Learning Multiple Visual Tasks while Discovering their Structure

Carlo Ciliberto[1,2], Lorenzo Rosasco[1,2], Silvia Villa[1]
[1]Laboratory for Computational and Statistical Learning, Istituto Italiano di Tecnologia. [2]Poggio Lab, Massachusetts Institute of Technology.

Several problems in computer vision and image processing, such as object detection/classification, image denoising, inpainting etc., require solving multiple learning tasks at the same time. In such settings a natural question is to ask whether it could be beneficial to solve all the tasks jointly, rather than separately. This idea is at the basis of the field of multi-task learning, where the joint solution of different problems has the potential to exploit tasks relatedness (structure) to improve learning. Indeed, when knowledge about task relatedness is available, it can be profitably incorporated in multi-task learning approaches for example by designing suitable embedding/coding schemes, kernels or regularizers, see [4, 6, 7].

The more interesting case, when knowledge about the tasks structure is not known a priori, has been the subject of recent studies. Largely influenced by the success of sparsity based methods, a common approach has been that of considering linear models for each task coupled with suitable parameterization/penalization enforcing task relatedness, for example encouraging the selection of features simultaneously important for all tasks [2] or for specific subgroups of related tasks [5]. One line of research has been devoted to the development of non-linear/non-parametric approaches using kernel methods [1, 3].

This paper follows this ideas, tackling in particular the development of a regularization framework to learn and exploit the tasks structure, which is not only important for prediction, but also for interpretation. Towards this end, we propose and study a family of matrix-valued reproducing kernels, parametrized so to enforce sparse relations among tasks. A novel algorithm dubbed Sparse Kernel MTL is then proposed considering a Tikhonov regularization approach.

Model

Following [7], we adopt the perspective of reproducing kernel Hilbert spaces for vector-valued functions (RKHSvv) to interpret the multiple tasks $f_1, \cdots, f_T$ that we aim to learn as the components of a vector valued predictor $f : \mathcal{X} \to \mathbb{R}^T$. A RKHSvv is a space $\mathcal{H}$ of vector-valued functions, equipped with an inner product and associated with a matrix-valued kernel $\Gamma : \mathcal{X} \times \mathcal{X} \to \mathbb{R}^{T\times T}$ for which a so-called "reproducing property" (which generalizes the reproducing property for scalar RKHS) holds. Our work focuses on "separable kernels", that is kernels of the form $\Gamma(\cdot,\cdot) = k(\cdot,\cdot)A$ with $k : \mathcal{X} \times X \to \mathbb{R}$ and $A \in S_+^T$ Positive Semidefinite (PSD) $T \times T$ matrix. In this setting we observe that, thanks to a generalization of the Representer theorem to the vector-valued case [7], each task f_t can be parametrized as

$$f_t(\cdot) = \sum_{i=1}^{n} k(\cdot, x_i)\langle A_t, c_i\rangle_{\mathbb{R}^T} = \sum_{s=1}^{T} A_{ts} g_s(\cdot) \qquad (1)$$

where the $c_i \in \mathbb{R}^T$ are coefficient vectors and $g_s(\cdot) = \sum_{i=1}^{n} k(\cdot, x_i)c_{is} \in \mathcal{H}_k$ for $s \in \{1, \ldots, T\}$. Eq. (1) shows that matrix A is encoding the tasks relations: The g_s can be interpreted as elements in a dictionary and each f_t factorizes as their linear combination. Therefore, any two predictors f_t and $f_{t'}$ are implicitly coupled by the subset of common g_s.

We consider the setting where the tasks structure is unknown and we aim to recover it from the available data in the form of a structure matrix A. Following a denoising/feature selection argument, our approach consists in imposing a sparsity penalty on the set of possible tasks structures, requiring each predictor f_t to be described by a small subset of g_s. Following the defacto standard choice of ℓ_1-norm regularization to impose sparsity in convex settings, the *Sparse Kernel MTL (SKMTL)* problem can be formulated as

$$\min_{f\in\mathcal{H}, A\in S_{++}^T} \frac{1}{n}\sum_{i=1}^{n} V(y_i, f(x_i)) + \lambda(\|f\|_{\mathcal{H}}^2 + \varepsilon\,\mathrm{tr}(A^{-1}) + \mu\,\mathrm{tr}(A) + (1-\mu)\,\|A\|_{\ell_1}) \qquad (2)$$

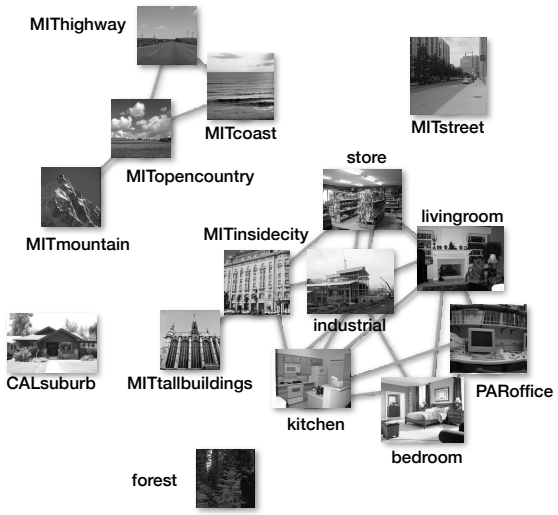

Figure 1: Tasks structure graph recovered by the Sparse Kernel MTL (SKMTL) proposed in this work on the 15-scenes dataset.

where $\|A\|_{\ell_1} = \sum_{t,s} |A_{ts}|$, $V : \mathcal{Y} \times \mathbb{R}^T \to \mathbb{R}_+$ is a loss function and $\lambda > 0$, $\varepsilon > 0$, and $\mu \in [0, 1]$ regularization parameters. Here $\mu \in [0, 1]$ regulates the amount of desired entry-wise sparsity of A with respect to the low-rank prior $\mathrm{tr}(A)$ (indeed notice that for $\mu = 1$ we recover the low-rank inducing framework of [2, 8]). This prior was empirically observed (see [2, 8]) to indeed encourage information transfer across tasks; the sparsity term can therefore be interpreted as enforcing such transfer to occur only between tasks that are strongly correlated. Finally the term $\varepsilon\,\mathrm{tr}(A^{-1})$ ensures the existence of a unique solution (making the problem strictly convex), and can be interpreted as a preconditioning of the problem.

The non-zero entries of the recovered A can be interpreted as relations between the corresponding tasks, see Fig. 1.

References

[1] M. Álvarez, N. Lawrence, and L. Rosasco. Kernels for vector-valued functions: a review. *Foundations and Trends in Machine Learning*, 4(3):195–266, 2012. URL http://dx.doi.org/10.1561/2200000036. see also http://arxiv.org/abs/1106.6251.

[2] A. Argyriou, T. Evgeniou, and M. Pontil. Convex multi-task feature learning. *Machine Learning*, 73, 2008.

[3] F. Dinuzzo, C. S. Ong, P. Gehler, and G. Pillonetto. Learning output kernels with block coordinate descent. *International Conference on Machine Learning*, 2011.

[4] Rob Fergus, Hector Bernal, Yair Weiss, and Antonio Torralba. Semantic label sharing for learning with many categories. *European Conference on Computer Vision*, 2010.

[5] Laurent Jacob, Francis Bach, and Jean-Philippe Vert. Clustered multi-task learning: a convex formulation. *Advances in Neural Information Processing Systems*, 2008.

[6] A.C. Lozano and V. Sindhwani. Block variable selection in multivariate regression and high-dimensional causal inference. *Advances in Neural Information Processing Systems*, 2011.

[7] C. A. Micchelli and M. Pontil. Kernels for multi-task learning. *Advances in Neural Information Processing Systems*, 2004.

[8] Yu Zhang and Dit-Yan Yeung. A convex formulation for learning task relationships in multi-task learning. In *Proceedings of the Twenty-Sixth Conference Annual Conference on Uncertainty in Artificial Intelligence (UAI-10)*, pages 733–742, Corvallis, Oregon, 2010. AUAI Press.

This is an extended abstract. The full paper is available at the Computer Vision Foundation webpage.