Selecting from an infinite set of features in SVM

Rémi Flamary, Florian Yger, Alain Rakotomamonjy

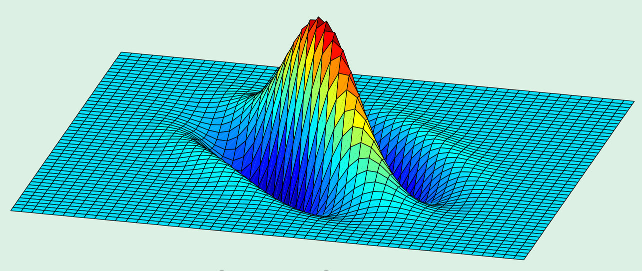
LITIS EA 4108, Université de Rouen
76800 Saint Etienne du Rouvray, France

How to extract features?

- Continuous parameters for feature extractions:
  $\Rightarrow$ Infinite set.
- Select from a finite number of values by Cross Validation or MKL [1]: limited to small number of parameters.
- Infinite MKL [2] for continuous parameters: limited to small scale datasets.
- We propose an active set algorithm for feature extraction and classifier learning: learning from continuously parametrized features for large scale datasets.

Examples of infinite sets

- 2D Gabor functions for texture recognition.



$$g(u,v) = e^{-(\frac{u^2}{2\sigma_1}+\frac{v^2}{2\sigma_2})} e^{i\pi f(u\cos\theta + v\sin\theta)}$$

  4 parameters: $\theta, f, \sigma_1, \sigma_2$.
- Signal filtering for Brain-Computer Interfaces. For Motor Imagery, a $[f_{min}, f_{max}]$ bandpass filtering is applied to the signals.
  2 parameters: f_{min}, f_{max}.

Framework

- n training examples $\{\mathbf{x}_i, y_i\}_{i=1}^n$ with $\mathbf{x}_i \in \mathcal{X}$ and $y_i \in \{-1, 1\}$.
- $\phi_\theta(\cdot)$ is a θ parametrized feature extraction.
- The decision function is:

$$f(\mathbf{x}) = \sum_{j=1}^{N} \langle \mathbf{w}_j, \phi_{\theta_j}(\mathbf{x}) \rangle_{\mathcal{X}_{\theta_j}} \qquad (1)$$

  where some of the $\mathbf{w}_j$ are 0.
- Φ is the matrix of feature maps, resulting from the concatenation of the N matrices $\{\Phi_{\theta_j}\}$.
- Φ is normalized to unit norm and $\widetilde{\Phi} = \text{diag}(\mathbf{y})\Phi$.

Fixed number of features

- Optimization problem:

$$\min_{\mathbf{w},b} \quad J(\mathbf{w}) = \frac{C}{2n}(\mathbb{I}-\widetilde{\Phi}\mathbf{w})_+^T(\mathbb{I}-\widetilde{\Phi}\mathbf{w})_+ + \Omega(\mathbf{w}) \qquad (2)$$

  where $[\widetilde{\Phi}\mathbf{w}]_i = f(\mathbf{x}_i)$, $\mathbb{I}$ is a unitary vector, $(\cdot)_+ = \max(0, .)$ is the element-wise positive part of a vector, Ω is a $\ell_1 - \ell_2$ norm.
- Optimality conditions are:

$$\begin{aligned} -\mathbf{r}_i + \frac{\mathbf{w}_i}{||\mathbf{w}_i||_2} &= \vec{0} \quad \forall i \ \mathbf{w}_i \neq \vec{0} \\ ||\mathbf{r}_i||_2 &\leq 1 \quad \forall i \ \mathbf{w}_i = \vec{0} \end{aligned} \qquad (3)$$

  with $\mathbf{r}_i = \frac{C}{n}\widetilde{\Phi}_i^T(\mathbb{I} - \widetilde{\Phi}\mathbf{w})_+$.

Active Set Algorithm

1: Set $\mathcal{A} = \emptyset$ initial active set
2: Set $\mathbf{w} = \vec{0}$
3: **repeat**
4: $\quad$ $\mathbf{w} \leftarrow$ solve problem (2) with features from $\mathcal{A}$
5: $\quad$ $r, i \leftarrow \max_{i \in \mathcal{A}^c} \quad ||\mathbf{r}_i||_2$
6: $\quad$ **if** $r > 1$ **then**
7: $\quad\quad$ $\mathcal{A} = \mathcal{A} \cup i$
8: $\quad$ **end if**
9: **until** $r \leq 1$

- The most violating feature is added for convergence speed (Line 5).
- Sub-problem solved quickly with an Fast Iterative Shrinkage Algorithm [3] (Line 4).

Extension to the infinite set

- Aim: find a finite set Θ of features minimizing $J(\mathbf{w})$.
- The new optimality conditions are:

$$\begin{aligned} -\mathbf{r}_i + \frac{\mathbf{w}_i}{||\mathbf{w}_i||_2} = \vec{0} \quad & \forall i \ \mathbf{w}_i \neq \vec{0} \\ ||\mathbf{r}_i||_2 \leq 1 \quad & \forall i \ \mathbf{w}_i = \vec{0} \\ ||\widetilde{\Phi}_{\theta_s}^T(\mathbb{I} - \widetilde{\Phi}\mathbf{w})_+||_2 \leq 1 \quad & \forall \ \theta_s \notin \Theta \end{aligned} \qquad (4)$$

- Not possible to check optimality $\forall \theta$.
- Optimality checked on a randomly drawn finite set of $\theta_s \notin \Theta$ (Line 5).
- Add the most violating feature from this random subset to the active set.

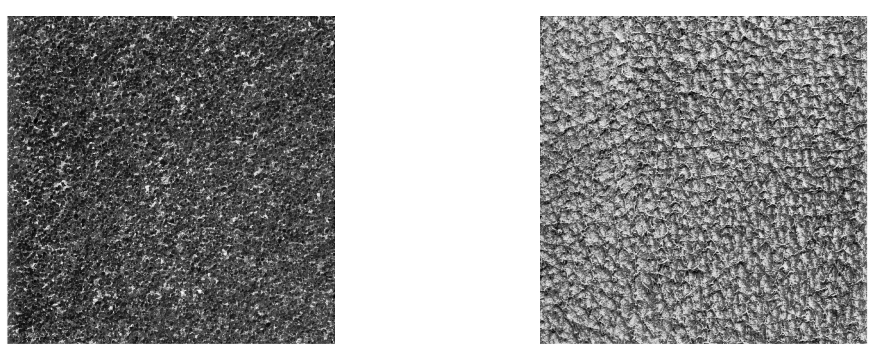


Figure 1: Textures D29 (left) and D92 (right) from the Brodatz Dataset

Texture Recognition Dataset

- Classifying 16×16 patches from Brodatz textures D29 and D92.
- Fixed and random 2D Gabor marginal features compared.
- C has been set to 10.

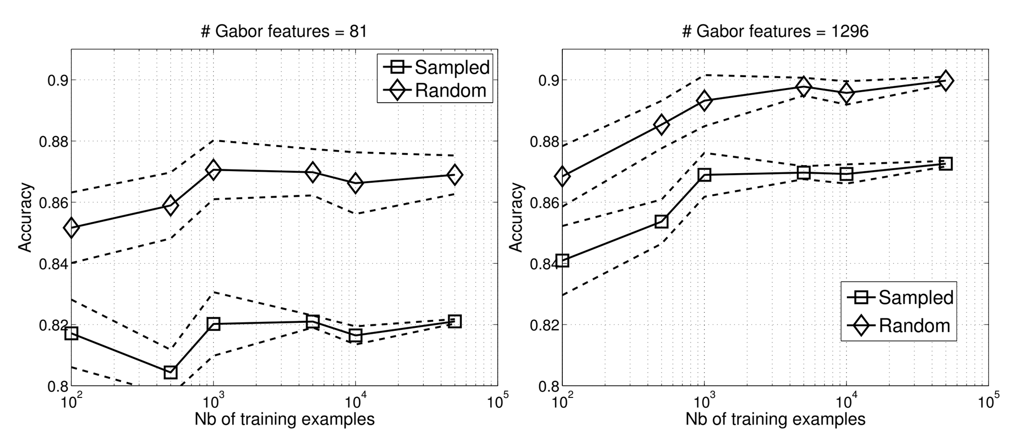


Figure 2: Accuracy performance with different numbers of sampled features (left) 81. (right) 1296.

BCI Dataset

- Dataset IIa from BCI Competition IV.
- Comparison between a fixed [8,30]Hz bandpass and a random bandpass of at least 20Hz inside [8,30]Hz.
- A CSP [4] is applied to the filtered signals and the most discriminant spatial filters are kept.
- The number of selected filters and C are chosen through Cross-Validation.

	Subjects									
Methods	S1	S2	S3	S4	S5	S6	S7	S8	S9	Avg
CSP [4]	88.89	51.39	**96.53**	70.14	54.86	71.53	81.25	93.75	**93.74**	78.01
Fixed	88.19	**53.47**	**96.53**	63.89	60.42	69.44	79.17	**97.92**	93.06	78.01
Random	**90.97**	52.78	95.14	**73.61**	**62.50**	**72.92**	**82.64**	97.22	92.36	**80.01**

Table 1: Classification accuracy on the test set for classical CSP approach, fixed and random bandpass filter for feature extraction on the BCI dataset.

Conclusion

- Active set algorithm.
- Handle large scale problems.
- Automated selection of continuous parameters.

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.

Peter Gehler and Sebastian Nowozin.
Let the kernel figure it out: Principled learning of pre-processing for kernel classifiers.
In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2009.

A. Beck and M. Teboulle.
A fast iterative shrinkage-thresholding algorithm for linear inverse problems.
SIAM Journal on Imaging Sciences, 2:183–202, 2009.

F. Lotte and C. Guan.
Regularizing common spatial patterns to improve bci designs: Unified theory and new algorithms.
IEEE Trans Biomed Eng, to appear, 2010.