Learning with infinitely many features

R. Flamary,
Joint work with A. Rakotomamonjy F. Yger, M. Volpi,
M. Dalla Mura, D. Tuia

Laboratoire Lagrange, Université de Nice Sophia Antipolis

December 2012

Example : image classification in remote sensing

Image classification for land-use mapping

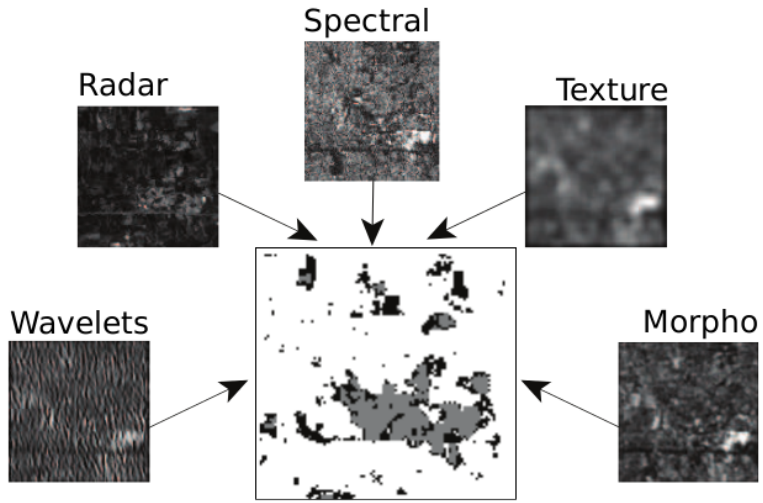
- Zurich, Switzerland
- 4 spectral bands
- 9 landuse classes
- $(329 \times 347 \times 4)$ datacube

Image (4 bands)

Labels examples

Features used in literature (parameters)

- Raw features (band)
- Morphological [4](opening/closing,shape,size,angle)
- Texture [5](mean/entropy/std,size)
- Attribute [6](area/diagonal/std)



Learning the feature extraction

Multiple-Kernel Learning [1]

The Good:

- Prediction performances.
- Learning of the feature extraction.

The Bad:

- Complexity (number of examples)

$\rightarrow$ Low rank kernels [3].

The Ugly:

- Fixed number of kernels.

$\rightarrow$ Infinite Kernel Learning [2].

Our approach

- Learn a linear classifier jointly with the feature extraction.
- Select a finite number of feature among many using sparsity promoting regularization.

Table of Contents

Notations

- $\mathcal{F}$ the set of all possible finite subset of features
- φ an element of $\mathcal{F}$ composed of d features $\{\Phi_{\theta_j}\}_{i=1}^{d}$, with θ being the feature parameter
- For an optimal $\varphi^\star$ with optimal parameters $\{\theta_j^\star\}$, the decision function writes:

$$f(\mathbf{x}) = \sum_{j=1} \mathbf{w}_j \Phi_{\theta_j^\star}(\mathbf{x}) = \mathbf{w}^T \Phi_\theta(\mathbf{x})$$

Examples of continuously parametrized features

- Wavelet or Gabor based features of the form $\langle \mathbf{x}, \psi_{j,k,\theta} \rangle$ or $\langle \mathbf{x}, \psi_{u,v,\sigma,\lambda} \rangle$
- Explicit features for kernel approximation of

$$k(\mathbf{x}, \mathbf{x}') = e^{-\sum_j \frac{(x_j - x'_j)^2}{2\sigma_j^2}}$$

Optimization problem

Formulation

$$\min_{\varphi \in \mathcal{F}} \quad \min_{\mathbf{w}} \sum_{i=1}^{n} L(y_i, \mathbf{w}^T \Phi_\theta(\mathbf{x}_i)) + \lambda \Omega(\mathbf{w})$$

- $L(\cdot, \cdot)$ convex and differentiable loss function
- $\Omega(\cdot)$ norm based sparsity-inducing regularizers
- λ : trade-off hyperparameter

Discussion

- Two-step optimization, bi-level optimization
- ERM with finite feature set φ
- Optimization over the feature set

Optimality conditions for $\Omega(\mathbf{w}) = \|\mathbf{w}\|_1$

Inner problem (restricted master):

$$\begin{aligned} \sum_i \Phi_{\theta_j}(\mathbf{x}_i) L'(y_i, \mathbf{w}^T \Phi_\theta(\mathbf{x}_i)) + \lambda sign(w_j) &= 0 \quad \text{if } w_j \neq 0 \\ \left|\sum_i \Phi_{\theta_j}(\mathbf{x}_i) L'(y_i, \mathbf{w}^T \Phi_\theta(\mathbf{x}_i))\right| &\leq \lambda \quad \text{if } w_j = 0 \text{ and } \Phi_\theta \in \varphi \end{aligned}$$

Full problem:

$$\begin{aligned} \sum_i \Phi_{\theta_j}(\mathbf{x}_i) L'(y_i, \mathbf{w}^T \Phi_{\theta_j}(\mathbf{x}_i)) + \lambda sign(w_j) &= 0 \quad \text{if } w_j \neq 0 \\ \left|\sum_i \Phi_{\theta_j}(\mathbf{x}_i) L'(y_i, \mathbf{w}^T \Phi_\theta(\mathbf{x}_i))\right| &\leq \lambda \quad \text{if } w_j = 0 \text{ and } \Phi_{\theta_j} \in \varphi \\ \left|\sum_i \Phi(\mathbf{x}_i) L'(y_i, \mathbf{w}^T \Phi_\theta(\mathbf{x}_i))\right| &\leq \lambda \quad \text{if } \Phi \notin \varphi \end{aligned}$$

Remarks

- $\leq \lambda$ constraint measure an alignement between a feature and the residue.
- Any feature $\Phi \notin \varphi$ violation the last constraint would lead to a decrease of the objective function if added to φ.
- Suggests the use of an active set algorithm.

Optimization algorithm

Active set algorithm

- Train the restricted master with a finite set of feature φ
- Select one feature ϕ violating constraints and update φ : $\varphi \leftarrow \varphi \cup \phi$
- Loop until convergence.

Practical problem

- Optimality of the full problem checked through $\max_{\Phi \notin \varphi} \left| \sum_i \Phi(\mathbf{x}_i) L'(y_i, \mathbf{w}^T \Phi_\theta(\mathbf{x}_i)) \right|$
- Depending on $L(\cdot, \cdot)$ and the structure of Φ_θ, the problem can be very difficult.

Searching for the feature

- Randomization, brute force, or clever search if applicable
  - Sample some values of θ
  - Select the feature that maximizes constraint violation.
  - sub-optimal but efficient
- ϵ-approximate solution in a finite time.

Solving the restricted master

Problem with square hinge loss

$$\min_{\mathbf{w}} \quad \sum_{i=1}^{n} \max(0, 1 - y_i \mathbf{w}^T \Phi_\theta(\mathbf{x}_i))^2 + \lambda \Omega(\mathbf{w})$$

Forward-Backward Splitting

- Use the proximal of $\Omega(\cdot)$
- Squared hinge loss differentiable and Gradient Lipschitz.
- Can be accelerated [7]

Alternating Direction Method of Multipliers

- Use the proximal of $\Omega(\cdot)$ and squared hinge loss
- variable splitting + Augmented Lagrangian [8].
- Include second order information

Extensions to other paradigms

Non-differentiable norm-based regularization term $\Omega(\mathbf{w})$

The violating constraint condition becomes

$$\Omega^\star\left(\sum_i \Phi(\mathbf{x}_i) L'(y_i, \mathbf{w}^T \Phi_{\theta_j}(\mathbf{x}_i))\right) \leq \lambda$$

with $\Omega^\star(\mathbf{w})$ being the dual norm of $\Omega(\mathbf{w})$.

Multi-task with shared features

$\ell_1 - \ell_q$ mixed-norm whose dual is $\ell_\infty - \ell_{q'}$

$$\|\mathbf{W}\|_{1,q} = \sum_{i=1}^{d} \|\mathbf{W}_{\cdot,t}\|_q$$

Dirty multitask

Shared features + mean classifier

$$\Omega(\mathbf{W}, \bar{\mathbf{w}}) = \|\mathbf{W}\|_{1,q} + \lambda_m \|\bar{\mathbf{w}}\|_1$$

with $\bar{\mathbf{w}}$ common to all tasks.

Application to kernel and multiple kernel approximation

Explicit feature map for kernel

Gaussian kernel $k(\mathbf{x}, \mathbf{x}')$

$$k(\mathbf{x}, \mathbf{x}') \approx \sum_{j=1}^{m} [\cos(\mathbf{v}_j^T \mathbf{x}) \cos(\mathbf{v}_j^T \mathbf{x}') + \sin(\mathbf{v}_j^T \mathbf{x}) \sin(\mathbf{v}_j^T \mathbf{x}')]$$

where $\{\mathbf{v}_j\}$ are random vectors samples according to the FT of the Gaussian kernel

Application for MKL in our framework

- Sample several values of the Gaussian kernel bandwidth
- For each value, draw direction vectors $\{\mathbf{v}_j\}$
- For all bandwidth and direction vectors, compute the constraint violation
- Select the pair of features violating the most their constraints.

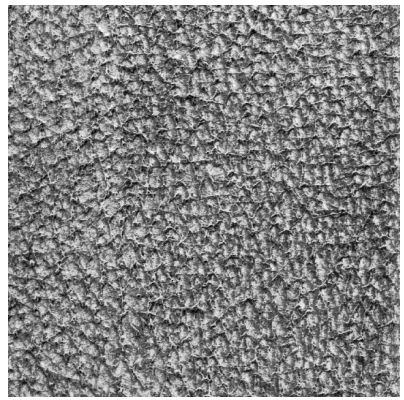
Experimental results

Image classification for land-use mapping

- Several features from literature
- Comparison with samples parameters

Texture recognition problem

- Gabor features on Brodatz dataset
- Comparison with sampled parameters



Large scale approximated kernel machines

- Using Fourier feature for approximate Gaussian kernel and MKL on Adult and IJCNN1 .
- Comparison with incomplete choleski decomposition

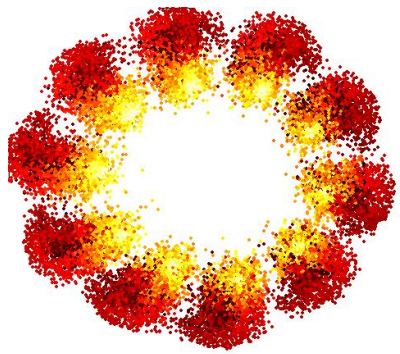
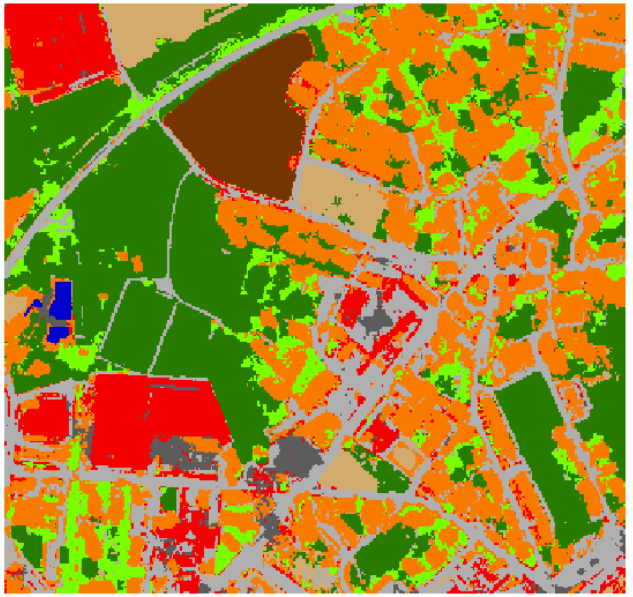


Image classification dataset

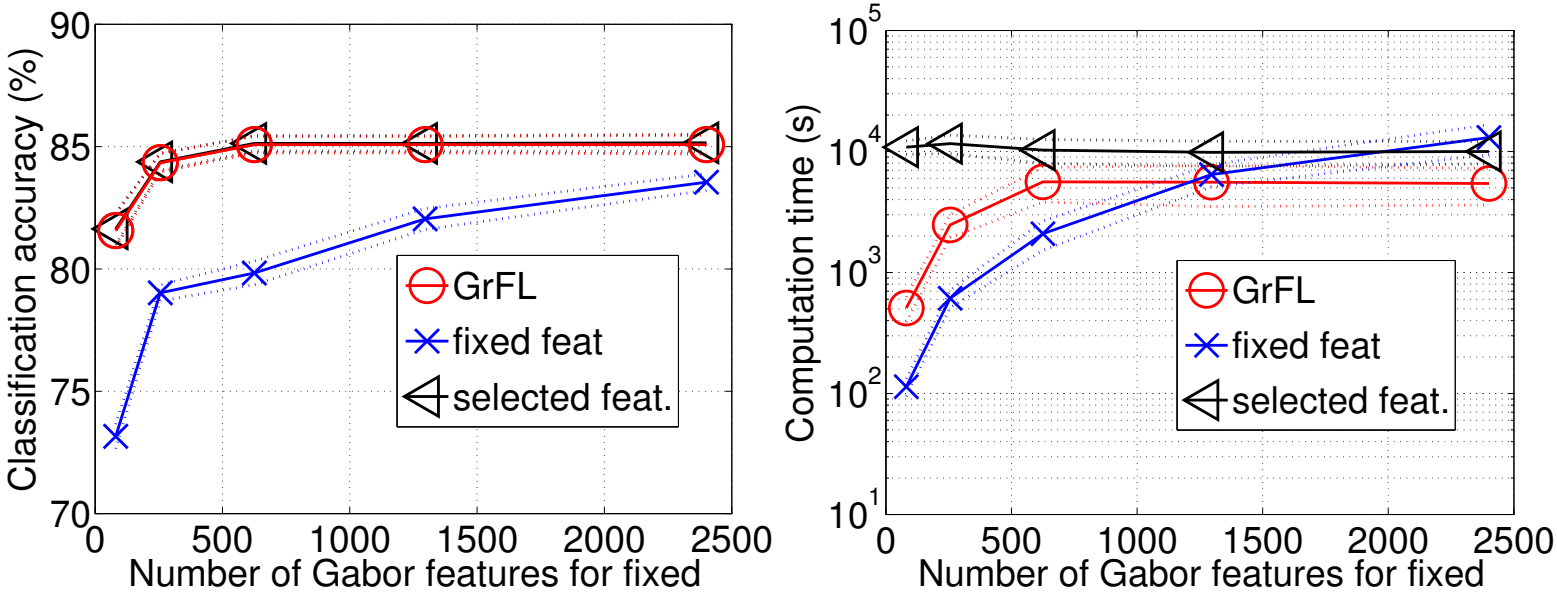
Model / Feature type	Predefined library				GrFL
	ℓ_1 SVM				
	Bands	MOR	ATT	All	
Overall accuracy	69.75	84.52	85.50	91.99	**92.46**
Cohen's Kappa	0.613	0.806	0.819	0.901	**0.907**
Residential	76.71	92.17	92.44	96.07	**96.71**
Commercial	51.49	74.02	66.42	79.65	**83.73**
Meadows	**99.93**	99.75	99.58	99.54	99.60
Harvested	0	30.47	83.24	**98.40**	97.51
Bare soil	49.53	**99.98**	99.41	99.93	99.91
Roads	88.92	84.50	84.32	88.95	**89.39**
Pools	21.09	95.47	**98.28**	97.42	96.40
Parkings	0	42.05	31.26	**56.41**	51.99
Trees	0	41.10	12.81	**65.98**	65.93
# Features	4	148	324	508	∞
# Selected	4	84.20	114.60	202.40	210.40

Classified map with GrFL



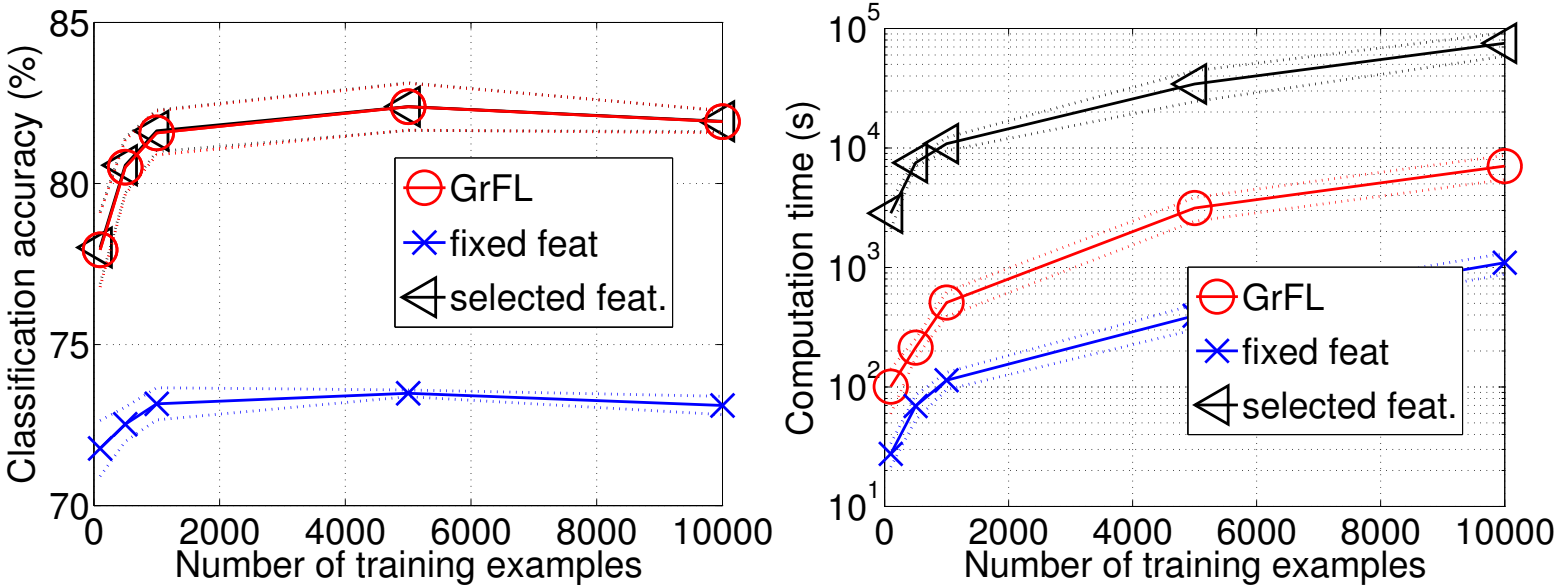
- Training, 2047 pixels, testing 38722 pixels.
- Best overall performances.
- Models using jointly several kind of features.

Gabor feature for texture recognition (1)



- 3 classes, 16×16 patches from the texture image (Brodatz)
- Increasing number of features and 1000 examples per class
- **Approaches**
  - GrFL : our method
  - Fixed feat : pre-defined features through discretization
  - Selected feat: Lasso with 3000 of the features visited by GrFL

Gabor feature for texture recognition (2)



- increasing number of training samples with 81 Gabor features

Lessons

- Learning with infinitely many cheaper than learning with many
- Do not sample parameters but take advantage of the continuous parameters

Large scale kernel machines

	Adult			IJCNN1		
# feat	GrFL	GrFL-M	IC	GrFL	GrFL-M	IC
10	**83.82**	83.77	83.38	**92.06**	91.96	91.03
50	84.76	**84.86**	84.58	**97.05**	96.97	92.19
100	84.98	**85.00**	84.84	97.97	**98.02**	93.29
500	85.24	**85.30**	85.04	–	–	–

	Adult			IJCNN1		
ratio	GrFL	GrFL-M	IC	GrFL	GrFL-M	IC
0.1	84.23	84.34	**84.54**	96.27	**96.67**	93.38
0.3	84.78	**84.87**	84.72	97.40	**97.77**	93.23
0.5	84.91	**84.95**	84.74	97.75	**97.96**	93.32
0.7	84.98	**85.00**	84.84	97.97	**98.02**	93.29

- Gaussian kernel with explicit and selected feature maps
- Datasets : *Adult* and *IJCNN1* (40k and 110k training examples)
- Sample kernel bandwidth and then sample vector direction
- Better performances than Incomplete Choleski decomposition
- Easy multiple Gaussian kernel

Conclusions

Learning with infinitely many features

- Framework is generic to loss functions and sparsity inducing regularizers.
- Works well in practice.
- Interpretability.

Questions and future works

- Theoretical guarantees when the algorithm stops at non-optimal solution?
- Data with many features, is random sampling enough?
- What is smart sampling?
- Learn directly in the Fourier feature space, fourier neural nets.

References

G. Lanckriet, N. Cristianini, L. El Ghaoui, P. Bartlett, and M. Jordan.

Learning the kernel matrix with semi-definite programming.

Journal of Machine Learning Research, 5:27–72, 2004.

P. Gehler and S. Nowozin.

Infinite kernel learning.

In *NIPS workshop on Automatic Selection of Kernel Parameters*, 2008.

F. Bach and M. Jordan.

Predictive low-rank decomposition for kernel methods.

In *Proceedings of the 22nd International Conference on Machine Learning*, 2005.

J.A. Benediktsson, J. A. Palmason, and J. R. Sveinsson.

Classification of hyperspectral data from urban areas based on extended morphological profiles.

IEEE Trans. Geosci. Remote Sens., 43(3):480–490, 2005.

F. Pacifici, M. Chini, and W.J. Emery.

A neural network approach using multi-scale textural metrics from very high-resolution panchromatic imagery for urban land-use classification.

Remote Sens. Environ., 113(6):1276–1292, 2009.

M. Dalla Mura, J. Atli Benediktsson, B. Waske, and L. Bruzzone.

Morphological attribute profiles for the analysis of very high resolution images.

IEEE Trans. Geosci. Remote Sens., 48(10):3747–3762, 2010.

A. Beck and M. Teboulle.

A fast iterative shrinkage-thresholding algorithm for linear inverse problems.

SIAM Journal on Imaging Sciences, 2:183–202, 2009.

Algorithmic implementation: ADMM (1)

- Randomization for feature searching
- minimization of empirical risk + sparse regularizer for the inner problem
  - fast proximal algorithm or alternate direction methods of multipliers
- Instantiation with square hinge loss of the ADMM approach

$$\min_{\mathbf{w}} \max(0, 1 - \mathbf{y}\Phi\mathbf{w})^T \max(0, 1 - \mathbf{y}\Phi\mathbf{w}) + \lambda\Omega(\mathbf{w})$$

- variable splitting

$$\begin{array}{ll} \min_{\mathbf{u},\mathbf{v},\mathbf{w}} & \max(0, \mathbf{u})^T \max(0, \mathbf{u}) + \lambda\Omega(\mathbf{v}) \\ & \mathbf{u} = 1 - \mathbf{y}\Phi\mathbf{w} \\ & \mathbf{v} = \mathbf{w} \end{array}$$

decouples the influence of the loss and the regularizer in the optimization problem.

Algorithmic implementation : ADMM (2)

- Lagrangian

$$\begin{aligned}\mathcal{L} &= \max(0,\mathbf{u})^T \max(0,\mathbf{u}) + \lambda\Omega(\mathbf{v}) + \alpha^T(\mathbf{u} - 1 + \mathbf{y}\Phi\mathbf{w}) \\ &\quad + \beta^T(\mathbf{v} - \mathbf{w}) + \frac{\nu}{2}\|\mathbf{u} - 1 + \mathbf{y}\Phi\mathbf{w}\|^2 + \frac{\nu'}{2}\|\mathbf{v} - \mathbf{w}\|^2\end{aligned}$$

- Iteration
  - minimization of the augmented Lagrangian wrt to each single primal variable
  - update of the dual variable α, β
- Steps :
  - linear system for $\mathbf{w}$
  - proximal operator update for $\mathbf{u}$ related to the loss function
  - proximal operator update for $\mathbf{v}$ related to the regularizer
- Nice points
  - simple and generic
  - convergence for inexact proximal operators
  - efficient