Sparsity-accuracy trade-off in MKL

Ryota Tomioka & Taiji Suzuki*
{tomioka, t-suzuki}@mist.i.u-tokyo.ac.jp

Abstract

We empirically investigate the best trade-off between sparse and uniformly-weighted multiple kernel learning (MKL) using the elastic-net regularization on real and simulated datasets. We find that the best trade-off parameter depends not only on the sparsity of the true kernel-weight spectrum but also on the linear dependence among kernels and the number of samples.

1 Introduction

Sparse MKL is often outperformed by the simple uniformly-weighted MKL in terms of accuracy. However the sparsity offered by the sparse MKL is helpful in understanding which feature is useful and can also save a lot of computation in practice. In this paper we investigate this trade-off between the sparsity and accuracy using an elastic-net type regularization term which is a smooth interpolation between the sparse (ℓ_1-) MKL and the uniformly-weighted MKL. In addition, we extend the recently proposed SpicyMKL algorithm [1] for efficient optimization in the proposed elastic-net regularized MKL framework. Based on real and simulated MKL problems with more than 1000 kernels, we show that:

1. Sparse MKL indeed suffers from poor accuracy when the number of samples is small.
2. As the number of samples grows larger, the difference in the accuracy between sparse MKL and uniformly-weighted MKL becomes smaller.
3. Often the best accuracy is obtained in between the sparse and uniformly-weighted MKL. This can be explained by the dependence among candidate kernels having neighboring kernel parameter values.

2 Method

Let us consider the following minimization problem:

$$\underset{\boldsymbol{\alpha}_m\in\mathbb{R}^N\,(i=1,\ldots,M),\,b\in\mathbb{R}}{\text{minimize}} \quad f_\ell\Big(\sum_{m=1}^M \boldsymbol{K}_m\boldsymbol{\alpha}_m + b\mathbf{1}\Big) + C\sum_{m=1}^M\Big((1-\lambda)\|\boldsymbol{\alpha}_m\|_{\boldsymbol{K}_m} + \frac{\lambda}{2}\|\boldsymbol{\alpha}_m\|^2_{\boldsymbol{K}_m}\Big), \tag{1}$$

where N is the number of samples, M is the number of kernels, $\boldsymbol{K}_m \in \mathbb{R}^{N\times N}$ is the m-th Gram matrix, $\boldsymbol{\alpha}_m$ is the weight vector for the m-th kernel, b is the bias term, and $\mathbf{1} \in \mathbb{R}^N$ is a vector of all one. $C > 0$ is the regularization constant. In addition, f_ℓ is a loss function which we assume to be convex and twice differentiable; in this paper we use the logistic loss function. Moreover, $\|\boldsymbol{\alpha}_m\|_{\boldsymbol{K}_m} = \sqrt{\boldsymbol{\alpha}_m{}^\top \boldsymbol{K}_m\boldsymbol{\alpha}_m}$.

Here the first regularization term is a linear sum of (finite-dimensional representation of) the RKHS norms, which is known to make only few $\boldsymbol{\alpha}_m$'s non-zero (i.e., sparse); the second regularization term is a squared sum of RKHS norms. The two regularization terms are balanced by the constant λ ($0 \le \lambda \le 1$); $\lambda = 0$ corresponds to sparse (ℓ_1-) MKL and $\lambda = 1$ corresponds to uniformly-weighted MKL; i.e., at the minimum of Eq. (1), there exist a vector $\boldsymbol{\beta} \in \mathbb{R}^N$ and a sequence of non-negative numbers $d_m \ge 0$ ($m = 1, \ldots, M$), which we call a *kernel-weight spectrum*, such that $\boldsymbol{\alpha}_m = d_m\boldsymbol{\beta}$ where $(d_m)_{m=1}^M$ is sparse if $\lambda = 0$, and $d_m = 1/M$

*Both authors contributed equality to this work.

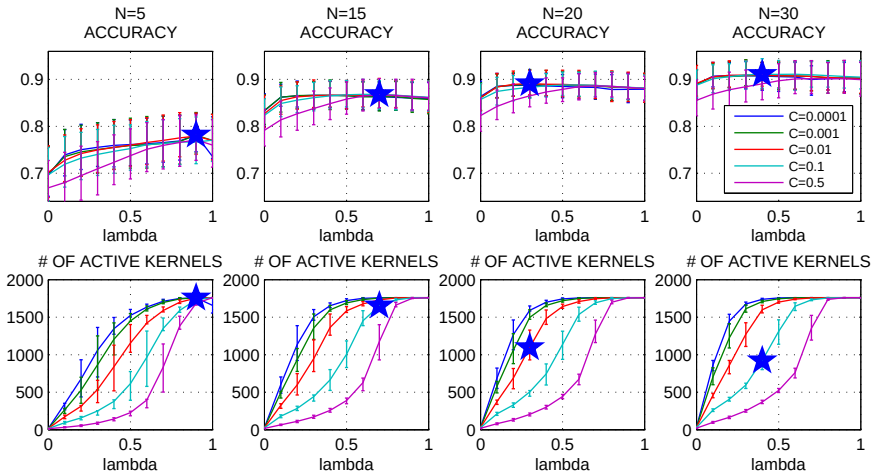

Figure 1: Image classification results from Caltech 101 dataset. The trade-off parameters λ that achieve the highest test accuracy are marked by stars.

if $\lambda = 1$. This type of regularization is known as the elastic-net regularization [2]. In the context of MKL Shawe-Taylor [3] proposed a similar approach that uses the square of the linear sum of norms in Eq. (1). Moreover, Kloft *et al.* [4] proposed another approach that connects the sparse MKL and uniformly-weighted MKL using the ℓ_p-norm. Our approach differs from [4] in that we can obtain different levels of *sparsity* for all $\lambda < 1$ (see bottom row of Fig. 1), whereas for all $p > 1$ the resulting kernel-weight spectrum is dense in [4]. Note also that uniformly-weighted MKL ($\lambda = 1$) corresponds to $p = \infty$ in [4].

3 Results

3.1 Real data

We computed 1760 kernels on 10 binary image classification problems (between "anchor", "ant", "cannon", "chair", and "cup") from Caltech 101 dataset [5] using four types of SIFT features, 22 spacial decompositions (including the spatial-pyramid kernel), two kernel functions (Gaussian and chi-squared) and 10 different kernel width parameters on each setting. See also [6] for a similar approach.

Figure 1 shows the average classification accuracy and the number of active kernels obtained at different values of the trade-off parameter λ. We can see that sparse MKL ($\lambda = 0$) can be significantly outperformed by simple uniformly-weight MKL ($\lambda = 1$) when the number of samples (N) is small. As the number of samples grows the difference between the two cases decreases. Moreover, the best accuracy is obtained at more and more sparse solutions as the number of samples grows larger.

3.2 Simulated data

In order to explain the results from the image-classification dataset in a simple setting, we generated three toy problems. In the first problem we placed one Gaussian RBF kernel over each input variable that was independently sampled from the standard normal distribution. The number of input variables was 100. We call this setting *Feature selection*. In the second problem we increased the variety of kernels by introducing 12 kernels with different band-widths on each input variable. The number of input variables was 10. We call this setting *Feature & Parameter selection*. In the third problem, we use the same 12 kernel functions with different band-widths but *jointly* over the same set of 10 input variables. We call this setting *Parameter selection*. The true kernel-weight spectrum $(d_m)_{m=1}^M$ was changed from sparse (only two non-zero d_m's), medium-dense (exponentially decaying spectrum) to dense (uniform spectrum).

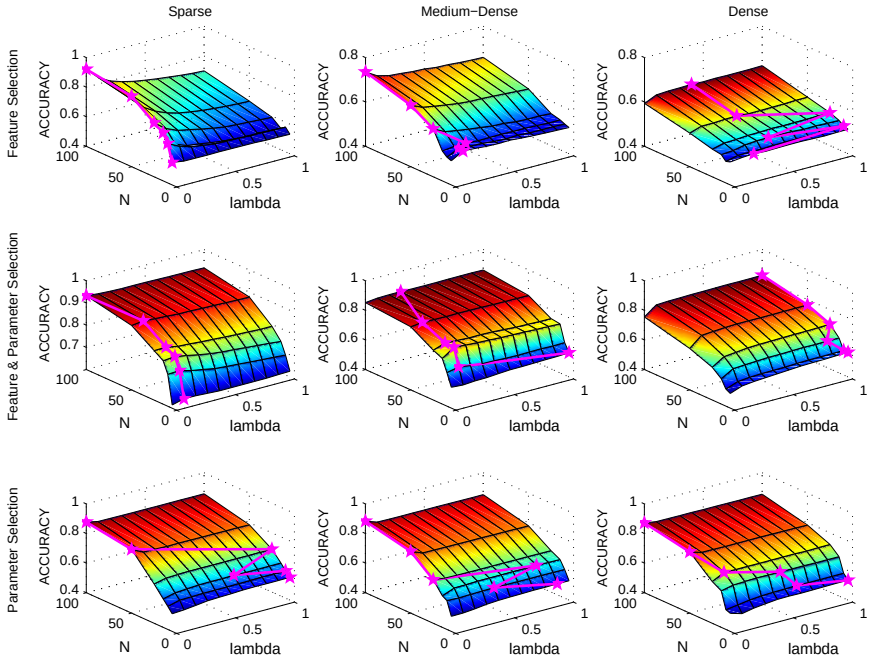

Figure 2: Classification accuracy obtained from the simulated datasets. The magenta colored curves with stars denote the value of trade-off parameters λ that yield the highest test accuracy.

Figure 2 shows the test classification accuracy obtained from training the proposed elastic-net MKL model to nine toy-problems with different goals and different true kernel-weight spectra. We choose the best regularization constant C for each plot. First we can observe that when the goal is to choose a subset of kernels from *independent* data-sources (top row), the best trade-off parameter λ is mostly determined by the true kernel-weight spectrum; i.e., small λ for sparse and large λ for dense spectrum. Remarkably the sparse MKL ($\lambda = 0$) performs well even when the number of samples is smaller than that of kernels if the true kernel-weight spectrum is sparse. On the other hand, if we also consider the selection of kernel parameter through MKL (middle row), the best trade-off parameter λ is often obtained in between zero and one and seems to depend less on the true kernel-weight spectrum. This finding seems to be consistent with the observation in [2] that the elastic-net ($0 < \lambda < 1$) performs well when the input variables are linearly dependent because kernels that only differ in the band-width can have significant dependency to each other. Furthermore, if we consider the selection of kernel parameter only (bottom row), the accuracy becomes almost flat for all λ regardless of the true kernel-weight spectrum. The behaviour in the Caltech dataset seems to be most similar to the second column of the second row (feature & parameter selection under medium sparsity).

4 Summary

In this paper, we have empirically investigated the trade-off between sparse and uniformly-weighted MKL using the elastic-net type regularization term for MKL. The sparsity of the solution is modulated by changing the trade-off parameter λ. We consistently found that, (a) often the uniformly-weighted MKL ($\lambda = 1$) outperforms sparse MKL ($\lambda = 0$); (b) the difference between the two cases decreases as the number of samples increases; (c) when the input kernels are independent, the sparse MKL seems to be favorable if the true kernel-weight spectrum is not too dense; (d) when the input kernels are linearly dependent (e.g., kernels with neighboring parameter values are included), intermediate λ value seems to be favorable. We have also observed that as the number of samples increases the sparser solution (small λ) is preferred. It was also observed (results not shown) that sparser solution is preferred when the noise in the training labels is small.

References

[1] T. Suzuki and R. Tomioka, "SpicyMKL", Technical Report arXiv:0909.5026, 2009.

[2] H. Zou and T. Hastie, "Regularization and variable selection via the elastic net", *Journal of the Royal Statistical Society Series B(Statistical Methodology)*, 67(2): 301–320, 2005.

[3] J. Shawe-Taylor, "Kernel Learning for Novelty Detection", In NIPS 08 Workshop: Kernel Learning – Automatic Selection of Optimal Kernels, 2008.

[4] M. Kloft, U. Brefeld, S. Sonnenburg, P. Laskov, K.-R. Müller, and A. Zien, "Efficient and Accurate Lp-norm Multiple Kernel Learning", in: *Advances in NIPS 22*, 2010.

[5] L. Fei-Fei, R. Fergus, and P. Perona, "Learning generative visual models from few training examples: an incremental Bayesian approach tested on 101 object categories", in: *IEEE. CVPR 2004 Workshop on Generative-Model Based Vision*, 2004.

[6] P. Gehler and S. Nowozin, "Let the kernel figure it out; principled learning of pre-processing for kernel classifiers", in: *IEEE CVPR 2009*, 2009.