Chapter 2

A Multi-Features Fusion of Multi-Temporal Hyperspectral Images Using a Cooperative GDD/SVM Method

Selim Hemissi and Imed Riadh Farah

Additional information is available at the end of the chapter

http://dx.doi.org/10.5772/56949

1. Introduction

Considering the emergence of hyperspectral sensors, feature fusion has been more and more important for images classification, indexing and retrieval. In this chapter, a cooperative fusion method GDD/SVM (Generalized Dirichlet Distribution/Support Vector Machines), which involves heterogeneous features, is proposed for multi-temporal hyperspectral images classification. It differentiates, from most of the previous approaches, by incorporating the potentials of generative models into a discriminative classifier. Therefore, the multi-features, including the 3D spectral features and textural features, can be integrated with an efficient way into a unified robust framework. The experimental results on a series of Hyperion images show that the precision is 92.64% and the recall is 91.87%. The experiments on AVIRIS dataset also confirm the improved performance and show that this cooperative fusion approach has consistence over different testing datasets.

2. Problem statement

The semantic categorization of remote-sensing images requires analysis of many features of the images such as texture, spectral profiles, etc. Current feature fusion approaches commonly concatenate different features. It gives, generally good results and several approaches have been proposed using this schema. However, most of them have various conditional constraints, such as noise and imperfection, which might retain the use of such systems under degraded performance. However, how to fuse heterogeneous features in a flexible way is still an open research question.

Similarly, in the area of Supervised Machine Learning (SML), diversity with respect to the errors committed by component classifiers has received much attention. Generative and discriminative approaches are two distinct schools of probabilistic machine learning. It has shown that discriminative approaches such as SVM [1] outperform model based approaches due to their flexibility in decision boundaries estimation. Conversely, since that discriminative methods are concerned with boundaries, all the classes need to be estimated conjointly [2]. Complementary, one of the interesting characteristics, that generative models have over discriminative ones, is that they are learnt independently for each class. Moreover, following their modeling power, generative models are able to deal with missing data. An ideal fusion method should combine these two approaches in order to improve the classification accuracy.

3. Theoretical background

3.1. Generalized dirichlet distribution

Priors based on Dirichlet location-scale mixture of normals are widely used to model densities as mixtures of normal kernels. A random density f arising from such a prior can be expressed as

$$f(y) = (\phi * P)(y) = \int \frac{1}{\sigma}\phi\left(\frac{y-\theta}{\sigma}\right) dP(\theta,\sigma), \tag{1}$$

where $\phi(\cdot)$ is the standard normal density and the mixing distribution P follows a Dirichlet process.

[3] initiated a theoretical study of these priors for the problem of density estimation. They showed that if a density f_0 satisfies certain conditions, then a Dirichlet location mixture of normals achieves posterior consistency at f_0. Their conditions can be best summarized as f_0 having a moment generating function on an open interval containing $[-1,1]$. Ghosal and van der Vaart (2001) extended these results to rate calculations for the more general Dirichlet location-scale mixture prior. However, they restricted the scale parameter σ to a compact interval $[\underline{\sigma},\overline{\sigma}] \subset (0,\infty)$.

3.1.1. Preliminaries

To make this chapter relatively self-contained, we recall the definitions of posterior consistency in the context of density estimation and regression. These definitions formalize the concept that in order to achieve consistency, the posterior should concentrate on arbitrarily small neighborhoods of the true model when more observations are made available.

Posterior consistency for density estimation: Suppose $X_1, X_2, \cdots$ are independent and identically distributed according to an unknown density f_0. We take the parameter space as $\mathcal{F}$ - a set of probability densities on the space of the observations and consider a prior distribution Π on $\mathcal{F}$. Then the posterior distribution $\Pi(\cdot|X_1,\cdots,X_n)$ given a sample $X_1,\cdots,X_n$ is obtained as,

http://dx.doi.org/10.5772/56949

$$\Pi(A|X_1,\cdots,X_n)=\frac{\int_A\prod_{i=1}^n f(X_i)d\Pi(f)}{\int_{\mathcal{F}}\prod_{i=1}^n f(X_i)d\Pi(f)}.$$

We say that the posterior achieves weak (or strong) posterior consistency at f_0 if for any weak (or L_1) neighborhood U of f_0, $\Pi(U|X_1,X_2,\cdots,X_n)\to 1$ almost surely as $n\to\infty$.

Posterior consistency for regression: Suppose one observes $Y_1,Y_2,\cdots$ from the model $Y_i=\alpha_0+\beta_0 x_i+\epsilon_i$, where x_i's are known non-random covariate values and ϵ_i's are independent and identically distributed with an unknown symmetric density f_0. The regression coefficients α_0,β_0 are also unknown. Here, it is appropriate to consider the parameter space as $\Theta=\mathcal{F}^*\times\mathbb{R}\times\mathbb{R}$, where $\mathcal{F}^*$ is a set of symmetric probability densities on $\mathbb{R}$ with a prior Π on Θ. The posterior distribution $\Pi(\cdot|Y_1,\cdots,Y_n)$ is then computed as,

$$\Pi(A|Y_1,\cdots,Y_n)=\frac{\int_A\prod_{i=1}^n f(Y_i-\alpha-\beta x_i)d\Pi(f,\alpha,\beta)}{\int_\times\prod_{i=1}^n f(Y_i-\alpha-\beta x_i)d\Pi(f,\alpha,\beta)}.$$

We say that the posterior achieves weak consistency at (f_0,α_0,β_0) if for any weak neighborhood U of f_0 and any $\delta>0$,

$$\Pi((f,\alpha,\beta):f\in U,|\alpha-\alpha_0|<\delta,|\beta-\beta_0|<\delta|Y_1,Y_2,\cdots,Y_n)\to 1$$

almost surely as $n\to\infty$.

3.1.2. Density estimation: weak consistency

We start with weak posterior consistency for the problem of density estimation. Our main tool is the following theorem due to Schwartz (1965).

A prior Π achieves weak posterior consistency at a density f_0, if

$$\forall\epsilon>0,\ \Pi\left(f\in\mathcal{F}:\int f_0(x)\log\frac{f_0(x)}{f(x)}dx<\epsilon\right)>0 \tag{2}$$

We would use the notation $f_0\in KL(\Pi)$ to indicate that a density f_0 satisfies (2).

General Mixture Priors First consider the case when the mixing distribution P in (1) follows some general distribution $\tilde{\Pi}$, not necessarily a Dirichlet process. It is *reasonable* to assume that the weak support of $\tilde{\Pi}$ contains all probability measures on $\mathbb{R}\times\mathbb{R}^+$ that are compactly supported. The next lemma reveals the implication of this property.

Consider an $f_0\in\mathcal{F}$ such that $\int x^2 f_0(x)dx<\infty$. Suppose $\tilde{f}=\phi*\tilde{P}$ is such that $\tilde{P}((-a,a)\times(\underline{\sigma},\overline{\sigma}))=1$ for some $a>0,0<\underline{\sigma}<\overline{\sigma}$. Then for any $\epsilon>0$, there exists a weak neighborhood W of $\tilde{P}$ such that for any $f=\phi*P$ with $P\in W$,

$$\int f_0(x)\log\frac{\tilde{f}(x)}{f(x)}dx<\epsilon \tag{3}$$

The proof of this lemma is similar to the proof of Theorem 3 of Ghosal *et al.* (1999) and we present it in the appendix. Here we state and prove the main result.

Let $f_0(x)$ be a continuous density on $\mathbb{R}$ satisfying:

1. f_0 is nowhere zero and bounded above by $M < \infty$.
2. $\left|\int_{\mathbb{R}} f_0(x) \log f_0(x) dx\right| < \infty$.
3. $\int_{\mathbb{R}} f_0(x) \log \frac{f_0(x)}{\psi_1(x)} dx < \infty$ where $\psi_1(x) = \inf_{t\in[x-1,x+1]} f_0(t)$.
4. $\exists \eta > 0$ such that $\int_{\mathbb{R}} |x|^{2(1+\eta)} f_0(x) dx < \infty$.

Then, $f_0 \in KL(\Pi)$.

Assumption *4* provides the important moment condition on f_0. Assumption *2* is satisfied by most of the common densities and assumption *3* can be viewed as a regularity conditions. The interval $[x-1, x+1]$ that appears in assumption *3* can be replaced by $[x-a, x+a]$ for any $a > 0$.

Proof. **of Theorem 3.1.2** Note that,

$$\int f_0(x) \log \frac{f_0(x)}{f(x)} dx = \int f_0(x) \log \frac{f_0(x)}{\tilde{f}(x)} dx + \int f_0(x) \log \frac{\tilde{f}(x)}{f(x)} dx. \tag{4}$$

Therefore, the result would follow if for any $\epsilon > 0$, we can find an $\tilde{f}$ which makes $\int f_0 \log \frac{f_0}{\tilde{f}} dx < \epsilon/2$ and also satisfies the condition of Lemma 3.1.2. Next we show how to construct such an $\tilde{f}$.

Consider the densities $f_n = \phi * P_n$, $n \geq 1$, with P_n's constructed as,

$$dP_n(\theta, \sigma) = t_n I_{(\theta\in[-n,n])} f_0(\theta) \delta_{\sigma_n}(\sigma) \tag{5}$$

where $\sigma_n = n^{-\eta}$, $t_n = (\int_{-n}^{n} f_0(y) dy)^{-1}$, I_A is the indicator function of a set A and δ_x is the point mass at a point x. Note that f_n can be simply written as,

$$f_n(x) = t_n \int_{-n}^{n} \frac{1}{\sigma_n} \phi\left(\frac{x-\theta}{\sigma_n}\right) f_0(\theta) d\theta. \tag{6}$$

Find a positive constant ξ such that $\int_{-\xi}^{\xi} \phi(t) dt > 1 - \epsilon$. Now fix an $x \in \mathbb{R}$. For sufficiently large n such that $[x - \xi\sigma_n, x + \xi\sigma_n] \subset [-n, n]$, one obtains,

$$\inf_{y\in(x-\xi\sigma_n, x+\xi\sigma_n)} f_0(y)(1-\epsilon) < \frac{f_n(x)}{t_n} < \sup_{y\in(x-\xi\sigma_n, x+\xi\sigma_n)} f_0(y) + M\epsilon \tag{7}$$

http://dx.doi.org/10.5772/56949

Since $t_n \to 1$ and $\sigma_n \to 0$, (7) would imply that $f_n(x) \to f_0(x)$ as $n \to \infty$ by continuity of f_0. Therefore one can conclude,

$$\log \frac{f_0(x)}{f_n(x)} \to 0 \text{ for all } x \in \mathbb{R} \tag{8}$$

Since t_n is a decreasing sequence and $f_0(\theta) < M$ for all $\theta \in \mathbb{R}$, one can readily see that for all $n \geq 1$ and all $x \in \mathbb{R}$,

$$f_n(x) = t_n \int_{-n}^{n} \frac{1}{\sigma_n} \phi\left(\frac{x-\theta}{\sigma_n}\right) f_0(\theta) d\theta \leq M t_n \leq M t_1. \tag{9}$$

Now, fix an $x \in \mathbb{R}$. Since, $|x-\theta| \leq |x| + n$ for all $\theta \in [-n, n]$ and $t_n \geq 1$, it follows that for all $n \leq |x|$,

$$f_n(x) \geq \frac{1}{\sigma_n} \phi\left(\frac{|x|+n}{\sigma_n}\right) = n^{\eta} \phi(n^{\eta}(|x|+n)) \geq |x|^{\eta} \phi(2|x|^{1+\eta}). \tag{10}$$

The last inequality follows from the fact that $\tau^{\eta}\phi(\tau^{\eta}(|x|+\tau))$ is decreasing in τ for $\tau \geq 1$.

Let $\psi_n(x) = \inf_{t \in [x-\sigma_n, x+\sigma_n]} f_0(t)$. It may be noted that the function $\psi_1(x)$ of assumption *3* is consistent with this definition. Let $A_n = [-n, n] \cap [x-\sigma_n, x+\sigma_n]$ and $c = \int_0^1 \phi(t)dt < 1$. Observe that for all $n > |x|$,

$$f_n(x) \geq t_n \int_{A_n} \frac{1}{\sigma_n} \phi\left(\frac{x-\theta}{\sigma_n}\right) f_0(\theta) d\theta \geq t_n \psi_n(x) \int_{A_n} \frac{1}{\sigma_n} \phi\left(\frac{x-\theta}{\sigma_n}\right) d\theta \tag{11}$$

Since $t_n \geq 1$, $\psi_n(x) \geq \psi_1(x)$ and $\int_{A_n} \frac{1}{\sigma_n}\phi(\frac{x-\theta}{\sigma_n})d\theta \geq \int_0^1 \phi(t)dt = c$ for all $n \geq 1$ and all $x \in \mathbb{R}$ it follows from (11) that $f_n(x) \geq c\psi_1(x)$ for all $n > |x|$.. Therefore,

$$f_n(x) \geq \begin{cases} c\psi_1(x) & |x| < 1 \\ \min(|x|^{\eta}\phi(2|x|^{1+\eta}), c\psi_1(x)) & |x| \geq 1 \end{cases} \tag{12}$$

A little algebraic manipulation with (9) and (12) obtains, $\forall n \geq 1$,

$$\left|\log \frac{f_0(x)}{f_n(x)}\right| \leq \log \frac{M t_1}{f_0(x)} + \log \frac{f_0(x)}{c\psi_1(x)} + I_{\{|x|>1\}} \log \frac{f_0(x)}{|x|^{\eta}\phi(2|x|^{1+\eta})} \tag{13}$$

From the assumptions of Theorem 3.2, it can be easily verified that the function on the right hand side of the above display is f_0 integrable. Therefore an application of DCT on (8) implies that,

$$\lim_{n \to \infty} \int f_0(x) \log \frac{f_0(x)}{f_n(x)} dx = 0. \tag{14}$$

Therefore we can simply choose $\tilde{f} = f_{n_0}$ for some large enough n_0.

□

3.2. Dirichlet mixture of normals

Next we consider $\tilde{\Pi} = Dir(\alpha G_0)$, a Dirichlet process with parameter αG_0. Here α is a positive constant and G_0 is a probability measure on $\mathbb{R} \times \mathbb{R}^+$.

Suppose $f_0 \in \mathcal{F}$ satisfies the following property: For any $0 < \tau < 1, \epsilon > 0$, there exist a set $\mathcal{A}$ and a positive number x_0 such that $\tilde{\Pi}(\mathcal{A}) > 1 - \tau$ and for any $f = \phi * P$ with $P \in \mathcal{A}$,

$$\int_{|x|>x_0} f_0(x) \log \frac{f_0(x)}{f(x)} dx < \epsilon. \tag{15}$$

Then, $f_0 \in KL(\Pi)$.

Note that the moment condition of Theorem 3.1.2 is substantially reduced.

Let f_0 be a density on $\mathbb{R}$ satisfying

1. $\int f_0(x) \log f_0(x) dx < \infty$.
2. $\exists\ \eta \in (0,1)$ such that $\int |x|^\eta f_0(x) dx < \infty$.

Further assume that there exist $\sigma_0 > 0$, $0 < \beta < \eta$, $\gamma > \beta$ and $b_1, b_2 > 0$ such that for large $x > 0$

3. $\max\left(G_0\left(\left[x - \sigma_0 x^{\frac{\eta}{2}}, \infty\right) \times [\sigma_0, \infty)\right), G_0\left([0,\infty) \times (x^{1-\frac{\eta}{2}}, \infty)\right)\right) \geq b_1 x^{-\beta}$
4. $G_0\left((-\infty, x) \times (0, e^{|x|^\eta - \frac{1}{2}})\right) > 1 - b_2 |x|^{-\gamma}$.

and for large $x < 0$,

3′. $\max\left(G_0\left(\left(-\infty, x + \sigma_0 |x|^{\frac{\eta}{2}}\right] \times [\sigma_0, \infty)\right), G_0\left((-\infty, 0] \times (|x|^{1-\frac{\eta}{2}}, \infty)\right)\right) \geq b_1 |x|^{-\beta}$
4.′ $G_0\left((x, \infty) \times (0, e^{|x|^\eta - \frac{1}{2}})\right) > 1 - b_2 |x|^{-\gamma}$.

then $f_0 \in KL(\Pi)$. Other than the important moment condition on f_0 this theorem also requires some regularity in the tail of the base measure G_0. For example, assumption *3,3′* requires the tail of G_0 not to decay faster than a polynomial rate for the scale parameter σ. This condition seems very reasonable since the Cauchy density itself can be written as a scale mixture of normals with the mixing density having a polynomial decay towards infinity.

A standard choice for G_0 is the conjugate normal-inverse gamma distribution (see Escobar and West 1995), under which, $\theta|\sigma \sim N(0, \xi\sigma^2)$ and $\sigma^{-2} \sim Gamma(r, \lambda)$, for some $\xi, r, \lambda > 0$. For such a G_0 with $r \in (1/2, 1)$, one can show that the conditions of Theorem 3.2 hold true with $\eta \in (2r/(1+r), 1)$, $\beta = r(2-\eta)$ and $\gamma = 2r$. For example, the conditions in *Assumptions 3, 3′* are satisfied since,

$$G_0\left([0,\infty) \times (x^{1-\frac{\eta}{2}}, \infty)\right) = \frac{1}{2} \Pr(\sigma^{-2} \leq x^{-(2-\eta)}) = c \int_0^{x^{-(2-\eta)}} v^{r-1} e^{-\lambda v} dv \leq c' x^{-r(2-\eta)},$$

http://dx.doi.org/10.5772/56949

for some positive constants c, c'. To see that the conditions of *Assumptions 4, 4'* also hold, note that,

$$1 - G_0\left((-\infty, x) \times (0, e^{|x|^\eta - \frac{1}{2}})\right) \leq \Pr(\theta > x) + \Pr(\sigma^{-2} < e^{-2|x|^\eta + 1}).$$

An argument similar to the one provided above shows that the second term, namely, $\Pr(\sigma^{-2} < e^{-2|x|^\eta+1})$ is bounded by a constant times $e^{-2r|x|^\eta + r}$. Therefore, this term can be made smaller than $c|x|^{-\gamma}$ for a suitable constant c. Now, using the inequality $1 - \Phi(X) \leq (1/x)\phi(x)$, where $\Phi(\cdot)$ and $\phi(\cdot)$ are the standard normal distribution and density functions, we obtain

$$\Pr(\theta > x) \leq \frac{c}{x}\int_0^\infty v^{r-1/2-1} e^{-(\frac{x^2}{2\xi^2}+\lambda)v} dv = \frac{c'}{x(\frac{x^2}{2\xi^2}+\lambda)^{r-1/2}} \leq \frac{c''}{x^{2r}}$$

for some positive constants c, c', c''. The desired inequality follows from these two bounds. Therefore, such a choice of G_0 would lead to posterior consistency, for example, when f_0 is a Cauchy density.

Proof. **of Theorem 3.2** We simply need show that such an f_0 satisfies the condition of Lemma 3.2. Let $w(x) = \exp(-x^\eta),\ x \geq 0$. Define a class of subsets of $\mathbb{R} \times \mathbb{R}^+$ indexed by $x \in \mathbb{R}$, as follows:

$$K_x = \left\{(\theta, \sigma) \in \mathbb{R} \times \mathbb{R}^+ : \frac{1}{\sigma}\phi\left(\frac{x-\theta}{\sigma}\right) \geq \frac{1}{\sqrt{2\pi}} w(|x|)\right\} \tag{16}$$

These sets are of particular interest, since for $f = \phi * P$,

$$\begin{aligned}
\int_{|x|>x_0} f_0(x) \log \frac{f_0(x)}{f(x)} dx &\leq \int_{|x|>x_0} f_0(x) \log \frac{f_0(x)}{\int_{K_x} \frac{1}{\sigma}\phi\left(\frac{x-\theta}{\sigma}\right) dP(\theta,\sigma)} dx \\
&\leq \int_{|x|>x_0} f_0(x) \log \frac{f_0(x)}{\frac{1}{\sqrt{2\pi}} w(|x|) P(K_x)} dx \\
&\leq \int_{|x|>x_0} f_0(x) \left\{\log f_0(x) + |x|^\eta + \log \frac{\sqrt{2\pi}}{P(K_x)}\right\} dx.
\end{aligned} \tag{17}$$

By the assumptions of the Theorem, this quantity can be made arbitrarily small for a suitably large x_0 if we can show that $P(K_x) > c_1 \exp(-c_2|x|^\eta)$ for all $|x| > x_0$ for some fixed constants $c_1, c_2 > 0$. Therefore it suffices to prove that, For any $\tau > 0$ there exists an $x_0 > 0$ and a set $\mathcal{A}$ with $\tilde{\Pi}(\mathcal{A}) > 1 - \tau$ such that $P \in \mathcal{A} \Rightarrow P(K_x) \geq (1/2)\exp(-2|x|^\eta / b_1)$ for all $|x| > x_0$.

The proof of this Lemma is fairly technical. It makes an extensive use of the tail behavior of a random probability P arising from a Dirichlet process. For clarity of reading, we present details of the proof in the Appendix.

□

4. Density estimation: strong consistency

We establish L_1-consistency of a Dirichlet location-scale mixture of normal prior Π by verifying the conditions of Theorem 8 of Ghosal *et al.* (1999). This theorem is reproduced below.

Let Π be a prior on $\mathcal{F}$ such that $f_0 \in KL(\Pi)$. If there is a $\delta < \epsilon/4, c_1, c_2 > 0, \beta < \epsilon^2/8$ and $\mathcal{F}_n \subseteq \mathcal{F}$ such that for all n large,

1. $\Pi(\mathcal{F}_n^c) < c_1 e^{-nc_2}$,
2. $J(\delta, \mathcal{F}_n) < n\beta$,

then Π achieves strong posterior consistency at f_0.

Here $J(\delta, \mathcal{G})$ denotes logarithm of the covering number of $\mathcal{G}$ by L_1 balls of radii δ.

We first show how to calculate $J(\delta, \mathcal{G})$ for certain type of sets $\mathcal{G}$. For some $a > 0,\ u > l > 0$ define

$$\mathcal{F}_{a,l,u} = \{f = \phi * P : P((-a, a] \times (l, u]) = 1\} \tag{18}$$

Then,

$$J(2\kappa, \mathcal{F}_{a,l,u}) \leq b_0 \left(b_1 \frac{a}{l} + b_2 \log \frac{u}{l} + 1\right). \tag{19}$$

where b_0, b_1 and b_2 depend upon κ but not on a, l or u.

Proof. Let $\phi_{\theta,\sigma}$ denote the normal density with mean θ and standard deviation σ. For $\sigma_2 > \sigma_1 > \sigma_2/2$, it can be shown that,

$$\begin{aligned} \left\|\phi_{\theta_1,\sigma_1} - \phi_{\theta_2,\sigma_2}\right\| &\leq \left\|\phi_{\theta_1,\sigma_2} - \phi_{\theta_2,\sigma_2}\right\| + \left\|\phi_{\theta_2,\sigma_1} - \phi_{\theta_2,\sigma_2}\right\| \\ &\leq \sqrt{\frac{2}{\pi}} \frac{|\theta_2 - \theta_1|}{\sigma_2} + 3\frac{\sigma_2 - \sigma_1}{\sigma_1}. \end{aligned} \tag{20}$$

Let $\zeta = \min(\kappa/6, 1)$. Define $\sigma_m = l(1+\zeta)^m,\ m \geq 0$. Let M be the smallest integer such that $\sigma_M = l(1+\zeta)^M \geq u$. This implies $M \leq (1+\zeta)^{-1} \log(u/l) + 1$. For $1 \leq j \leq M$, let $N_j = \left\lceil \sqrt{\frac{32}{\pi}} a/(\kappa\sigma_{j-1}) \right\rceil$. For $1 \leq i \leq N_j$; $1 \leq j \leq M$, define

$$E_{ij} = \left(-a + \frac{2a(i-1)}{N_j},\ -a + \frac{2ai}{N_j}\right] \times (\sigma_{j-1}, \sigma_j]. \tag{21}$$

Then, $(\theta, \sigma),\ (\theta', \sigma') \in E_{ij} \Rightarrow \left\|\phi_{\theta,\sigma} - \phi_{\theta',\sigma'}\right\| < \kappa$. Take $N = \sum_{j=1}^{M} N_j$ and let

$$\mathcal{P}_N = \left\{(P_{11}, \cdots, P_{N_1 1}, \cdots, P_{1M}, \cdots, P_{N_M M}) : P_{ij} \geq 0, \sum_{ij} P_{ij} = 1\right\} \tag{22}$$

http://dx.doi.org/10.5772/56949

be the N dimensional probability simplex and $\mathcal{P}_N^*$ be a κ-net in $\mathcal{P}_N$. Let τ_j's be as before and $\theta_{ij} = -a + 2a(i - 1/2)/N_j$, $1 \le i \le N_j$, $1 \le j \le M$. So $(\theta_{ij}, \sigma_j) \in E_{ij}\ \forall i, j$. It can be shown by following an argument similar to the one presented in the proof of Lemma 1 of Ghosal *et al.* (1999) that ,

$$\mathcal{F} = \left\{ \sum_{j=1}^{M} \sum_{i=1}^{N_j} P_{ij}^* \phi_{\theta_{ij},\sigma_j} : P^* \in \mathcal{P}_N^* \right\} \tag{23}$$

is a 2κ-net in $\mathcal{F}_{a,l,u}$ and consequently, $J(2\kappa, \mathcal{F}_{a,l,u}) \le J(\kappa, \mathcal{P}_N) \le N\left(1 + \log \frac{1+\kappa}{\kappa}\right)$. But,

$$\begin{aligned} N &\le \sum_{j=1}^{M} \left(\sqrt{\frac{32}{\pi}} \frac{a}{\sigma_{j-1}\kappa} + 1 \right) = \sqrt{\frac{32}{\pi}} \frac{a}{l\kappa} \sum_{j=0}^{M-1} (1+\zeta)^{-j} + M \\ &\le \sqrt{\frac{32}{\pi}} \frac{a}{l} \frac{1+\zeta}{\kappa\zeta} + \frac{1}{1+\zeta} \log \frac{u}{l} + 1 \\ &= b_1 \frac{a}{l} + b_2 \log \frac{u}{l} + 1 \end{aligned} \tag{24}$$

From this the result follows with $b_0 = 1 + \log \frac{1+\kappa}{\kappa}$. □

Let $\mathcal{F}_{a,l,u}^{\kappa} = \{f = \phi * P : P((-a, a] \times (l, u]) \ge 1 - \kappa\}$. Then $J(3\kappa, \mathcal{F}_{a,l,u}^{\kappa}) \le J(\kappa, \mathcal{F}_{a,l,u})$.

Proof. Let $f = \phi * P \in \mathcal{F}_{a,l,u}^{\kappa}$. Consider the probability measure defined by $P^*(A) = P(A \cap (-a, a] \times (l, u])/P((-a, a] \times (l, u])$. Then the density $f^* = \phi * P^*$ clearly belongs to $\mathcal{F}_{a,l,u}$ and further satisfies $\|f - f^*\| < 2\kappa$. This proves the lemma.

□

Suppose for each $\kappa > 0, \beta > 0$, there exist sequences of positive numbers $a_n, u_n \uparrow \infty, l_n \downarrow 0$ with $l_n < u_n$ and constant β_0, all depending on κ and β such that

1. $\tilde{\Pi}(\{P : P((-a_n, a_n] \times (l_n, u_n]) < 1 - \kappa\}) < e^{-n\beta_0}$,
2. $a_n/l_n < n\beta$, $\log(u_n/l_n) < n\beta$.

then $f_0 \in KL(\Pi)$ implies that Π achieves strong posterior consistency at f_0.

Proof. Take $\mathcal{F}_n = \mathcal{F}_{a_n,l_n,u_n}^{\kappa}$. Then the conditions of Theorem 4 are easily verified using Lemma 4 for a suitable choice of $\kappa > 0$. □

If $\tilde{\Pi} = Dir(\alpha G_0)$, verification of conditions 1 and 2 becomes particularly simple. For example, if G_0 is a product of a normal on θ and an inverse gamma on σ^2, then the conditions of theorem 4 are satisfied if $a_n = O(\sqrt{n}), l_n = O(1/\sqrt{n})$ and $u_n = O(e^n)$.

4.1. Support vector machines

We give, in this section, a very brief presentation of Support Vector Machines (SVMs) that is needed for the definition of their functional versions. We refer the reader to e.g. [4] for a more comprehensive presentation. As stated in section **??**, $\mathcal{X}$ denotes an arbitrary Hilbert space. Our presentation of SVM departs from the standard introduction because it assumes that the observations belong to $\mathcal{X}$ rather than to a d. This will make clear that the definition of SVM on arbitrary Hilbert spaces is not the difficult part in the construction of functional SVM. We will discuss problems related to the functional nature of the data in section 4.1.5.

Our goal is to classify data into two predefined classes. We assume given a learning set, i.e. N examples $(x_1, y_1), \ldots, (x_N, y_N)$ which are i.i.d. realizations of the random variable pair (X, Y) where X has values in $\mathcal{X}$ and Y in $\{-1, 1\}$, i.e. Y is the class label for X which is the observation.

4.1.1. Hard margin SVM

The principle of SVM is to perform an affine discrimination of the observations with maximal margin, that is to find an element $w \in \mathcal{X}$ with a minimum norm and a real value b, such that $y_i(\langle w, x_i\rangle + b) \geq 1$ for all i. To do so, we have to solve the following quadratic programming problem:

$$(P_0) \quad \min_{w,b} \langle w, w\rangle, \text{ subject to } y_i(\langle w, x_i\rangle + b) \geq 1,\ 1 \leq i \leq N.$$

The classification rule associated to (w, b) is simply $(x) = \text{sign}(\langle w, x\rangle + b)$. In this situation (called hard margin SVM), we request the rule to have zero error on the learning set.

4.1.2. Soft margin SVM

In practice, the solution provided by problem (P_0) is not very satisfactory. Firstly, perfectly linearly separable problems are quite rare, partly because non linear problems are frequent, but also because noise can turn a linearly separable problem into a non separable one. Secondly, choosing a classifier with maximal margin does not prevent overfitting, especially in very high dimensional spaces (see e.g. [5] for a discussion about this point).

A first step to solve this problem is to allow some classification errors on the learning set. This is done by replacing (P_0) by its soft margin version, i.e., by the problem:

$$\begin{aligned}(P_C)\ &\min_{w,b,\xi} \langle w, w\rangle + C\sum_{i=1}^{N} \xi_i,\\ &\text{subject to } y_i(\langle w, x_i\rangle + b) \geq 1 - \xi_i,\ 1 \leq i \leq N,\\ &\qquad\qquad \xi_i \geq 0,\ 1 \leq i \leq N.\end{aligned}$$

Classification errors are allowed thanks to the slack variables ξ_i. The C parameter acts as an inverse regularization parameter. When C is small, the cost of violating the hard margin constraints, i.e., the cost of having some $\xi_i > 0$ is small and therefore the constraint on w dominates. On the contrary, when C is large, classification errors dominate and (P_C) gets closer to (P_0).

http://dx.doi.org/10.5772/56949

4.1.3. Non linear SVM

As noted in the previous section, some classification problems don't have a satisfactory linear solution but have a non linear one. Non linear SVMs are obtained by transforming the original data. Assume given an Hilbert space $\mathcal{H}$ (and denote $\langle .,. \rangle_{\mathcal{H}}$ the corresponding inner product) and a function ϕ from $\mathcal{X}$ to $\mathcal{H}$ (this function is called a *feature map*). A linear SVM in $\mathcal{H}$ can be constructed on the data set $(\phi(x_1), y_1), \ldots, (\phi(x_N), y_N)$. If ϕ is a non linear mapping, the classification rule $(x) = \text{sign}(\langle w, \phi(x) \rangle_{\mathcal{H}} + b)$ is also non linear.

In order to obtain the linear SVM in $\mathcal{H}$ one has to solve the following optimization problem:

$$\begin{aligned}(P_{C,\mathcal{H}})\ &\min_{w,b,\xi} \langle w, w \rangle_{\mathcal{H}} + C \sum_{i=1}^{N} \xi_i, \\ &\text{subject to } y_i(\langle w, \phi(x_i) \rangle_{\mathcal{H}} + b) \geq 1 - \xi_i,\ 1 \leq i \leq N, \\ &\qquad \xi_i \geq 0,\ 1 \leq i \leq N.\end{aligned}$$

It should be noted that this feature mapping allows to define SVM on almost arbitrary input spaces.

4.1.4. Dual formulation and Kernels

Solving problems (P_C) or $(P_{C,\mathcal{H}})$ might seem very difficult at first, because $\mathcal{X}$ and $\mathcal{H}$ are arbitrary Hilbert spaces and can therefore have very high or even infinite dimension (when $\mathcal{X}$ is a functional space for instance). However, each problem has a dual formulation. More precisely, (P_C) is equivalent to the following optimization problem (see [6]):

$$\begin{aligned}(D_C)\ &\max_{\alpha} \sum_{i=1}^{N} \alpha_i - \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \alpha_j y_i y_j \langle x_i, x_j \rangle, \\ &\text{subject to } \sum_{i=1}^{N} \alpha_i y_i = 0, \\ &\qquad 0 \leq \alpha_i \leq C,\ 1 \leq i \leq N.\end{aligned}$$

This result applies to the original problem in which data are not mapped into $\mathcal{H}$, but also to the mapped data, i.e., $(P_{C,\mathcal{H}})$ is equivalent to a problem $(D_{C,\mathcal{H}})$ in which the x_i are replaced by $\phi(x_i)$ and in which the inner product of $\mathcal{H}$ is used. This leads to:

$$\begin{aligned}(D_{C,\mathcal{H}})\ &\max_{\alpha} \sum_{i=1}^{N} \alpha_i - \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \alpha_j y_i y_j \langle \phi(x_i), \phi(x_j) \rangle_{\mathcal{H}}, \\ &\text{subject to } \sum_{i=1}^{N} \alpha_i y_i = 0, \\ &\qquad 0 \leq \alpha_i \leq C,\ 1 \leq i \leq N.\end{aligned}$$

Solving $(D_{C,\mathcal{H}})$ rather than $(P_{C,\mathcal{H}})$ has two advantages. The first positive aspect is that $(D_{C,\mathcal{H}})$ is an optimization problem in N rather than in $\mathcal{H}$ which can have infinite dimension (the same is true for $\mathcal{X}$).

The second important point is linked to the fact that the optimal classification rule can be written $(x) = \text{sign}(\sum_{i=1}^{N} \alpha_i y_i \langle \phi(x_i), \phi(x) \rangle_{\mathcal{H}} + b)$. This means that both the optimization problem and the classification rule do not make direct use of the transformed data, i.e. of the $\phi(x_i)$. All the calculations are done through the inner product in $\mathcal{H}$, more precisely through the values $\langle \phi(x_i), \phi(x_j) \rangle_{\mathcal{H}}$. Therefore, rather than choosing directly $\mathcal{H}$ and ϕ, one can provide a so called *Kernel function* K such that $K(x_i, x_j) = \langle \phi(x_i), \phi(x_j) \rangle_{\mathcal{H}}$ for a given pair $(\mathcal{H}, \phi)$.

In order that K corresponds to an actual inner product in a Hilbert space, it has to fulfill some conditions. K has to be symmetric and positive definite, that is, for every N, $x_1, \ldots, x_N$ in $\mathcal{X}$ and $\alpha_1, \ldots, \alpha_N$ in , $\sum_{i=1}^{N}\sum_{j=1}^{N} \alpha_i \alpha_j K(x_i, x_j) \geq 0$. If K satisfies those conditions, according to Moore-Aronszajn theorem [**?**], there exists a Hilbert space $\mathcal{H}$ and feature map ϕ such that $K(x_i, x_j) = \langle \phi(x_i), \phi(x_j) \rangle_{\mathcal{H}}$.

4.1.5. The case of functional data

The short introduction to SVM proposed in the previous section has clearly shown that defining linear SVM for data in a functional space is as easy as for data in d, because we only assumed that the input space was a Hilbert space. By the dual formulation of the optimization problem (P_C), a software implementation of linear SVM on functional data is even possible, by relying on numerical quadrature methods to calculate the requested integrals (inner product in $L^2(\mu)$, cf section **??**).

However, the functional nature of the data has some effects. It should be first noted that in infinite dimensional Hilbert spaces, the hard margin problem (P_0) has always a solution when the input data are in general positions, i.e., when N observations span a N dimensional subspace of $\mathcal{X}$. A very naive solution would therefore consists in avoiding soft margins and non linear kernels. This would not give very interesting results in practice because of the lack of regularization (see [5] for some examples in very high dimension spaces, as well as section **??**).

Moreover, the linear SVM with soft margin can also lead to bad performances. It is indeed well known (see e.g. [7]) that problem (P_C) is equivalent to the following unconstrained optimization problem:

$$(R_\lambda) \min_{w,b} \frac{1}{N} \sum_{i=1}^{N} \max\left(0, 1 - y_i(\langle w, x_i \rangle + b)\right) + \lambda \langle w, w \rangle,$$

with $\lambda = \frac{1}{CN}$. This way of viewing (P_C) emphasizes the regularization aspect (see also [8–10]) and links the SVM model to ridge regression [**?**]. As shown in [11], the penalization used in ridge regression behaves poorly with functional data. Of course, the loss function used by SVM (the *hinge loss*, i.e., $h(u, v) = \max(0, 1 - uv)$) is different from the quadratic loss used in ridge regression and therefore no conclusion can be drawn from experiments reported in [11]. However they show that we might expect bad performances with the linear SVM applied directly to functional data. We will see in sections **??** and **??** that the efficiency of the ridge regularization seems to be linked with the actual dimension of the data: it does not behave very well when the number of discretization points is very big and thus leads to approximate the ridge penalty by a dot product in a very high dimensional space (see also section **??**).

It is therefore interesting to consider non linear SVM for functional data, by introducing adapted kernels. As pointed out in e.g. [10], $(P_{C,\mathcal{H}})$ is equivalent to

$$(R_{\lambda,\mathcal{H}}) \min_{f \in \mathcal{H}} \frac{1}{N} \sum_{i=1}^{N} \max\left(0, 1 - y_i f(x_i)\right)) + \lambda \langle f, f \rangle_{\mathcal{H}}.$$

http://dx.doi.org/10.5772/56949

Using a kernel corresponds therefore both to replace a linear classifier by a non linear one, but also to replace the ridge penalization by a penalization induced by the kernel which might be more adapted to the problem (see [9] for links between regularization operators and kernels). The applications presented in **??** illustrate this fact.

5. Proposed approach

5.1. Overview of the proposed fusion schema

In this chapter, we propose a new technique in remote-sensing images classification by fusing heterogeneous representations. The proposed approach involve several steps including preprocessing; features extraction; features fusion; matching and classification stages. The block diagram of the proposed technique is shown in Fig. 1. In our previous work [12], we proposed a novel 3D model which design the spectral signature as a three dimensional function which are the time, reflectance, and wavelength band (equation 1). For each pixel, we generated a surface (3D Mesh) which generalizes the usual signature by adding a time dimension. We call this new representation the *multi-temporal spectral signature*. Interested readers can refer to [12].

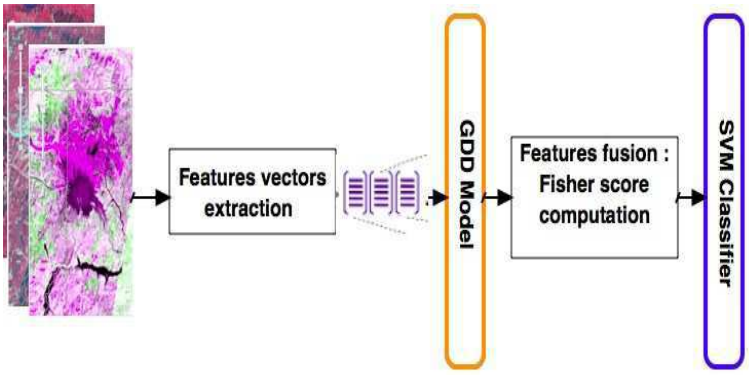


Figure 1. General workflow of the proposed approach

5.2. Images pre-processing and features extraction

In this study multi-temporal hyperspectral images constitutes the source data. Spectral and textural features are the foundational data for this kind of images. The 3D spectral features are extracted from the relative mesh of a given pixel (multi-temporal spectral signature) while the textural ones are derived directly from images. Mainly, two features vectors are generated for each pixel as follows:

Heat kernel signature (HKS) : The HKS is a signature computed only from the intrinsic geometry of an object. Suppose (m, g) is a compelte Riemannian manifold, g is the Riemannian metric. δ is the Laplace-Beltrami operator. The eigenvalues $\{\lambda_n\}$ and eigenfunctions $\{\phi_n\}$ of δ are $\delta\phi_n = \lambda_n\phi_n$, where ϕ_n is normalized to be orthonormal in $L^2(M)$. The Laplace spectrum is given by $0 = \lambda_0 < \lambda_1 \leq \lambda_2 \leq \ldots, \lambda_n \to \infty$. $\triangle$ is the

Laplace-Beltrami operator. As a local shape descriptor, Sun et al. [**?**] defined the heat kernel signature (HKS) by :

$$h(x,t) = K_{x,t}(x,x) = \sum_{i=0}^{\infty} e^{-\lambda_t} \phi_i^2(x) \tag{25}$$

where $\lambda_0, \lambda_1, \cdots \geq O$ are eigenvalues and $_0, \phi_1, \ldots$ are the corresponding eigenfuctions on the Laplace-Beltrami operator, satisfying $\delta_X \phi_i = \lambda_i \phi_i$. Let's denote this vector by Y.

Spatio-temporal Gabor filters: Texture is one of the important characteristics used in identifying objects or regions of interest. It contains important information about the structural arrangement of surfaces. Fusing texture with 3D spectral information is conducive to the interpretation of remote seeing image [13]. We use a method for dynamic texture modeling based on spatio-temporal Gabor filters. Briefly, the sequence of images is convolved with a bank of spatiotemporal Gabor filters and a feature vector is constructed with the energy of the responses as components. Let's denote this vector by Y'.

5.3. Multi-Features fusion based on a cooperative GDD/SVM classifier

In this section, we present an approach that combines an SVM classifier [1] with a generatively trained GDD model and profits, accordingly, from the advantages of both techniques. The key idea here is to concatenate the extracted features into one vector and to project it in a new space. First, a straightforward feature combination approach is used to concatenate feature vectors (Y and Y') to a single feature vector $X = (X_{i1}, \ldots, X_{idim})$. The *dim* size may differ from one pixel to another making the fusion and classification a challenging tasks. To overcome this limit, we use the Generalized Dirichelet Distribution (GDD) model [14] to map each feature vector into its Fisher score. Therefore, the Fisher kernel function from the GDD is used to replace the Gaussian kernel in the classical SVM.

Let $(X_1, \ldots, X_N)$ denote a collection of N multi-temporal hyperspectral pixels. Each data X_i is assumed to have *dim* size, $X = (X_{i1}, \ldots, X_{idim})$. Each data X_i is assumed to be drawn from the following finite mixture model :

$$p(X_i/\theta) = \sum_{j=1}^{M} p(X_i/j, \theta_j) P(j) \tag{26}$$

where M is the number of components, the $P(j)$, $(0 < P(j) < 1$ and $\sum_{j=1}^{dim} P(j) = 1)$ are the mixing proportions and $p(X/j, \theta_j)$ is the Probability Density Function PDF. θ is the set of parameters to be estimated : $\theta = (\alpha_1, \ldots, \alpha_M, P(1), \ldots, P(M))$.

If the random vector $X = (X_{i1}, \ldots, X_{idim})$ follows a Dirichelet distribution, the joint density function is given by :

$$X = (X_{i1}, \ldots, X_{idim}) = \frac{\tau(|\alpha|)}{\prod_{i=1}^{dim+1} \tau(\alpha_i)} \prod_{i=1}^{dim+1} X_i^{\alpha_i - 1} \tag{27}$$

http://dx.doi.org/10.5772/56949

Since that each feature vector X may has an arbitrary dimension, the proposed method defines the fusion as a projection from one feature vector space (spectral bands) to another with a fixed dimentionnality. Accordingly, the feature-level fusion is done by projecting the vector X combining into one vector in the Fisher space. Thus, the generative model will have its impact on the final classification result through the projection of the extracted features in this new space.

SVM classifier is used to classify the fused features and the multi-temporal dataset of images. Given the generative model obtained by GDD with parameters θ, we compute for each sample X the Fisher score $U_d = \bigtriangledown_\theta log P(x|\theta)$ (the gradient of the log likehood of x for model θ). The Fisher kernel operates in the gradient space of the generative mode and provides a natural similarity measure between data samples. For each sample, this score is a vector of fixed dimentionality. Using this score, the Fisher Information matrix is defined as $\mathbb{I} = E_{X_i}\left\{{U_{X_i}}^T U_{X_i}\right\}$. After Fisher score normalization, we compute the Fisher kernel function on the basis of the Euclidean distance between the scores of the new sample and the training samples :

$$K(X, X^{'}) = U_{X_i}\, \mathbb{I}^{-1} U_{X_i^{'T}} \tag{28}$$

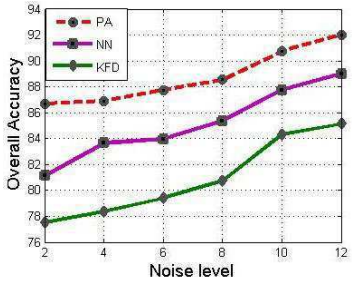
In the second stage, suppose our training set S consists of labels input vectors $(X_i, z_j), i = 1, \ldots, m$ where $X_i \in \mathbb{R}^n$ and $z_i \in \{\pm 1\}$. Given a kernel matrix and a set of labels z_i for each sample, the SVM proceeds to learn a classifier of the form,

$$z(x) = sign(\sum_i \alpha_i z_i) K(X_i, X)) \tag{29}$$

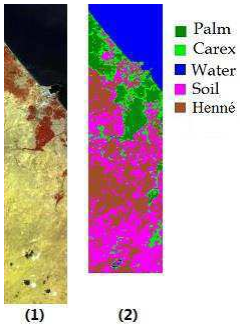
where the coefficients α_i are determined by solving a constrained quadratic program which aims to maximize the margin between the classes. In our experiments we used the LIBSVM package. Our research deals with multi-class problem. The One-Vs-One approach is adopted to extend the proposed approach to multi-temporal hyperspectral classification.

6. Results and discussion

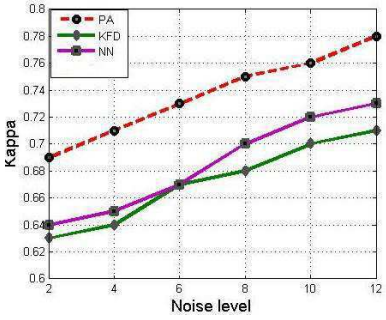
The proposed approach was tested on two different data sets. The datasets involve several types of information with dimensions ranging from 176 to 183 bands. The first dataset, *Hyperion*, contains vegetation type data, is divided into five classes, has 183 spectral bands and has a pixel size of $30m$. The second set is from an airborne sensor (*AVIRIS*), divided into 7 classes, has 176 spectral bands and a pixel size of $18m$. First, we present experiments that assess the classification accuracy of the proposed approach (PA). We also included the direct SVM fusion and a probabilistic fusion approach in our comparison as a baseline. Figure (2) summarizes the results obtained. At each level of label noise we carry out four experiments, and the figures show the mean performance. The strength of this approach is that it combines the rich modeling power of GDD with the discriminative power of the SVM algorithm.



(a) Overall accuracy of the EKFD [Both two sets]



(b-1) Map of ground truth
(b-2) Result of classificationwith EKFD [First set]



(c) Overall accuracy of the EKFD [Two sets]

Figure 2. Experimental results.

Author details

Selim Hemissi and Imed Riadh Farah

RIADII-SIIVT, Tunisia

References

[1] Christopher J. C. Burges. A tutorial on support vector machines for pattern recognition. *Data Min. Knowl. Discov.*, 2(2):121–167, June 1998.

[2] Ilkay Ulusoy and Christopher M Bishop. Comparison of generative and discriminative techniques for object detection and classification. *Toward CategoryLevel Object Recognition*, pages 173–195, 2006.

[3] John Paisley and Lawrence Carin. Dirichlet process mixture models with multiple modalities. In *Acoustics, Speech and Signal Processing, 2009. ICASSP 2009. IEEE International Conference on*, pages 1613–1616, 2009.

http://dx.doi.org/10.5772/56949

[4] Nello Cristianini and John Shawe-Taylor. *An Introduction to Support Vector Machines*. Cambridge University Press, Cambridge, UK, 2000.

[5] Trevor Hastie, Saharon Rosset, Robert Tibshirani, and Ji Zhu. The entire regularization path for the support vector machine. *Journal of Machine Learning Research*, 5:1391–1415, October 2004.

[6] Chih-Jen Lin. Formulations of support vector machines: a note from an optimization point of view. *Neural Computation*, 2(13):307–317, 2001.

[7] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer-Verlag, 2001.

[8] Alexander Smola and Bernhard SchÅąlkopf. On a kernel-based method for pattern recognition, regression, approximation and operator inversion. *Algorithmica*, 22(1064):211–231, 1998.

[9] Alexander Smola, Bernhard SchÅąlkopf, and Klaus-Robert MÅÿller. The connection between regularization operators and support vector kernels. *Neural Networks*, 11:637–649, 1998.

[10] T. Evgeniou, M. Pontil, and T. Poggio. Regularization networks and support vector machines. *Advances in Computational Mathematics*, 13(1):1–50, 2000.

[11] T. Hastie, A. Buja, and R. Tibshirani. Penalized discriminant analysis. *Annals of Statistics*, 23:73–102, 1995.

[12] Imed Riadh Farah, Selim Hemissi, Karim Saheb Ettabaa, and Bassel Souleiman. Multi-temporal Hyperspectral Images Unmixing and Classification Based on 3D Signature Model and Matching. *Piers Online*, 6:480–484, 2010.

[13] Y Wang and C Chua. Face recognition from 2d and 3d images using 3d gabor filters. *Image and Vision Computing*, 23(11):1018–1028, 2005.

[14] Nizar Bouguila and Djemel Ziou. A dirichlet process mixture of generalized dirichlet distributions for proportional data modeling. *IEEE Transactions on Neural Networks*, 21(1):107–122, 2010.