in Mathematics from the Philipps-Universität Marburg with a thesis in algebraic geometry. A major emphasis of his research program has been in the field of kernel methods for data fusion (multiple kernel learning). His current research focus lies, furthermore, in the field of machine learning with non-i.i.d. data and applications to genomic cancer data.

Address: Courant Institute of Mathematical Sciences, 251 Mercer Street, New York, NY 10012, USA, and Memorial Sloan-Kettering Cancer Center, 415 E 68th street, New York, NY 10065, USA, e-mail: mkloft@cs.nyu.edu, kloftm@mskcc.org

Preview on issue 3/2013

The topic of our next issue will be "High Performance Computing" (Editors: J. Teich and P. Molitor) and it will contain the following articles:

- *H. Köstler and U. Rüde:* The CSE software challenge – covering the complete stack
- *J. Reinders:* Hardware and Systems
- *D. Keyes:* The Miracle, Mandate, and Mirage of High Performance Computing
- *M. Horsch et al.:* Computational Molecular Engineering as an emerging technology in process engineering