

Pyramides de noyaux

Marie SZAFRANSKI¹, Yves GRANDVALET²

¹Laboratoire d'Informatique Fondamentale
Centre de Mathématiques et Informatique, 39 rue Joliot-Curie, 13453 Marseille Cedex 13, France

²Laboratoire HeuDiaSyC
Université de Technologie de Compiègne, Centre de Recherches de Royallieu, BP 20529, 60205 Compiègne Cedex, France
marie.szafranski@lif.univ-mrs.fr, yves.grandvalet@utc.fr

Résumé – L'apprentissage statistique vise à prédire, mais aussi analyser ou interpréter un phénomène. Pour réaliser ces objectifs, le choix de la représentation des données joue un rôle crucial. Nous proposons de guider le processus d'apprentissage en y intégrant une connaissance relative à la façon dont sont organisées les similarités entre exemples. La connaissance est représentée par une "pyramide de noyaux", une structure arborescente qui permet d'organiser des groupes et sous-groupes distincts de similarités. Quand peu de (groupes de) similarités sont pertinentes pour discriminer les observations, notre approche fait émerger ces similarités. Nous proposons ici la première solution complète à ce problème, permettant l'apprentissage d'un séparateur à vaste marge (SVM) sur des pyramides de noyaux de hauteur arbitraire. Les pondérations des (groupes de) similarités sont apprises conjointement avec les paramètres du SVM, par optimisation d'un critère que nous montrons être une formulation variationnelle d'un problème régularisé par une norme mixte. Nous illustrons notre approche sur un problème de reconnaissance d'expressions faciales, où les caractéristiques des images sont décrites par une pyramide représentant l'organisation spatiale et l'échelle des filtres d'ondelettes appliqués sur des patches d'images.

Abstract – Statistical learning aims at predicting, but also at analyzing and interpreting a phenomenon. In this process, data representation is a crucial issue. We propose to guide the learning process by providing it with a prior knowledge describing how similarities between examples are organized. This knowledge is encoded as a "kernel pyramid", that is, a tree structure that represents nested groups and sub-groups of similarities. Provided few (groups of) similarities are relevant for the classifying the observations, our approach identifies these similarities. We propose herein the first complete solution to this problem, enabling to learn a Support Vector Machine (SVM) on pyramids of arbitrary heights. A weighted combination of (groups of) similarities is learned jointly with the SVM parameters, by optimizing a criterion that is shown to be a variational formulation of the original fitting problem penalized by a mixed norm. We illustrate our approach on the recognition of facial expressions from still images, whose features are described by a pyramid depicting the spatial and scale parameters of the wavelet decomposition of the original image.

1 Méthodes à noyaux pour la classification supervisée

L'objectif de la classification supervisée est d'estimer une fonction de décision prédisant la classe y d'une observation \mathbf{x} . Pour la classification binaire, on dispose d'un ensemble d'apprentissage $S = \{(\mathbf{x}_i, y_i)\}_{i=1}^n$, où $(\mathbf{x}_i, y_i) \in \mathcal{X} \times \{\pm 1\}$. Dans les méthodes à noyaux, comme les Séparateurs à Vaste Marge (SVM), les observations sont implicitement projetées dans un espace de caractéristiques par le biais d'une transformation $\phi : \mathcal{X} \rightarrow \mathcal{H}$ où \mathcal{H} est un Espace de Hilbert à Noyau Reproduisant (EHNR) et $K : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$ est le noyau associé. Le rôle premier de K consiste à définir une fonctionnelle d'évaluation dans $\mathcal{H} : \forall f \in \mathcal{H}, f(\mathbf{x}) = \langle f, K(\mathbf{x}, \cdot) \rangle_{\mathcal{H}}$, et il définit également :

i) l'espace \mathcal{H} puisque $\forall f \in \mathcal{H}, f(\mathbf{x}) = \sum_{i=1}^{\infty} \alpha_i K(\mathbf{x}, \mathbf{x}_i)$;

ii) une métrique : $\|f\|_{\mathcal{H}}^2 = \sum_{i=1}^{\infty} \sum_{j=1}^{\infty} \alpha_i \alpha_j K(\mathbf{x}_i, \mathbf{x}_j)$;

iii) la transformation $\phi(\mathbf{x}) = K(\mathbf{x}, \cdot)$ et une similarité entre observations par $K(\mathbf{x}, \mathbf{x}') = \langle \phi(\mathbf{x}), \phi(\mathbf{x}') \rangle_{\mathcal{H}}$.

Ces propriétés montrent à la fois l'importance et la difficulté du choix du noyau dans les SVM. Ce rôle crucial a motivé l'introduction de techniques d'apprentissage du noyau, en particulier celle de Lanckriet *et al.* [1].

Dans le cadre SVM standard, on estime la fonction discriminante f dans un EHNR prédéterminé \mathcal{H} , en résolvant

$$(f^*, b^*) = \arg \min_{(f, b)} \frac{1}{2} \|f\|_{\mathcal{H}}^2 + C \sum_{i=1}^n [1 - y_i(f(\mathbf{x}_i) + b)]_+,$$

où $[u]_+ = \max(0, u)$, et $C > 0$ est le paramètre de régularisation qui contrôle le compromis entre la taille de la marge et l'adéquation aux données. La fonction de décision résultante est de la forme $\text{sign}(f^*(\mathbf{x}) + b^*)$.

L'apprentissage de noyaux multiples (*Multiple Kernel Learning*) vise à atténuer le problème du choix de noyau en utilisant une combinaison convexe de noyaux adaptée aux données. Le problème se formalise par l'optimisation jointe des

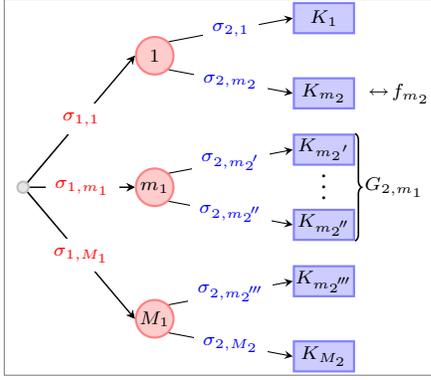


Figure 1: Exemple d'arborescence ($H = 2$).

paramètres du SVM avec ceux de la combinaison convexe de M noyaux prédéfinis. Dans ce but, l'approche de Rakotomamonjy *et al.* [2] résout :

$$(1) \begin{cases} \min_{f_1, \dots, f_M, b, \sigma} & \frac{1}{2} \sum_{m=1}^M \frac{1}{\sigma_m} \|f_m\|_{\mathcal{H}_m}^2 + \\ & C \sum_{i=1}^n [1 - y_i (\sum_{m=1}^M f_m(\mathbf{x}_i) + b)]_+, \\ \text{t.q.} & \sum_{m=1}^M \sigma_m \leq 1, \sigma_m \geq 0, \forall m, \end{cases}$$

où \mathcal{H}_m sont des EHNR de noyaux associés K_m , et où la contrainte ℓ_1 imposée aux coefficients σ_m favorise une solution parcimonieuse en f_m , et donc en K_m .

Dans [3], nous avons introduit l'apprentissage de noyaux composites, qui permet de prendre en compte une structure de groupe sur les noyaux. Dans la section suivante, nous étendons ce cadre à une structure plus profonde, mieux adaptée à la représentation de groupes de similarités pour les problèmes de traitement de signal ou de données génomiques dans lesquels une hiérarchie de similarités est définie par différents niveaux de granularité.

2 Pyramides de noyaux

Nous montrons ici comment étendre la formulation (1) à des problèmes dans lesquels les noyaux sont organisés dans une arborescence, comme celle illustrée sur la figure 1. Chaque feuille représente un noyau élémentaire, et à chaque nœud correspond le noyau formé récursivement par la combinaison des noyaux des nœuds fils. Notre objectif est d'obtenir un noyau global parcimonieux, ne faisant intervenir que peu de groupes et/ou sous-groupes de noyaux.

La variable h indice la hauteur dans l'arborescence : $h = 0$ pour la racine et $h = H$ pour les feuilles. Les nœuds de hauteur h sont également indicés par m à valeur dans $\{1, \dots, M_h\}$, et $G_{h+1,m}$ représente l'ensemble des fils (de hauteur $h + 1$) du nœud m de hauteur h . À chaque branche de l'arbre, on as-

socie le coefficient $\sigma_{h,m}$ de pondération du noyau m de hauteur h dans le calcul du noyau père. En notant $K_{h,m}$ le noyau équivalent au nœud m de hauteur h , on a $K_{H,m} = K_m$ et $K_{h-1,m} = \sum_{q \in G_{h,m}} \sigma_{h,q} K_{h,q}$. Ainsi, sur la figure 1, le noyau correspondant au nœud m_1 de hauteur 1 est défini par $\sum_{m_2 \in G_{2,m_1}} \sigma_{2,m_2} K_{m_2}$ et sa contribution dans la combinaison globale est pondérée par σ_{1,m_1} .

L'apprentissage de noyaux consiste à apprendre l'ensemble des coefficients $\sigma_{h,m}$, et donc le noyau effectif, conjointement avec les paramètres du SVM. Nous étendons ainsi la formulation du problème (1), associée à une arborescence à 1 niveau, et celle des noyaux composites [3] limitée à 2 niveaux. Le problème s'écrit :

$$(2) \begin{cases} \min_{\{f_m\}, b, \sigma} & \frac{1}{2} \sum_{m_1} \frac{1}{\sigma_{1,m_1}} \dots \sum_{m_H} \frac{1}{\sigma_{H,m_H}} \|f_{m_H}\|_{\mathcal{H}_{m_H}}^2 \\ & + C \sum_{i=1}^n [1 - y_i (\sum_{m=1}^{M_H} f_m(\mathbf{x}_i) + b)]_+ \\ \text{t.q.} & \sum_{m=1}^{M_h} \sigma_{h,m}^{p_h} \leq 1, \forall h \in \{1, \dots, H\} \\ & \sigma_{h,m} \geq 0, \forall h \in \{1, \dots, H\}, \forall m \in \{1, \dots, M_h\}, \end{cases}$$

où les indices m_h prennent leurs valeurs dans $G_{h,m_{h-1}}$, et les paramètres p_h introduits ici permettent de contrôler le degré de parcimonie associé au niveau h . Ce sont des paramètres libres, dont le choix sera discuté à la fin de cette section.

Le problème (2) peut être reformulé en terme de norme mixte sur les éléments f_m , pour expliciter la nature exacte de la pénalité appliquée à chaque niveau de la pyramide.

Proposition 1 *Le problème (2) équivaut à minimiser :*

$$\frac{1}{2} \left(\sum_{m_1} \left(\sum_{m_2} \dots \left(\sum_{m_H} \|f_{m_H}\|_{\mathcal{H}_{m_H}}^{\gamma_H} \right)^{\frac{\gamma_{H-1}}{\gamma_H}} \dots \right)^{\frac{\gamma_1}{\gamma_2}} \right)^{\frac{2}{\gamma_1}} + C \sum_{i=1}^n [1 - y_i (\sum_{m=1}^{M_H} f_m(\mathbf{x}_i) + b)]_+, \quad (3)$$

$$\text{où } \gamma_k = 2 \left(1 + \sum_{h=k}^H p_h \right)^{-1}.$$

Les propriétés des normes nous permettent d'établir les conditions de convexité et de parcimonie de (3).

Proposition 2 *La fonction objectif (3) est convexe si et seulement si $\gamma_h \geq 1$ pour tout $h \in \{1, \dots, H\}$.*

Proposition 3 *Le minimum de (3) est potentiellement parcimonieux au niveau h (i.e les coefficients $\sigma_{h,q}$ définissant le(s) noyau(x) $K_{h-1,m}$ sont potentiellement nuls) si et seulement si $\gamma_h \leq 1$ pour tout $h \in \{1, \dots, H\}$.*

Les propositions 2 et 3 indiquent que (3) ne peut à la fois être convexe et parcimonieux sur chaque niveau que si $\gamma_h = 1$ pour tout h . Dans ce cas, la structure de groupe n'est alors pas

prise en compte puisque la pénalité est insensible à l'hérédité représentée dans l'arborescence. Si nous voulons assurer la convexité du problème d'optimisation, grouper des noyaux permet donc à favoriser la sélection jointe des éléments d'un groupe ou sous-groupe par rapport à une sélection plus sévère mais plus dispersée sur l'arborescence, et il y a un compromis à faire entre structure de groupe et parcimonie, gouverné par les exposants p_h qui déterminent γ_h .

3 Algorithme

Pour résoudre le problème (2), nous utilisons un algorithme de type wrapper. Le schéma consiste à considérer deux problèmes imbriqués :

$$(4) \begin{cases} \min_{\sigma} J(\sigma) \\ \text{t.q.} \sum_{m=1}^{M_h} \sigma_{h,m}^{p_h} \leq 1, \forall h \in \{1, \dots, H\} \\ \sigma_{h,m} \geq 0, \forall h \in \{1, \dots, H\}, \forall m \in \{1, \dots, M_h\}, \end{cases}$$

où $J(\sigma)$ est défini comme le minimum de la fonction objectif du problème (2) par rapport à $\{f_m\}$ et b :

$$J(\sigma) = \min_{\{f_m\}, b} \frac{1}{2} \sum_{m_1} \frac{1}{\sigma_{1,m_1}} \dots \sum_{m_H} \frac{1}{\sigma_{H,m_H}} \|f_{m_H}\|_{\mathcal{H}_{m_H}}^2 + C \sum_{i=1}^n \left[1 - y_i \left(\sum_{m=1}^{M_H} f_m(x_i) + b \right) \right]_+ \quad (5)$$

Dans une boucle interne, la fonction objectif est minimisée par rapport aux paramètres $\{f_m\}$ et b , les paramètres σ étant fixés, définissant ainsi $J(\sigma)$. Dans une boucle externe, correspondant au problème (4), $J(\sigma)$ est optimisé par rapport à σ pour les paramètres $\{f_m\}, b$ calculés précédemment.

La solution du problème (5) peut être calculée par n'importe quel algorithme de résolution d'un problème SVM; celle du problème (4) peut être déduite des relations permettant d'exprimer σ_{h,m_h} en fonction de $\|f_{m_H}\|_{\mathcal{H}_{m_H}}$. Nous donnons ici les expressions obtenues pour une pyramide de hauteur 3 : ¹

$$\begin{aligned} \sigma_{1,m_1} &= c \times (s_{m_1})^{\frac{\gamma_1}{\gamma_2}} \\ \sigma_{2,m_2} &= c \times (s_{m_1})^{-\frac{p_1 \gamma_1}{2}} \times (s_{m_2})^{\frac{\gamma_2}{\gamma_3}} \\ \sigma_{3,m_3} &= c \times (s_{m_1})^{-\frac{p_1 \gamma_1}{2}} \times (s_{m_2})^{-\frac{p_2 \gamma_2}{2}} \times \|f_{m_3}\|_{\mathcal{H}_{m_3}}^{\gamma_3}, \end{aligned}$$

où $s_{m_2} = \sum_{m_3} \|f_{m_3}\|_{\mathcal{H}_{m_3}}^{\gamma_3}$, $s_{m_1} = \sum_{m_2} (s_{m_2})^{\frac{\gamma_2}{\gamma_3}}$, et $c = \left(\sum_{m_1} (s_{m_1})^{\frac{\gamma_1}{\gamma_2}} \right)^{-1}$, et les indices m_h prennent leurs valeurs dans $G_{h,m_{h-1}}$.

Afin d'optimiser les paramètres de la boucle externe, nous utilisons un algorithme de point fixe. L'ensemble de la procédure est résumée dans l'algorithme 1.

¹Ces relations sont obtenues par les conditions d'optimalité du premier ordre associées au problème (2). Les dérivations de ces expressions seront développées dans une version étendue.

Algorithme 1 : Pyramide de noyaux

```

initialiser  $\sigma$ 
résoudre le problème SVM  $\rightarrow J(\sigma)$ 

répéter
  répéter
    pour  $h = 1, \dots, H$ , et  $m = 1, \dots, M_h$  faire
      mettre à jour  $\sigma_{h,m}$ 
      // dont l'expression est déduite des
      // conditions d'optimalité de (2)
    jusqu'à la convergence
  résoudre le problème SVM  $\rightarrow J(\sigma)$ 
jusqu'à la convergence

```

4 Illustration

Nous illustrons la mise en œuvre de notre approche sur un problème de reconnaissance d'expression faciale sur des images. Ces images ont été normalisées par alignement des positions des yeux, du nez et de la bouche. L'ensemble d'apprentissage comporte 375 images, avec des visages exprimant la joie, la surprise, le dégoût, la tristesse, la peur ou la colère. Nous montrons ici la reconnaissance de la joie.

Chaque image de taille 128×96 pixels a été divisée en 48 imagettes de 16×16 pixels (cf. figure 2 (a)). Une transformation en ondelettes de Haar sur 3 niveaux de détails a ensuite été appliquée sur chaque imagette. Les coefficients de l'approximation et les 3 niveaux de détails forment ainsi 4 groupes de coefficients. Finalement, nous décomposons notre problème sur une pyramide de 3 niveaux : (1) le niveau 1 représente la position des 48 imagettes; (2) pour chaque imagette, le niveau 2 représente les groupes de coefficients associés aux 4 échelles ; (3) le niveau 3 différencie chaque coefficient de la décomposition en ondelettes pour chaque niveau de détails et chaque imagette.

Les images ont été réparties en deux ensembles : l'ensemble d'apprentissage est constitué de 250 images (35 positives et 215 négatives), tandis que l'ensemble de test contient 125 images (17 positives et 108 négatives). Nous avons testé 2 pénalités sur ce jeu de données : un SVM classique, correspondant à une pénalité ℓ_2 , et une version convexe de la formulation (3), correspondant à une pénalité $\ell(1; \frac{6}{5}; \frac{3}{2})$ ($p_1 = p_2 = p_3 = 3$), où 1 représente la pénalité appliquée au niveau 1, $6/5$ celle appliquée au niveau 2, et $3/2$ celle appliquée au niveau 3. Le paramètre de régularisation C a été sélectionné par validation croisée sur 5 blocs. Les deux méthodes obtiennent des performances similaires, à savoir 5 erreurs (4%) pour la pénalité ℓ_2 et 3 erreurs (2.4%) pour la pénalité $\ell(1; \frac{6}{5}; \frac{3}{2})$.

La figure 2 représente la pertinence des coefficients d'ondelettes pondérés par les paramètres des trois niveaux de la pyramide de noyaux (la position, les échelles et les poids associés aux coefficients de la décomposition). On constate ainsi que l'influence des coefficients localisés autour de la bouche et des yeux est plus importante que celles des autres zones.

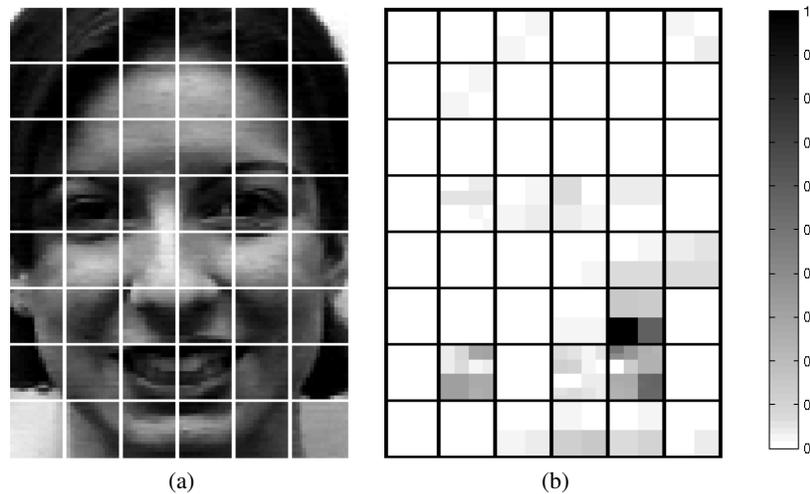


Figure 2: À gauche, une image sur laquelle sont délimitées les 48 imagettes. À droite, le masque obtenu à partir des coefficients d'ondelettes pondérés par les paramètres de la pyramide de noyaux. Les valeurs des coefficients ont été normalisées entre 0 et 1.

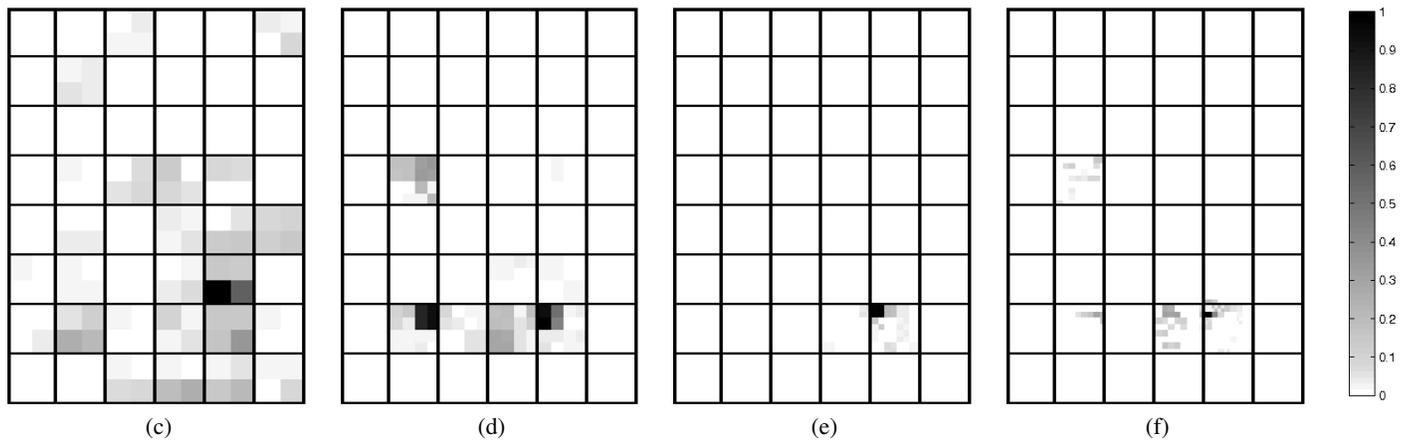


Figure 3: Influence des pondérations associées au second niveau de la pyramide de noyaux sur la définition du masque (cf. figure 2 (b)). Les figures (c), (d), (e), et (f) représentent respectivement la pertinence des coefficients de l'approximation, du premier, du second, et du troisième niveau de détail, en fonction des zones de l'image. Les valeurs des coefficients ont été normalisées entre 0 et 1.

La figure 3 montre comment influent les niveaux de détails de la décomposition en ondelettes : les niveaux les plus précis se concentrent sur les zones de la bouche et des yeux.

5 Conclusion

La méthode proposée dans cet article est à la frontière de l'apprentissage de noyaux et de la sélection de caractéristiques. D'une part, les pyramides de noyaux étendent l'approche du *Multiple Kernel Learning* en intégrant une connaissance relative à la façon dont les similarités entre exemples sont organisées. D'autre part, elle généralise les méthodes paramétriques de sélection de groupes de variables (telles que le *Group-lasso* [4]) à des espaces de fonctions noyaux (EHNR). Nous avons mis en évidence ces deux aspects sur un problème de reconnaissance d'expressions faciales.

References

- [1] G. R. G. Lanckriet, N. Cristianini, P. Bartlett, L. El Ghaoui, and M. I. Jordan. Learning the kernel matrix with semi-definite programming. *Journal of Machine Learning Research*, 5:27–72, 2004.
- [2] A. Rakotomamonjy, F. R. Bach, S. Canu, and Y. Grandvalet. SimpleMKL. *Journal of Machine Learning Research (JMLR)*, 9:2491–2521, 2008.
- [3] M. Szafranski, Y. Grandvalet, and A. Rakotomamonjy. Composite Kernel Learning. In *Proceedings of the 25th Annual International Conference on Machine Learning (ICML 2008)*, pages 1040–1047. Omnipress, 2008.
- [4] M. Yuan and Y. Lin. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society, Series B*, 68(1):49–67, 2006.