

Apprentissage à partir de noyaux multiples et indéfinis

Marie SZAFRANSKI^{1,2}, Matthieu KOWALSKI³, Liva RALAIVOLA⁴

¹ÉNSIIE, F-91025 Évry Cedex, France

²IBISC, EA 4526 – Université d'Évry Val d'Essonne, F-91037 Évry Cedex, France

³L2S, UMR 8506 CNRS – SUPELEC – Univ Paris-Sud, F-91192 Gif-sur-Yvette Cedex, France

⁴LIF, UMR 7279 CNRS – Aix-Marseille Université, F-13288, Marseille Cedex 9, France

marie.szafranski@ibisc.univ-evry.fr, matthieu.kowalski@lss.supelec.fr,
liva@.ralaivola@lif.univ-mrs.fr

Résumé – Nous présentons un cadre général permettant d'apprendre un classifieur à partir de plusieurs noyaux, qui ne sont potentiellement pas semi-définis positifs. Le problème d'optimisation associé fait intervenir une norme mixte qui permet de considérer différents types de dépendances entre les noyaux et / ou les données. Pour résoudre ce problème d'optimisation convexe, nous avons recours aux algorithmes proximaux qui permettent de traiter les non différentiabilités induites par les normes mixtes utilisées. Nous illustrons différentes utilisations de notre approche sur plusieurs jeux de données UCI ainsi que sur une application en classification d'images.

Abstract – We present a general framework to learn a classifier based on several kernels, which may induces kernels which are not semi-definite positive. The use of mixed-norms in the corresponding optimization problem allows to consider different kinds of dependancies between the kernels and / or the examples. We solve the convex problem by employing proximal minimization algorithms which allow to deal with the non differentiabilitys induced by the mixed-norms. We show different illustrations of our approach though numerical simulations on UCI datasets and on a image processing problem.

1 Introduction

Ces dernières années, la communauté de l'apprentissage statistique s'est intéressée au problème consistant à construire un classifieur à partir de différentes sources d'information. L'émergence de ces travaux est plus particulièrement liée aux problématiques issues d'applications bioinformatiques ou du traitement de données multimédias. Le *Multiple Kernel Learning* (MKL), introduit par Lanckriet *et al.* [4], est une des principales approches correspondant à cette tâche d'apprentissage.

Dans ce cadre, chaque source de données disponible est codée par un *noyau de Mercer*, c'est-à-dire une matrice symétrique et semi-définie positive. Les propriétés liées à ce type de noyaux permettent de projeter implicitement les exemples de \mathcal{X} dans un espace de caractéristiques \mathcal{H} , où \mathcal{H} est un Espace de Hilbert à Noyau Reproduisant (EHNR). C'est à travers ce cadre théorique que les méthodes dites « à noyaux » ont pris de l'essor, notamment avec les Séparateurs à Vaste Marge (SVM).

Cependant, il n'est pas toujours possible de coder l'information en terme de matrices semi-définies positives. En effet, les similarités utilisées dans certains domaines d'applications ne garantissent pas nécessairement les conditions énoncées plus haut. C'est par exemple le cas en bioinformatique, lorsque le score représentant l'alignement de deux séquences d'ADN est établi par l'algorithme de Smith-Waterman : la propriété de symétrie n'est pas respectée. Plus généralement, les mesures

de divergence ne peuvent être utilisées directement par les méthodes à noyaux classiques. Ong *et al.* [6] ont proposé un cadre théorique permettant d'utiliser des noyaux indéfinis induits par des espaces fonctionnels appelés Espaces de Krein à Noyaux Reproduisants (EKNR).

Pour traiter simultanément les aspects liés à la multiplicité des noyaux et au fait que ces derniers sont potentiellement indéfinis, nos travaux adoptent un autre point de vue, assez courant dans la communauté du traitement du signal. Notre approche consiste à formaliser l'apprentissage de plusieurs noyaux comme l'apprentissage d'une représentation des données issue de l'union de plusieurs dictionnaires.

Dans la première partie, nous définirons le cadre proposé pour apprendre un classifieur à partir de plusieurs noyaux indéfinis régularisés par des normes mixtes ; nous spécifierons le problème d'optimisation ainsi que le comportement attendu en fonction des normes mixtes utilisées. En effet, en plus d'utiliser plusieurs noyaux pour tirer partie de différentes sources d'information, nous souhaitons que l'algorithme d'apprentissage soit aussi capable de faire ressortir les informations pertinentes, c'est-à-dire les données ou les sources de données qui interviennent de façon significative dans le processus décisionnel. Nous présenterons dans la seconde partie un algorithme de résolution du problème fondé sur les méthodes proximales. Enfin, nous illustrerons notre approche sur des jeux de données UCI ainsi que sur une application en traitement d'images.

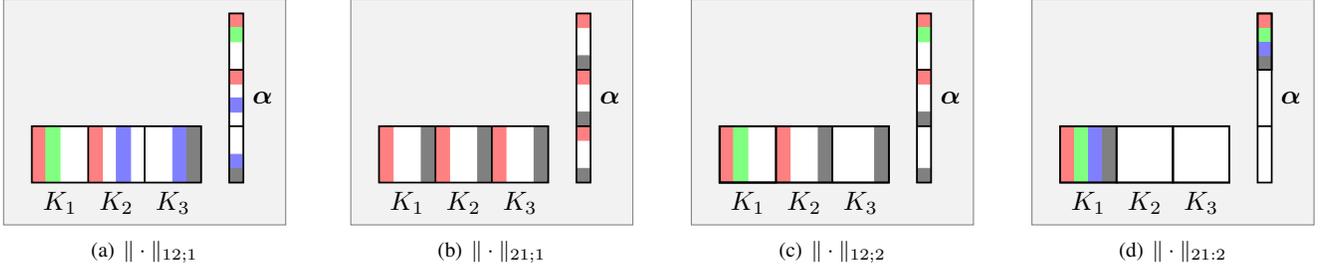


FIGURE 1 – Nature de la parcimonie attendue sur le vecteur α en fonction des différentes normes $\|\cdot\|_{pq;r}$. Les carrés blancs de α correspondent à des coefficients dont la valeur est nulle. Les exemples sont identifiés par une couleur spécifique, reprise dans la colonne associée de K qui est utilisée dans la fonction de décision du classifieur. Lorsque $r = 1$ ((a) et (b)), la structure est définie par rapport aux exemples. (a) : pour chaque exemple, un sous-ensemble de noyaux est sélectionné. (b) : un sous-ensemble commun d'exemples est sélectionné, et ce pour tous les noyaux. Lorsque $r = 2$ ((c) et (d)), la structure est définie par rapport aux noyaux. (c) : pour chaque noyau, un sous-ensemble d'exemples est sélectionné. (d) : un sous-ensemble de noyaux, contenant tous les exemples, est sélectionné.

2 MIKL et les normes mixtes

2.1 Cadre

Nous nous intéressons à un problème de classification binaire, pour lequel nous disposons d'un ensemble d'apprentissage $S = \{(\mathbf{x}_i, y_i)\}_{i=1}^n$ de n paires d'exemples étiquetés, où $\forall i, (\mathbf{x}_i, y_i) \in \mathcal{Z} = \mathcal{X} \times \mathcal{Y}$, avec \mathcal{X} l'espace dit *des entrées* et $\mathcal{Y} = \{-1, +1\}$ l'ensemble des étiquettes.

Nous considérons des *classifieurs à noyaux multiples* dans lesquels les noyaux utilisés ne sont pas nécessairement semi-définis positifs. Ici, un noyau k est simplement un élément de $\mathbb{R}^{\mathcal{X} \times \mathcal{X}}$, sans autre forme de condition. Nous disposons donc d'un ensemble $\mathcal{K} = \{k_1, \dots, k_\tau\}$ de τ noyaux, et les classifieurs à noyaux multiples sont les versions signées de fonctions $f \in \mathcal{F}_S$ définies à partir de l'ensemble d'apprentissage S comme

$$\mathcal{F}_S = \left\{ \mathbf{x} \mapsto \sum_{i,t=1}^{n,\tau} \alpha_{it} k_t(\mathbf{x}_i, \mathbf{x}) : \alpha \in \mathbb{R}^{n\tau}, k_t \in \mathcal{K} \right\}.$$

Ainsi, la sortie prédite pour \mathbf{x} par $f \in \mathcal{F}_S$ est $\text{sign}(f(\mathbf{x}))$.

Dans ce cadre général¹, apprendre un classifieur à partir de S revient à rechercher le vecteur α permettant de minimiser l'erreur de généralisation. Dans ce but, nous proposons de résoudre le problème d'optimisation régularisé suivant :

$$\min_{\alpha \in \mathbb{R}^{n\tau}} \sum_{i=1}^n \max(1 - y_i \mathbf{k}_i \alpha, 0)^2 + \frac{\lambda}{q} \|\alpha\|_{pq;r}^q \quad (1)$$

avec $\lambda \in \mathbb{R}^{+*}$, $p, q \in \{1, 2\}$ et $r \in \{1, 2\}$. Ici, α est le vecteur colonne associé à l'ensemble doublement indexé des coefficients $(\alpha_{it})_{\substack{1 \leq i \leq n \\ 1 \leq t \leq \tau}}$ et \mathbf{k}_i est la i^{e} ligne de

$$K = [K_1 \dots K_\tau] \in \mathbb{R}^{n \times n\tau}, \quad (2)$$

1. Nous pouvons montrer qu'une telle famille englobe les classifieurs à noyaux multiples définis dans le cadre du MKL.

où K_t est la matrice $K_t = (k_t(\mathbf{x}_i, \mathbf{x}_j))_{1 \leq i, j \leq n}$ associée au noyau k_t . Quant-aux normes mixtes, la norme $\|\cdot\|_{pq;1}$ est définie par :

$$\|\alpha\|_{pq;1} = \left[\sum_{i=1}^n \left[\sum_{t=1}^{\tau} |\alpha_{it}|^p \right]^{q/p} \right]^{1/q}, \quad (3)$$

tandis que la norme $\|\cdot\|_{pq;2}$ est définie ainsi :

$$\|\alpha\|_{pq;2} = \left[\sum_{t=1}^{\tau} \left[\sum_{i=1}^n |\alpha_{it}|^p \right]^{q/p} \right]^{1/q}. \quad (4)$$

Notons que l'ordre dans lequel s'effectue les sommes dans (3) puis (4) change. Le choix des valeurs de p et q , lorsqu'on les fixe à 1, induit différentes structures de parcimonie sur le vecteur α dans (1). La figure 1 décrit les différentes structures de parcimonie attendues sur les coefficients α en fonction des paramètres p et q , mais aussi du paramètre r .

Le premier terme de la fonction objectif du problème (1) est le *coût charnière* au carré, qui utilise la norme 2 des SVM. Le second terme correspond à la partie régularisée, qui permet de contrôler la complexité de la famille de classifieurs considérée. L'ensemble des valeurs de p et q que nous considérons, à savoir $p, q \in \{1, 2\}$, nous permet d'obtenir une fonction de régularisation (la norme mixte) convexe en α . Puisque $\lambda > 0$, la fonction objectif dans sa globalité est également convexe en α . Néanmoins, notons que cette fonction objectif devient non-différentiable dès que p ou q valent 1, ce qui correspond précisément à la situation qui nous intéresse puisque ce choix induit une structure parcimonieuse sur α qui nous pourrait nous permettre d'identifier les exemples et les sources pertinentes d'un problème. Cette non-différentiabilité sera élégamment traitée par un algorithme proximal que nous décrivons dans la partie suivante.

3 Algorithme

3.1 Algorithmes proximaux

Les problèmes traités par les algorithmes proximaux prennent la forme générale suivante

$$\min_{\alpha} f_1(\alpha) + f_2(\alpha), \quad (5)$$

dans laquelle f_1 et f_2 sont des fonctions convexes et semi-continues inférieurement. La résolution de ce type de problème se fonde sur les opérateurs de proximités introduits dans [5]. Combettes et Pesquet [3] ont montré que l'algorithme *Forward-Backward* converge vers un minimum global du problème (5) lorsque f_2 est convexe et semi-continue inférieurement et que f_1 est de plus différentiable et de gradient β -Lipschitz.

Pour résoudre le problème (1), nous utiliserons donc le cadre des algorithmes proximaux dans lequel le rôle de f_1 sera joué par *coût charnière* au carré tandis que celui de f_2 sera assuré par la régularisation induite sur les coefficients α .

3.2 Opérateur de proximité

Afin de pouvoir évoluer dans ce cadre, il nous faut définir les opérateurs de proximité.

Définition 1 (Opérateur de proximité). Soit $\varphi : \mathbb{R}^P \rightarrow \mathbb{R}$ une fonction convexe semi-continue inférieurement. L'opérateur de proximité $\text{prox}_{\varphi} : \mathbb{R}^P \rightarrow \mathbb{R}^P$ associé à φ est défini par

$$\text{prox}_{\varphi}(\mathbf{u}) = \underset{\alpha \in \mathbb{R}^P}{\text{argmin}} \frac{1}{2} \|\mathbf{u} - \alpha\|_2^2 + \varphi(\alpha).$$

Nous nous intéressons ici aux opérateurs de proximité associés aux normes mixtes utilisées dans le problème (1). Dans ce cas, $f_2(\alpha) = q^{-1} \lambda \|\alpha\|_{pq}^q$ est une fonction convexe et semi-continue inférieurement pour $p, q \geq 1$; néanmoins elle n'est pas différentiable en 0. Nous concentrons ici notre étude des opérateurs de proximité sur les normes mixtes suivantes : $\|\cdot\|_1$, $\|\cdot\|_2$, $\|\cdot\|_{12}$, $\|\cdot\|_{21}$, c'est-à-dire lorsque $p, q \in \{1, 2\}$.

Notations. Avant de poursuivre, nous avons besoin de définir quelques notations. Soient $\lambda \in \mathbb{R}^{+*}$, $L \in \mathbb{N}$ et $M \in \mathbb{N}$. Soit $\mathbf{u} \in \mathbb{R}^{N \times M}$, un ensemble de coefficients doublement indexés par (ℓ, m) , avec $1 \leq \ell \leq L$ et $1 \leq m \leq M$. Nous notons $\mathbf{u}_{\ell \bullet}$ le vecteur $[u_{\ell 1}, \dots, u_{\ell M}]^T$. Enfin, $|u_{\ell m}|_+ = \max(0, u_{\ell m})$ et $\|\mathbf{u}_{\ell \bullet}\|_+^2 = \sum_m |u_{\ell m}|_+^2$.

Proposition 1 (Opérateurs de proximité pour les normes mixtes). L'opérateur de proximité pour $\lambda \|\cdot\|_{pq}$, avec $p, q \in \{1, 2\}$, défini par

$$\hat{\alpha} = \text{prox}_{\lambda \|\cdot\|_{pq}^q}(\mathbf{u}) = \underset{\alpha \in \mathbb{R}^P}{\text{argmin}} \frac{1}{2} \|\mathbf{u} - \alpha\|_2^2 + \frac{\lambda}{q} \|\alpha\|_{pq}^q,$$

se formule composante par composante :

– lorsque $p = 1$ et $q = 1$,

$$\hat{\alpha}_{\ell, m} = \text{sign}(u_{\ell, m}) |u_{\ell, m}| - \lambda|_+,$$

l'opérateur de seuillage doux associé au problème bien connu du *basis-pursuit denoising* [2].

– lorsque $p = 2$ et $q = 2$,

$$\hat{\alpha}_{\ell, m} = \frac{1}{1 + \lambda} u_{\ell, m};$$

– lorsque $p = 2$ et $q = 1$,

$$\hat{\alpha}_{\ell, m} = u_{\ell, m} \left| 1 - \frac{\lambda}{\|\mathbf{u}_{\ell \bullet}\|_2} \right|_+;$$

– lorsque $p = 1$ et $q = 2$,

$$\hat{\alpha}_{\ell, m} = \text{sign}(u_{\ell, m}) \left| u_{\ell, m} - \frac{\lambda \sum_{m_{\ell}=1}^{M_{\ell}} \tilde{u}_{\ell, m_{\ell}}}{(1 + \lambda M_{\ell}) \|\mathbf{u}_{\ell \bullet}\|_2} \right|_+,$$

où $\tilde{u}_{\ell, m_{\ell}}$ représente les coefficients $|u_{\ell, m_{\ell}}|$ ordonnés de façon décroissante pour un ℓ fixé, et où M_{ℓ} représente le nombre pour lequel

$$\tilde{u}_{\ell, M_{\ell}+1} \leq \lambda \sum_{m_{\ell}=1}^{M_{\ell}+1} (\tilde{u}_{\ell, m_{\ell}} - \tilde{u}_{\ell, M_{\ell}+1}),$$

et

$$\tilde{u}_{\ell, M_{\ell}} > \lambda \sum_{m_{\ell}=1}^{M_{\ell}} (\tilde{u}_{\ell, m_{\ell}} - \tilde{u}_{\ell, M_{\ell}}).$$

3.3 Optimisation proximale du problème (1)

Le *coût charnière* au carré peut se récrire sous forme matricielle : $\|\mathbf{1} - K_y \alpha\|_+^2$, où $K_y = \text{diag}([y_1, \dots, y_n]) K$. Puisque

- i) $f_1(\alpha) = \|\mathbf{1} - K_y \alpha\|_+^2$ est une fonction différentiable et de gradient β -Lipschitz,
- ii) $f_2(\alpha) = q^{-1} \lambda \|\alpha\|_{pq}^q$, pour $p, q \in \{1, 2\}$, est une fonction convexe, semi-continue inférieurement et non différentiable en 0,
- iii) nous avons établi comment calculer les opérateurs de proximités pour des normes $\|\cdot\|_{pq; r}$, lorsque $p, q \in \{1, 2\}$,

nous pouvons donc utiliser l'algorithme *Forward-Backward* pour résoudre le problème (1). Pour cela, il faut calculer le gradient de $\|\mathbf{1} - K_y \alpha\|_+^2$ qui vaut $-K_y^T \|\mathbf{1} - K_y \alpha\|_+$ et qui est β -Lipschitz pour $\beta = \|K_y^T K_y\|$. La procédure adaptée au problème (1) est donnée dans l'algorithme 1. En pratique, nous utilisons une version accélérée de type FISTA [1].

Algorithme 1: Forward-Backward pour *coût charnière* au carré

Entrées : K_y, λ, γ , avec $\gamma < 2/\|K_y^T K_y\|$

Initialisations : $\alpha^{(0)} \in \mathbb{R}^{n \times \tau}$, par exemple $\mathbf{0}$

répéter

$$\left| \alpha^{(s+1)} = \text{prox}_{\gamma \lambda \|\cdot\|_{pq}^q} \left(\alpha^{(s)} + \gamma K_y^T [\mathbf{1} - K_y \alpha]_+ \right) \right.$$

jusqu'à ce que l'algorithme converge;

4 Expériences

4.1 Simulations

Cette simulation numérique compare les résultats obtenus avec différentes normes sur 3 jeux de données issus de la base UCI. Pour chaque jeu de données, nous avons utilisé un noyau général composé de 9 noyaux : des noyaux polynomiaux avec 3 degrés différents et des noyaux gaussiens avec 6 largeurs de bandes différentes. Nous utilisons ici les méthodes de combinaison de noyaux comme alternative à la validation croisée. Par ce biais, nous évitons une procédure supplémentaire pour sélectionner le(s) paramètre(s) approprié(s) des noyaux utilisés. Le paramètre de régularisation λ est sélectionné par une procédure de validation croisée à 5 blocs.

Le tableau 1 reporte les résultats moyens obtenus sur 10 ensembles d'apprentissage différents. Nous pouvons constater qu'une structure adaptée peut être bénéfique. En particulier, sur *Liver*, une structure permettant de sélectionner les exemples les plus influents (MIKL $\|\cdot\|_{12;2}$) permet d'augmenter significativement les performances. Sur *Sonar*, les structures permettant d'éliminer les noyaux non pertinents (MKL et MIKL $\|\cdot\|_{21;2}$, ou MIKL $\|\cdot\|_{12;1}$) semblent plus efficaces. En pratique, le choix de la norme utilisée dépendra du type de structure inhérente à l'application ciblée et de l'objectif attendu.

	Ionosphere	Liver	Sonar
MKL	93.7 \pm 2.0	64.3 \pm 4.2	84.1 \pm 4.4
MIKL $\ \cdot\ _2$	92.4 \pm 1.8	66.4 \pm 4.0	82.5 \pm 6.2
MIKL $\ \cdot\ _1$	86.2 \pm 3.5	65.2 \pm 3.5	83.2 \pm 5.0
MIKL $\ \cdot\ _{12;1}$	91.0 \pm 1.7	65.6 \pm 3.6	84.3 \pm 5.1
MIKL $\ \cdot\ _{21;1}$	88.4 \pm 3.1	65.9 \pm 3.9	81.0 \pm 6.2
MIKL $\ \cdot\ _{12;2}$	86.2 \pm 3.5	69.5 \pm 2.8	81.4 \pm 4.8
MIKL $\ \cdot\ _{21;2}$	87.9 \pm 3.5	65.3 \pm 2.9	84.6 \pm 4.8

TABLE 1 – Résultats moyens (en % de classification correcte) pour les simulations.

4.2 Classification d'images

Nous nous intéressons ici à un problème de classification d'images réparties dans 10 classes. Nous considérons un problème de classification binaire (classe 1 contre classes 2 à 10). Les similarités entre images sont représentées par une divergence de *Kullback-Leibler* (non symétrique, cf. figure 2) sur les distributions marginales de coefficients d'ondelettes pour 3 niveaux de résolution différents, selon le procédé proposé dans [7]. Le paramètre de régularisation λ est sélectionné par une procédure de validation croisée à 5 blocs.

Le tableau 2 présente les résultats moyens obtenus sur 10 ensembles d'apprentissage différents. Les performances obtenues par les différentes méthodes sont assez similaires, avec un léger avantage pour une méthode non structurée (MIKL $\|\cdot\|_2$). Notons qu'en terme de sélection de noyaux (MIKL $\|\cdot\|_{21;2}$), un seul essai les conserve tous. Sur les 9 essais restants, K_1 est éliminé 3 fois tandis que K_2 est éliminé 5 fois.

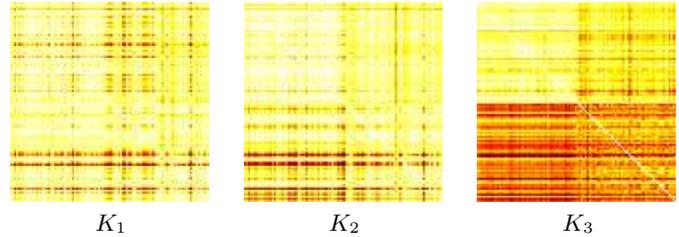


FIGURE 2 – Noyaux non symétriques issus du calcul des similarités entre les images sur 3 niveaux de résolution différents.

Sans structure	MIKL $\ \cdot\ _2$	4.89 \pm 3.5
	MIKL $\ \cdot\ _1$	5.89 \pm 3.7
Structure sur les exemples	MIKL $\ \cdot\ _{12;1}$	5.56 \pm 3.6
	MIKL $\ \cdot\ _{21;1}$	5.44 \pm 3.9
Structure sur les noyaux	MIKL $\ \cdot\ _{12;2}$	6.33 \pm 3.9
	MIKL $\ \cdot\ _{21;2}$	5.56 \pm 3.2

TABLE 2 – Résultats moyens (en % de taux d'erreur) pour le problème de classification d'image.

5 Conclusion

Nous avons présenté un cadre permettant d'apprendre un classifieur à partir de plusieurs noyaux potentiellement indéfinis. Une structure sur les exemples et les noyaux, intégrée par le biais d'une norme mixte sur le terme de régularisation, peut permettre de faire ressortir les informations pertinentes.

Nous pouvons également montrer que ces classifieurs présentent des garanties sur leur capacité de généralisation, c'est-à-dire leur capacité à prédire des réponses correctes à partir d'exemples inédits. Cet aspect n'a cependant pu être développé faute de place.

Références

- [1] BECK, A. et TBOULLE, M. (2009). A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM Journal on Imaging Sciences*, 2(1):183–202.
- [2] CHEN, S. S., DONOHO, D. L., MICHAEL et SAUNDERS, A. (1998). Atomic decomposition by basis pursuit. *SIAM Journal on Scientific Computing*, 20:33–61.
- [3] COMBETTES, P. L. et PESQUET, J.-C. (2007). A Douglas–Rachford splitting approach to nonsmooth convex variational signal recovery. *IEEE Journal of Selected Topics in Signal Processing*, 1(4):564–574.
- [4] LANCKRIET, G., CRISTIANINI, N., BARTLETT, P., EL GHAOU, L. et JORDAN, M. (2004). Learning the Kernel Matrix with Semidefinite Programming. *Journal of Machine Learning Research*, 5:27–72.
- [5] MOREAU, J.-J. (1965). Proximité et dualité dans un espace hilbertien. *Bulletin de la Société Mathématique de France*, 93:273–299.
- [6] ONG, C. S., MARY, X., CANU, S. et SMOLA, A. (2004). Learning with non-positive kernels. In *Proceedings of the 21st International Conference on Machine Learning*, Banf, Alberta, Canada.
- [7] PIRO, P., ANTHOINE, S., DEBREUVE, E. et BARLAUD, M. (2008). Image retrieval via Kullback-Leibler divergence of patches of multiscale coefficients in the kNN framework. In *Content-Based Multimedia Indexing*, London, UK.