LEARNING THE RELEVANT IMAGE FEATURES WITH MULTIPLE KERNELS

Devis Tuia [1], *Giona Matasci* [1], *Gustavo Camps-Valls* [2], *Mikhail Kanevski* [1]

[1] Institute of Geomatics and Analysis of Risk, University of Lausanne, Switzerland
[2] Image Processing Laboratory (IPL), Universitat de València, Spain

ABSTRACT

This paper proposes to *learn* the relevant features of remote sensing images for automatic spatio-spectral classification with the automatic optimization of multiple kernels. The method consists of building dedicated kernels for different sets of bands, contextual or textural features. The optimal linear combination of kernels is optimized through gradient descent on the support vector machine (SVM) objective function. Since a naïve implementation is computationally demanding, we propose an efficient model selection procedure based on *kernel alignment*. The result is a weight –learned from the data– for each kernel where both relevant and meaningless image features emerge after training. Excellent results are observed in both multi and hyperspectral image classification, improving standard SVM and other spatio-spectral formulations.

Index Terms— Support vector machine (SVM), Multiple kernel learning (MKL), SimpleMKL, kernel alignment, image classification.

1. INTRODUCTION

The increase in spatial and spectral resolution of satellite sensors has provided new tools for describing and modelling the Earth's surface. This allows us to indentify materials on the land cover analyzing the acquired data. These theoretical advantages also pose some hard problems and new challenges in terms of remote sensing image processing: 1) the high number of redundant bands induce collinearity problems and the well-known overfitting phenomenon and, 2) since images are also spatially redundant, this knowledge must be included in the classifier through careful spatial processing techniques. Nevertheless, the evaluation of the relevance of the extracted (both spectral and spatial, contextual or textural) features is a difficult problem. Building classifiers in such scenarios involve high-dimensional data processing and thus create the need for i) classifiers that are efficient and robust in high dimensional spaces and for ii) feature selection routines capable to select features that are discriminative to solve the problem.

Regarding classifiers, kernel methods in general and support vector machines (SVM) in particular have been shown to be robust methods capable of handling high dimensional input spaces. SVM have been successfully applied to a wide range of remote sensing problems dealing with spectral [1, 2], contextual [3] and multi-temporal and multi-source [4, 5] information. Regarding feature selection, several strategies have been discussed in the remote sensing literature. Two main strategies exist: *filter* [6] and *wrapper* [7] methods. Filter methods use an indirect measure of the quality of the selected features, e.g. evaluating the correlation or the mutual information between each input feature and the observed output (class). A faster convergence of the algorithm is thus obtained but i) they may fail to select the right subset of features if the used criterium deviates from the one used for training the learning machine, and ii) the combination of features to explain (*learn*) a problem is not considered. On the other hand, wrapper methods use as selection criteria the goodness-of-fit between the observed and provided output by the learning machine under consideration. This approach guarantees that, in each step of the algorithm, the selected subset improves performance of the previous one.

In this paper we propose an embedded (wrapper) method integrating feature selection and classification within the framework of multiple kernel learning (MKL) [8, 9, 10]. The goal is to *learn* the relevant features of remote sensing images for automatic spatio-spectral automatic classification. The method consists of building dedicated kernels for different features. The optimal linear combination of kernels is optimized through gradient descent on the support vector machine objective function. The result is a weight –learnt from the data– for each kernel where both relevant and meaningless image features emerge. Since a naïve implementation is computationally demanding, we propose an efficient model selection procedure based on kernel alignment [11]. The result is a weight –learnt from the data– for each kernel where both relevant and meaningless image features emerge after training.

The remainder of the paper is organized as follows. Section 2 revises the MKL framework. Noting that a critical point is the definition of the family of 'kernels on features', a detailed discussion on model selection and design is given in Section 3. Section 4 shows the experimental setup and obtained results. Finally Section 5 concludes the paper.

2. EFFICIENT MULTIPLE KERNEL LEARNING

This section reviews the main formulation for learning with a linear combination of kernels. This is known as multiple kernel learning (MKL). Then an efficient method for solving the problem, named SimpleMKL, is described in detail.

This work has been partly supported by the Swiss National Science Foundation (grant no.100012-113506) and by the Spanish Ministry of Education and Science under project CONSOLIDER/CSD2007-00018.

2.1. Multiple kernel learning

Kernel methods are state-of-the-art algorithms for learning from data and have demonstrated excellent results in remote sensing [5]. Both for training and prediction only defining a distance metric (the kernel) is needed. Success of kernel methods depends strongly on the data representation encoded into the kernel function: such a function defines the similarity between examples and must be chosen carefully, in order to be as discriminative as possible. For classification, the form of the decision function is: $f(x) = \sum_{i=1}^{l} y_i \alpha_i^* K(x, x_i) + b^*$, where α_i^* and b^* are coefficients to be learnt from data, y_i are training data labels and $K(\cdot, \cdot)$ is a positive definite kernel function. Common kernels, like the polynomial or the radial basis function, are rigid representations of the data, that may be replaced by more flexible and data-adapted kernels. The use of *multiple kernels* can enhance the performance of the model and, more importantly, the interpretability of the results [8]. A mutiple kernel $K(x, x')$ is a convex combination of basis kernels:

$$K(x, x') = \sum_{m=1}^{M} d_m K_m(x, x') \tag{1}$$

Multiple kernel learning aims at simultaneously optimize the α_i and the d_m subject to $d_m \geq 0$ and $\sum_{m=1}^{M} d_m = 1$. Even being a very attractive formulation, it becomes rapidly intracable with the increase of training examples and number of kernels.

2.2. Simple Multiple Kernel Learning

Simple MKL is an efficient algorithm to solve the problem [12]. Similarly to [9], Simple MKL wraps a SVM solver with a single kernel, which is already the linear combination in Eq. (1). Essentially, the algorithm is based on a gradient descent on the SVM objective value:

$$\min_{d} J(d) \quad \text{such that} \quad \sum_{m=1}^{M} d_m = 1 \, , d_m \geq 0 \tag{2}$$

where

$$J(d) = \begin{cases} \min_{\{f\},b,\xi} & \frac{1}{2} \sum_m \frac{1}{d_m} \|f_m\|^2_{\mathcal{H}_m} + C \sum_i \xi_i \\ \text{s.t.} & y_i \sum_m^M f_m(x_i) + y_i b \geq 1 - \xi_i \\ & \xi_i \geq 0 \end{cases} \tag{3}$$

Note that $J(d)$ is the formulation of a classical SVM for a kernel composed by a linear combination of sub-kernels. It can be proven that, for positive definite kernel matrices K_m, $J(\cdot)$ is convex and differentiable and can be resolved by computing the reduced gradient $\nabla_{red} J$ for each non-zero entry of $\mathbf{d}$. Such a gradient allows to compute the descent direction D for updating d. The usual update scheme for this gradient is $d \leftarrow d + \gamma D$, where γ is the optimal step size. The algorithm iterates the following steps until convergence:

1. Set $\mathbf{d} = \frac{1}{M}$. Compute $J(d)$.
2. Compute $\nabla_{red} J$. Find the descent directions D_m.
3. i. Find the maximum admissible step size, which is met when a component d_v is set to 0.
   ii. Compute the new $J(\mathbf{d})$
   iii. If $J(\mathbf{d})$ decreased, update d_v, set D_v to 0 and normalize $\mathbf{d}$.
   iv. Repeat i. to iii. until $J(\mathbf{d})$ stops decreasing.
4. Find the optimal γ by line search.

During the algorithm, the SVM solver is called several times, which may seem very costly. Nonetheless, if the α values are initialized with the values found previously, for small variations of d the SVM is solved very fast. Besides, the gradient is not computed after each update of d, but only once the objective function stops decreasing.

3. MODEL SELECTION WITH KERNEL ALIGNMENT

In the original implementation of SimpleMKL [12], model selection was performed by building several kernels with different kernel parameters. This way, the optimization of the d_m's weights allowed to find automatically the best combination by finding the non-zero weights. No *stricto-sensu* model selection is performed. It also occurs that the same variables are used several times in different kernels, providing a natural multiscale solution.

For the purpose of feature selection in multidimensional data, we may build as many kernels as we have initial variables. Thus, in the case of remote sensing images and when confronted to datasets carrying more than one hundred bands, such a strategy would imply the creation (and storage in memory) of several hundreds of kernels. This problem precludes the use of the original formulation of SimpleMKL for problems including thousands of training pixels.

To overcome this problem, in the experiments below we estimate the optimal values of kernel parameters before training the model. To do this, we use the concept of *kernel alignment* [11], a measure of similarity between matrices. Given a kernel matrix K, and a vector of labels $\mathbf{y} \in \{-1, 1\}$, the alignment between them can be written as:

$$A = \frac{\langle K, \mathbf{y}\mathbf{y}^\top \rangle_F}{\sqrt{\langle K, K \rangle_F \langle \mathbf{y}\mathbf{y}^\top, \mathbf{y}\mathbf{y}^\top \rangle_F}}, \tag{4}$$

where $\langle \cdot, \cdot \rangle_F$ stands for a Frobenius product between matrices such as $\langle M, N \rangle_F = \sum_{i,j} m_{ij} n_{ij}$. Since the kernel shows high values for similar points, the alignment can be seen as a correlation coefficient between the kernel values and the correct labels assignments. It can take values in the range $[-1, 1]$. In the case of multiclass classification, the ideal kernel $\mathbf{y}\mathbf{y}^\top$ must be replaced by a kernel returning the value 1 if the considered pixels belong to the same class and 0 otherwise. The advantage of this solution is that it speeds up consistently the analysis, because a $(n \times n \times N)$ kernel is stored and analyzed, despite of a $(n \times n \times Nk)$, where n is the number of training pixels, N is the number of features and k is the number of kernel parameters to chose upon.

In [11], kernel alignment was used to evaluate combinations of kernels: if two kernels are aligned with the labels vector and not aligned with each other, their combination will be valuable to solve the problem, because both kernels contain independent information. In our setting, we select the best candidates for SimpleMKL by maximizing the alignment of each feature's kernels K_m with the output vector. Then, SimpleMKL selects the best combination to solve the problem.

4. DATA AND EXPERIMENTS

This section describes the multi and hyperspectral image datasets, the experimental setup and shows the feature selection and classification results.

4.1. Contextual Multispectral Image Classification

The first image considered is a 0.6 m multispectral scene taken in 2004 by the QuickBird sensor over a part of the city of Zürich, Switzerland. Five classes of interest are considered: Building, Road, Vegetation, Shadow and Water. Four multispectral bands, accounting for RGB and near infrared channels and 18 spatial features extracted using opening and closing by reconstruction morphological filters are used. Three experiments are carried out:

(OM) Each of the 22 features is encoded into a separate kernel. The model selection is done as proposed in [12] (see Sect. 3). Four values of σ are considered for each feature, resulting in a total of 88 kernels.

(OA) The 22 features are considered separately, but the model selection is carried out by evaluating the alignment with the ideal kernel $K_{ideal} = \mathbf{y}\mathbf{y}^\top$.

(GM) Three groups of features are used to build 3 kernels: multispectral (4 features), opening (9) and closing (9). Model selection is carried out as in the OM experiment, for a total of 12 kernels.

In all cases, RBF kernels are used. Regularization parameter C is found by cross-validation. Experiments using $n = \{2, 5, 10, 20, 50, 70, 100\}$ labeled pixels *per* class are shown below. Validation of the models is carried out on 97,000 pixels.

4.2. Physically-based Hyperspectral Image Classification

The second case study is a hyperspectral image acquired in 1999 by the HyMap airborne spectrometer during the DAISEX99 campaign. The image has 128 bands in the region $0.4\mu m$ - $2.5\mu m$ and a resolution of 5m. The six classes of interest are: Corn, Sugar beets, Barley, Wheat, Alfalfa, and Soil [1].

Four experiments are carried out: for the OM and OA experiments, the 128 bands are used separately, in the same way as for the Zürich image. For the GM experiment, three groups of bands are chosen with respect to the physical properties of the bands [1]: leaf pigments (bands $1 - 23$), cell structure $(24 - 57)$, and leaf water content $(58 - 128)$. These three experiments' results are omitted for lack of space. In the fourth experiment, named 6MKL, the six most useful bands are first selected with SimpleMKL (using the OA strategy, see Fig. 2) experiment and then an additional OA model with these six features only is trained. Thus, in this experiments SimpleMKL is used as pure a feature selection algorithm. Results of 6MKL experiment are compared with a OA model trained using the six most important bands highlighted after physical analysis in [1]. This last model is called 6CART.

Table 1. Representative bands extracted in [1].

Bands	λ [μm]	**Characteristics**
6	0.5030	Leaf pigments (carotenes and chlorophylls).
17	0.6710	Chlorophylla-a maximum absorption.
22	0.7470	Red edge (change Visible-Near Infrared). Leaf Area Index.
24	0.7770	Beginning of Near InfraRed (NIR) with high reflectance and low absorbance. Leaf biomass and structure.
99	1.9860	Water absorption. Soil moisture and leaf water content.
118	2.3210	Water absorption. Dry matter and soil minerals.

RBF kernels are used. Regularization parameter C is found by crossvalidation. Experiments using $n = \{2, 5, 10, 20, 50, 70\}$ labeled pixels per class are considered. Validation of the models is carried out on 900 labeled pixels.

4.3. Results and discussion

Overall accuracy curves are reported in Fig. 1. This figure shows the average and standard deviations over 10 independent runs of the algorithm. For the Zürich image, the experiments with Simple MKL clearly outperform the standard SVM, showing that by weighting the importance of the features, we can construct efficient, yet *ad-hoc*, kernel machines encoding the relationship between the observed data. For the hyperspectral image, righside of Fig. 1 show the comparison between a standard SVM trained using the six bands highlighted in [1] and models built using the six most important bands highlighted by SimpleMKL. Fig. 2 illustrates the iterative optimization of weights **d** for the OM and OA experiments for the two images.

In terms of overall accuracy, the GM experiment shows the best results for the QuickBird image. The simplicity of this solution confirms the intuition that each type of information may be related to different parameters. The OA experiment shows good performances, even if inferior than the GM's: the pre-computation of the alignment avoids optimizing an 88-dimensional vector and the benefits of using such methods can be observed when few labeled examples are available. Finally, the OM experiment shows slower convergence to optimal results than the GM, but allows to visualize the chosen features (Fig. 2, top row): starting by a uniform configuration of weights ($d_m = 1/M = 0.011, \forall m \in M$), the near infrared band is given a strong weight after 5 iterations. The blue and green bands are also selected in the following steps. Regarding the morphological features, closing features related to large structuring elements are retained and all the openings is concentrated into a single feature.

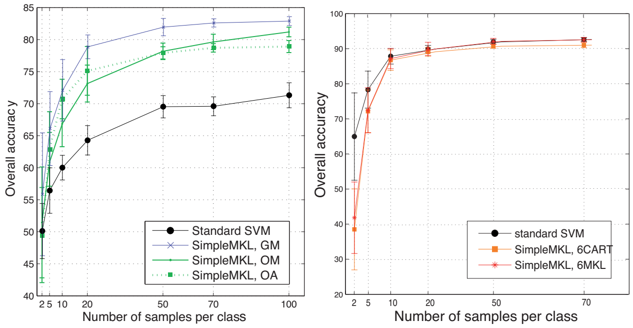

Fig. 1. Overall accuracy curves for (left) the QuickBird image and (right) the HyMap image.

Note that the solution shown in Fig. 2 is *multiscale* in the sense that two kernels using different σ are retained into the final solution for the blue or NIR bands. These kernels encode short- and middle signal relationship between training data.

Results in the HyMap data confirms the ability of SimpleMKL to select the relevant image features. The 6MKL experiment equals the performance of both the standard SVM using the 128 bands and the 6CART experiment. Nonetheless, a strong decrease in performance is observed for very small number of training pixels (2 and 5 per class), showing a tendence to overfitting for ill-posed situations. However, the overall results are similar to the ones obtained by the 6CART experiment, which is driven by physical knowledge of the problem. It is worth noting that, in this problem, the selected features by the method (Fig. 2, bottom row) are essentially the six most important spectral bands identified in [1]. Along them, some other features of interest are selected (49 and $86-88$). This fact not only confirms the correctness of the proposed model selection, but also (and even more importantly) offers a way to obtain trustable insight on the model.

5. CONCLUSION

This paper proposed multiple kernel learning to *learn* the relevant features of remote sensing images for automatic image classification. Noting the high computational cost involved, we have introduced an efficient model selection procedure based on the alignment with the ideal kernel. Excellent classification results were observed in both multi and hyperspectral images. Additionally, the model returns automatically a rank of the most relevant features and hence the most important physical signal characteristics are discovered.

6. REFERENCES

[1] G. Camps-Valls, L. Gómez-Chova, J. Calpe, and E. Soria, "Robust support vector method for hyperspectral data classification and knowledge discovery," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 42 (7), pp. 1530–1542, 2004.

[2] G. Camps-Valls and L. Bruzzone, "Kernel-based methods for hyperspectral image classification," *IEEE Trans. on Geosc. Rem. Sens.*, vol. 43, no. 6, pp. 1351– 1362, 2005.

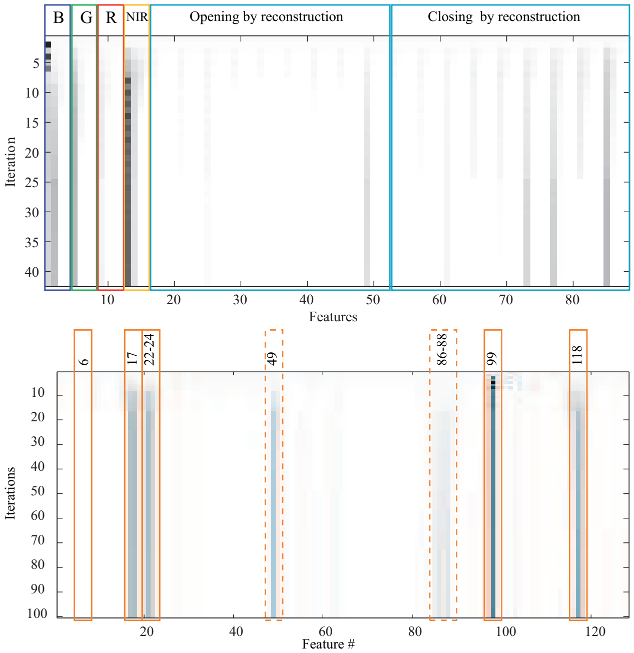

Fig. 2. Optimization of the weight vector **d**: each line corresponds to an iteration of Simple MKL. Top: QuickBird (OM experiment, 88 kernels). Bottom: HyMap image (OA experiment, 128 kernels).

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

[4] G. Camps-Valls, L. Gómez-Chova, M. Muñoz Marí, J.and Martínez-Ramón, and J. L. Rojo-Álvarez, "Kernel-based framework for multi-temporal and multi-source remote sensing data classification and change detection," *IEEE Trans. on Geosc. Rem. Sens.*, vol. 46, no. 6, pp. 1822–1835, Jun 2008.

[5] G. Camps-Valls, "New machine-learning paradigm provides advantages for remote sensing," *SPIE Newsroom*, July 2008.

[6] A. Blum and P. Langley, "Selection of relevant features and examples in machine learning," *Artificial Intelligence*, vol. 97, pp. 245–271, 1998.

[7] R. Kohavi and G. H. John, "Wrappers for features subset selection," *Int J Digit Libr*, vol. 1, pp. 108–121, 1997.

[8] G. Lancricket, T. De Bie, N. Cristianini, M. Jordan, and W. Noble, "A statistical framework for genomic data fusion," *Bioinformatics*, vol. 20, pp. 2626–2635, 2004.

[9] S. Sonnenburg, C. Schaefer G. Rätsch, and B. Schölkopf, "Large scale multiple kernel learning," *Journal of Machine Learning Research*, vol. 7, pp. 1531–1565, 2006.

[10] A. Villa, M. Fauvel, J. Chanussot, P. Gamba, and J. A. Benediktsson, "Gradient optimization for multiple kernel parameters in support vector machines classification," in *Proc. of Geosc. Rem. Sens. Symp., IGARSS08*, 2008.

[11] N. Cristianini, J. Kandola, A. Elisseeff, and J. Shawe-Taylor, "On kernel target alignment," Tech. Rep. 2001-087, NeuroCOLT, 2001.

[12] A. Rakotomamonjy, F. R. Bach, S. Canu, and Y. Grandvalet, "Simple MKL," *Journal of Machine Learning Research*, vol. 9, pp. 2491–2521, 2008.