TRANSDUCTIVE KERNEL MATRIX LEARNING WITH HIERARCHIC BAYESIAN MODEL, APPLICATION TO HYPERSPECTRAL IMAGES

A. Ferrari, C. Richard, I. Smith and C. Theys

UMR 6525 H. Fizeau, Université de Nice-Sophia Antipolis, CNRS, Observatoire de la Côte d'Azur
Campus Valrose, F-06108 Nice cedex, France
{ferrari,cedric.richard,isabelle.smith,theys}@unice.fr

1. INTRODUCTION

In recent years, kernel methods have demonstrated their performance in hyperspectral imaging. Among the reasons their ability to handle large input spaces is essential. However for this type of applications a critical problem is the choice of the kernel which must combine spectral and spatial information [1] and of course achieve good generalization performance.

The kernel design stage is generally defined as the optimization of a distance metric when the kernel is chosen in a particular subspace. The alignment criterion between a parametrized kernel and a target kernel [2, 3] belongs to this family. However this solution does not embed the kernel learning problem in a particular kernel-based algorithms such as the support vector machine (SVM). This is not the case of [4] where both problems of kernel learning and SVM estimation are jointly solved in a transductive setting using semidefinite programming.

The Bayesian formalism is another powerful framework for kernel learning. In [5] a hierarchical model is used to achieve a transductive learning of the kernel matrix. In the Bayesian learning context, relevance vector machine (RVM) is the natural choice to tackle jointly the Kernel learning and estimation of the classification algorithm parameters. A solution to this problem is proposed in [6] where linear composite kernel learning and RVM regression coefficients estimation are performed using a global hierarchic Bayesian model.

This contributions proposes a general formalism for *joint transductive learning of the kernel matrix and regression coefficients estimation in a Bayesian context*.

2. HIERARCHICAL BAYESIAN MODEL

2.1. Regression model

We work in a transduction setting where some of the data are labeled and the remainder are unlabeled. The training set is $\{(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_{N_{tr}}, y_{N_{tr}})\}$ and the test set is $\{\boldsymbol{x}_{N_{tr}+1}, \ldots, \boldsymbol{x}_N\}$. The prediction is based on the classical RVM regression model $y(\boldsymbol{x})$ which employs a kernel $k(.,.)$:

$$y(\boldsymbol{x}) = \sum_{m=0}^{N_{t_r}} w_m k(\boldsymbol{x}, \boldsymbol{x}_m) \tag{1}$$

Given the training set, we assume that the targets are sample from the model with noise:

$$\forall\, n \in \{1, \ldots, N_{tr}\},\ y(\boldsymbol{x}_n) = \sum_{m=0}^{N_{tr}} w_m k(\boldsymbol{x}_n, \boldsymbol{x}_m) + e(n) \tag{2}$$

In this setting which is analogous to [4], optimization of the kernel corresponds to estimate the kernel matrix K with elements $K_{ij} = k(\boldsymbol{x}_i, \boldsymbol{x}_j)$ which can be partitioned has:

$$K = \begin{pmatrix} K_{tr} & K_{tr,t} \\ K_{tr,t}^t & K_t \end{pmatrix} \tag{3}$$

where K_{tr} defines the so-called $n_{t_r} \times n_{t_r}$ training matrix and K_t the test matrix. Denoting as $\boldsymbol{w}$ the weights vector, the model response vector is $\boldsymbol{y} = K_{tr}\boldsymbol{w} + \boldsymbol{e}$.

The a-priori on the regression model, i.e. on the noise vector and the regression coefficients, is the same as in the RVM [7] in order to guaranty sparsity on the weights vector.

- The error vector is Gaussian:
$$\boldsymbol{y}|K, \boldsymbol{w}, \sigma^2 \sim \mathcal{N}(K_{tr}\boldsymbol{w}, \sigma^2 I) \tag{4}$$
with a a uniform scale prior on σ^2.

- The weights w_m are independent zero-mean Gaussian distributed:
$$w_m|\alpha_m \sim \mathcal{N}(0, \alpha_m^{-1}) \tag{5}$$
with independent and uniform scale prior on α_m.

According to this model, we have:

$$p(\boldsymbol{y}|\boldsymbol{\alpha}, \sigma^2, K) = (2\pi)^{-N/2}|\sigma^2 I + K_{tr}A^{-1}K_{tr}^t|^{-1/2}\text{etr}\left(-\frac{1}{2}(\sigma^2 I + K_{tr}A^{-1}K_{tr}^t)^{-1}\boldsymbol{y}\boldsymbol{y}^t\right) \tag{6}$$

where $\text{etr}(.) \equiv \exp(\text{trace}(.))$ and $A = \text{diag}(\boldsymbol{\alpha})$.

2.2. Kernel model

We assume that a realization $\tilde{K}$ of the matrix K is available. This realization is obtained applying the "base" kernel $k^\circ(.,.)$ on $\mathcal{X} \times \mathcal{X}$ where $\mathcal{X} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N\}$. A classical distribution on the cone of positive semidefinite matrices is the Wishart distribution. We will assume the model $\tilde{K}|K \sim \mathcal{W}_N(\nu, K)$. This choice was justified in the context of kernel learning in [5]. As proved in [4], every positive semidefinite and symmetric matrix is a kernel matrix. This motivates the choice for the classical inverse-Wishart prior $K|\Phi \sim \mathcal{IW}_N(\eta, \Phi)$ wher Φ is assumed to be known. Straightforward computation shows that these assumptions imply $K|\tilde{K} \sim \mathcal{W}_N^{-1}(\ell, \Phi + \tilde{K})$ where $\ell = \eta + \nu$:

$$p(K|\tilde{K}, \Phi) = \frac{|\Phi + \tilde{K}|^{\frac{\ell}{2}}|K|^{-\frac{\ell+N+1}{2}}}{2^{\frac{\ell N}{2}}\Gamma_N(\frac{\ell}{2})}\text{etr}\left(-\frac{1}{2}K^{-1}(\Phi + \tilde{K})\right) \tag{7}$$

2.3. Inference

Combining Eqs. (6) and (7) we obtain the posterior density $p(K|\boldsymbol{\alpha}, \boldsymbol{Y}, \tilde{K}, \sigma^2, \Phi)$.

$$p(K|\boldsymbol{\alpha}, \sigma^2, \boldsymbol{y}, \tilde{K}) \propto |K|^{-\frac{\ell+N+1}{2}}|\sigma^2 I + K_{tr}A^{-1}K_{tr}^t|^{-1/2}\text{etr}\left(-\frac{1}{2}\left((\sigma^2 I + K_{tr}A^{-1}K_{tr}^t)^{-1}\boldsymbol{y}\boldsymbol{y}^t + K^{-1}(\Phi + \tilde{K}\right)\right) \tag{8}$$

Marginalization of this density to obtain $p(K_{tr}|\boldsymbol{\alpha}, \sigma^2, .)$ is straightforward. It is worthy to note that if $\forall m$, $\alpha_m \to \infty$ we obtain:

$$p(K_{tr}|\sigma^2, .) \propto |K_{tr}|^{-\frac{\ell+N+1}{2}}\text{etr}\left(-\frac{1}{2}K_{tr}^{-1}(\Phi_{tr} + \tilde{K}_{tr})\right) \tag{9}$$

which is maximised for $K_{tr} = (\ell + N + 1)^{-1}(\Phi_{tr} + \tilde{K}_{tr})$. The same result holds when $\sigma^2 \to \infty$. In all the others cases, the poster density of K_{tr} depends also on the "ideal kernel" $\boldsymbol{y}\boldsymbol{y}^t$.

We propose to estimate the parameters using a Metropolis Hasting MCMC procedure.

$$K^{[k+1]} \sim p(K|\boldsymbol{\alpha}^{[k]}, \sigma^{2\,[k]}, \boldsymbol{y}, \tilde{K}, \Phi) \tag{10}$$

$$\boldsymbol{\alpha}^{[k+1]} \sim p(\boldsymbol{\alpha}|K^{[k]}, \sigma^{2\,[k]}, \boldsymbol{y}) \tag{11}$$

$$\sigma^{2\,[k+1]} \sim p(\sigma^{2\,[k]}|K^{[k]}, \boldsymbol{\alpha}^{[k]}, \boldsymbol{y}) \tag{12}$$

$$\tag{13}$$

The sampling of the positive definite matrix $K^{[k+1]}$ in (10) is performed using the fact that (8) is the product of an inverse Wishart distribution on K and an inverse Wishart distribution on $\sigma^2 I + K_{tr}A^{-1}K_{tr}^t$.

3. SUMMARY

The final version will present a detailed analysis of the algorithm and a complete description of the MCMC procedure. Experimental results obtained using hyperspectral images of the Multi Unit Spectroscopic Explorer MUSE (see `http://muse.univ-lyon1.fr/`) will be presented.

4. REFERENCES

[1] G. Camps-Valls, L. Gomez-Chova, J. Munoz-Mari, J. Vila-Frances, and J. Calpe-Maravilla, "Composite kernels for hyperspectral image classification," *Geoscience and Remote Sensing Letters, IEEE*, vol. 3, no. 1, pp. 93–97, Jan. 2006.

[2] John Shawe-Taylor and Nello Cristianini, *Kernel Methods for Pattern Analysis*, Cambridge University Press, New York, NY, USA, 2004.

[3] J.-B. Pothin and C. Richard, "Optimal feature representation for kernel machines using kernel-target alignment criterion," in *ICASSP*, Honolulu, HI, 2007.

[4] Gert R. G. Lanckriet, Nello Cristianini, Peter Bartlett, Laurent El Ghaoui, and Michael I. Jordan, "Learning the kernel matrix with semidefinite programming," *J. Mach. Learn. Res.*, vol. 5, pp. 27–72, 2004.

[5] Zhihua Zhang, Dit-Yan Yeung, and James T. Kwok, "Bayesian inference for transductive learning of kernel matrix using the tanner-wong data augmentation algorithm," in *ICML '04: Proceedings of the twenty-first international conference on Machine learning*, New York, NY, USA, 2004, p. 118, ACM.

[6] Mark Girolami and Simon Rogers, "Hierarchic bayesian models for kernel learning," in *ICML '05: Proceedings of the 22nd international conference on Machine learning*, New York, NY, USA, 2005, pp. 241–248, ACM.

[7] Michael E. Tipping, "Sparse bayesian learning and the relevance vector machine," *J. Mach. Learn. Res.*, vol. 1, pp. 211–244, 2001.