Available online at www.sciencedirect.com

SCIENCE DIRECT®

Neurocomputing 69 (2006) 714–720

NEUROCOMPUTING

www.elsevier.com/locate/neucom

Kernel methods and the exponential family

Stéphane Canu[a,*], Alex Smola[b,c]

[a] 1-PSI-FRE CNRS 2645, INSA de Rouen, France, St Etienne du Rouvray, France
[b] Statistical Machine Learning Program, National ICT, Australia
[c] RSISE, Australian National University, Canberra, 0200 ACT, Australia

Available online 25 January 2006

Abstract

The success of support vector machine (SVM) has given rise to the development of a new class of theoretically elegant learning machines which use a central concept of kernels and the associated reproducing kernel Hilbert space (RKHS). Exponential families, a standard tool in statistics, can be used to unify many existing machine learning algorithms based on kernels (such as SVM) and to invent novel ones quite effortlessly. A new derivation of the novelty detection algorithm based on the one class SVM is proposed to illustrate the power of the exponential family model in an RKHS.

© 2005 Published by Elsevier B.V.

Keywords: Kernel methods; Exponential families; Novelty detection

1. Introduction

Machine learning is providing increasingly important tools in many fields such as text processing, machine vision, speech, to name just a few. Among these new tools, kernel based algorithms have demonstrated their efficiency on many practical problems. These algorithms performed function estimation, and the functional framework behind these algorithms is now well known [3]. But still too little is known about the relation between these learning algorithms and more classical statistical tools such as likelihood, likelihood ratio, estimation and test theory. A key model to understand this relation is the generalized or non parametric exponential family. This exponential family is a generic way to represent any probability distribution since any distribution can be well approximated by an exponential distribution. The idea here is to retrieve learning algorithm by using the exponential family model with classical statistical principle such as the maximum penalized likelihood estimator or the generalized likelihood ratio test.

Outline. To do so the paper is organized as follows. The first section presents the functional framework and Reproducing Kernel Hilbert Space (RKHS). Then the exponential family on an RKHS is introduced and classification as well as density estimation and regression kernel-based algorithms such as SVM are derived. In the final section new material is presented establishing the link between the kernel-based one-class SVM novelty detection algorithm and classical test theory. It is shown how this novelty detection can be seen as an approximation of a generalized likelihood ratio and, therefore, as an optimal test.

2. Functional framework

Learning can be seen as retrieving a relevant function among a large class of possible hypotheses. One strength of kernel-based learning algorithm is the ability to express the set of hypotheses (functions) in terms of a kernel representing the way to evaluate a function at a given point. This is possible because the relationship between the kernel and the underlying functional space is constituent. There exists a bijection between a large class of useful kernels (positive ones) and interesting functional spaces: the so-called reproducing kernel Hilbert spaces. Details regarding this bijection are now precised. Let $\mathscr{X}$ be the learning domain (typically $\mathscr{X} \subseteq \mathbb{R}^d$).

Definition 1 (*Reproducing Kernel Hilbert Space—RKHS*). A Hilbert space $(\mathscr{H}, \langle .,. \rangle_{\mathscr{H}})$ of functions on a domain $\mathscr{X}$

*Corresponding author. Fax: +33 2 32 95 97 08.
E-mail addresses: Stephane.Canu@insa-rouen.fr, scanu@insa-rouen.fr (S. Canu), Alex.Smola@nicta.com.au (A. Smola).

0925-2312/$ - see front matter © 2005 Published by Elsevier B.V.
doi:10.1016/j.neucom.2005.12.009

(defined pointwise) is an RKHS if the evaluation functional is continuous on $\mathcal{H}$.

For instance $\mathbb{R}^n$, the set $\mathcal{P}_k$ of polynomials of order k, as any finite dimensional set of genuine functions form an RKHS. The space of sequences ℓ_2 is also an RKHS, the evaluation function in this case being the value of the series at location $x \in \mathcal{X} = \mathbb{N}$. Note that the usual L_2 spaces, such as $L_2(\mathbb{R}^n)$ (using the Lebesgue measure) are not RKHS since their evaluation functionals are not continuous (in fact, L_2 elements are not even defined in a pointwise way). For more details see [3] and references therein.

Definition 2 (*Positive Kernel*). A mapping k from $\mathcal{X} \times \mathcal{X}$ to $\mathbb{R}$ is a positive kernel if it is symmetric and if for any finite integer n, any finite subset $\{x_i\}, i = 1, n$ of $\mathcal{X}$ and any sequence of scalar coefficients $\{\alpha_i\}, i = 1, n$ the following inequality holds:

$$\sum_{i=1}^{n}\sum_{j=1}^{n} \alpha_i\alpha_j K(x_i, x_j) \geqslant 0. \tag{1}$$

This definition is equivalent to the one of Aronszajn [2]. The following corollary arises from [11, Proposition 23] and [13, Theorem 1.1.1].

Proposition 3 (*Bijection between RKHS and Kernel*). *There is a bijection between the set of all possible RKHS and the set of all positive kernels.*

Thus Mercer kernels (as defined in [10]) are a particular case of a more general situation since every Mercer kernel is positive in the Aronszajn sense (Definition 2) while the converse need not be true.

It is always possible to associate with any positive kernel a set of function and a dot product (an RKHS) on which the evaluation functional is continuous. Conversely, the continuity of the evaluation functional guarantee the existence of a positive kernel. This continuity is the key property of the hypothesis set. It is associated with a useful property to be used hereafter: the reproducing property of the kernel k in an RKHS. It is closely related to the fact that, in RKHS, functions are pointwise-defined and the evaluation functional is continuous. Thus, because of this continuity, Riesz theorem can be stated as follows

$$\text{for all} f \in \mathcal{H} \text{and for all} x \in \mathcal{X}$$
$$\text{we have} f(x) = \langle f(.), k(x,.)\rangle_{\mathcal{H}}. \tag{2}$$

This continuity implies that two closed functions also have their pointwise values closed, for any x. In the framework of learning machines this means that if two hypotheses are closed, then you do not want their prediction at any point to differ too much.

In the remainder of the paper we assume that we are given the RKHS, its dot product and its kernel k. When appropriate we will refer to $k(x,.)$ as a map of x into the so-called "feature space". The dot product considered is the one of the RKHS.

3. Kernel approaches for the exponential family

3.1. Parametric exponential families

We begin by reviewing some basic facts of exponential families. Denote by $\mathcal{X}$ a domain, $\mu(\mathcal{X})$ a (not necessarily finite) measure on $\mathcal{X}$ and let $\phi(x)$ be a map from $\mathcal{X}$ into $\mathbb{R}^p$ (it is called the sufficient statistics). Then a probability measure on $\mathcal{X}$ with respect to ϕ and $\theta \in \Theta \subseteq \mathbb{R}^p$ is given by

$$\mathbb{P}(x;\theta) = \exp(\theta^\top \phi(x) - g(\theta)), \tag{3a}$$

where

$$g(\theta) = \log \int_{\mathcal{X}} \exp(\theta^\top \phi(x))\,\mathrm{d}\mu(x). \tag{3b}$$

Here it is understood that $\mathbb{P}(x;\theta)$ is taken with respect to the underlying measure $\mu(\mathcal{X})$. A large class of distributions can be described this way, as can be seen in Table 1. This table also shows examples of carrier measures μ and parameter domains Θ.

The function $g(\theta)$ is typically referred to as the log-partition function or, since its derivatives generate the cumulants of $\mathbb{P}(x;\theta)$, as the cumulant generating function. It is analytic and convex on the domain on which (3b) exists. Exponential representation can be seen as a change of parameterization of the distribution. The main interest of this transformation is, through the convexity of the log-partition function, the convexity of the associated likelihood with respect to the exponential parameters.

3.2. Nonparametric exponential families

Unfortunately, not all integrals in (3b) can be computed explicitly in closed form. Moreover, the set of distributions which can be modeled by mapping x into a finite-dimensional $\phi(x)$ is rather limited. We now consider an extension of the definition to nonparametric distributions.

Assume there exists a reproducing kernel Hilbert space $\mathcal{H}$ embedded with the dot product $\langle .,. \rangle_{\mathcal{H}}$ and with a reproducing kernel k such that kernel $k(x,.)$ is a sufficient statistics of x. Then in exponential families the density $\mathbb{P}(x;\theta)$ is given by

$$\mathbb{P}(x;\theta) = \mu(x)\exp(\langle \theta(.), k(x,.)\rangle_{\mathcal{H}} - g(\theta)), \tag{4a}$$

where

$$g(\theta) = \log \int_{\mathcal{X}} \exp(\langle \theta(.), k(x,.)\rangle_{\mathcal{H}})\,\mathrm{d}\mu(x), \tag{4b}$$

where μ is the *carrier density*, which can be absorbed into the underlying measure as above (and set to 1). Here θ is the natural parameter and $g(\theta)$ is the log-partition function, also called the cumulant generating function. All we changed from before is that now θ is an element of an RKHS and $\phi(x)$ is also a map into such a space, given by $\phi(x) = k(x,\cdot)$ thus

$$\langle \theta(.), k(x,.)\rangle_{\mathcal{H}} = \theta(x) \quad \text{and} \quad \mathbb{P}(x;\theta) = \mu(x)\exp^{\theta(x)-g(\theta)}.$$

Table 1
Common parametric exponential families used for estimating univariate and discrete distributions

Name	Domain $\mathcal{X}$	Measure	$\phi(x)$	$g(\theta)$	Domain Θ
Binomial	$\{0,1\}$	Counting	x	$\log(1+e^{\theta})$	$\mathbb{R}$
Multinomial	$\{1..N\}$	Counting	e_x	$\log\sum_{i=1}^{N} \mathrm{e}^{\theta_i}$	$\mathbb{R}^N$
Exponential	$\mathbb{N}_0^+$	Counting	x	$-\log(1-e^{\theta})$	$(-\infty,0)$
Poisson	$\mathbb{N}_0^+$	$\frac{1}{x!}$	x	e^{θ}	$\mathbb{R}$
Laplace	$[0,\infty)$	Lebesgue	x	$\log\theta$	$(-\infty,0)$
Gaussian	$\mathbb{R}$	Lebesgue	$(x,-\frac{1}{2}x^2)$	$\frac{1}{2}\log 2\pi-\frac{1}{2}\log\theta_2+\frac{1}{2}\frac{\theta_1^2}{\theta_2}$	$\mathbb{R}\times(0,\infty)$
	$\mathbb{R}^n$	Lebesgue	$(x,-\frac{1}{2}xx^{\top})$	$\frac{n}{2}\log 2\pi-\frac{1}{2}\log\|\theta_2\|+\frac{1}{2}\theta_1^{\top}\theta_2^{-1}\theta_1$	$\mathbb{R}^n\times\mathrm{Cone}\,\mathbb{R}^{n^2}$
Inv. Normal	$[0,\infty)$	$x^{-\frac{3}{2}}$	$\left(-x,-\frac{1}{x}\right)$	$\frac{1}{2}\log\pi-2\sqrt{\theta_1\theta_2}-\frac{1}{2}\log\theta_2$	$(0,\infty)^2$
Beta	$[0,1]$	$\frac{1}{x(1-x)}$	$(\log x,\log(1-x))$	$\log\frac{\Gamma(\theta_1)\Gamma(\theta_2)}{\Gamma(\theta_1+\theta_2)}$	$\mathbb{R}^2$
Gamma	$[0,\infty)$	$\frac{1}{x}$	$(\log x,-x)$	$\log\Gamma(\theta_1)-\theta_1\log\theta_2$	$(0,\infty)^2$
Wishart	$\mathrm{Cone}\,\mathbb{R}^{n^2}$	$\|X\|^{-\frac{n+1}{2}}$	$(\log\|x\|,-\frac{1}{2}x)$	$-\theta_1\log\|\theta_2\|+\theta_1 n\log 2+\sum_{i=1}^{n}\log\Gamma\left(\theta_1+\frac{1-i}{2}\right)$	$\mathbb{R}\times\mathrm{Cone}\,\mathbb{R}^{n^2}$
Dirichlet	$\\|x\\|_1=1$ $x_i\geqslant 0$	$\prod_{i=1}^{n}x_i^{-1}$	$(\log x_1,\ldots,\log x_n)$	$\sum_{i=1}^{n}\log\Gamma(\theta_i)-\log\Gamma(\sum_{i=1}^{n}\theta_i)$	$(\mathbb{R}^+)^n$

The notation is geared towards an inner product setting, that is, using $\phi(x)$ and $g(\theta)$ in an explicit form.

When we are concerned with estimating conditional probabilities, the exponential families framework can be extended to conditional densities

$$\mathbb{P}(y|x;\theta)=\mu(y|x)\exp(\langle\theta(.),k(x,y,.)\rangle_{\mathscr{H}}-g(\theta|x)) \tag{5a}$$

and

$$g(\theta|x)=\log\int_{\mathscr{Y}}\exp(\langle\theta(.),k(x,y,.)\rangle_{\mathscr{H}})\,\mathrm{d}\mu(y|x). \tag{5b}$$

$g(\theta|x)$ is commonly referred to as the conditional log-partition function. Both $g(\theta)$ and $g(\theta|x)$ are convex C^{∞} functions in θ and they can be used to compute moments of a distribution

$$\begin{aligned}\partial_\theta g(\theta(.))&=\mathbf{E}_{p(x;\theta)}[k(x,.)]\\ \partial_\theta g(\theta|x)&=\mathbf{E}_{p(x,y;\theta)}[k(x,y)|x]\quad\text{Mean,}\end{aligned} \tag{6}$$

$$\begin{aligned}\partial_\theta^2 g(\theta(.))&=\mathrm{Var}_{p(x;\theta)}[k(x,.)]\\ \partial_\theta^2 g(\theta|x)&=\mathrm{Var}_{p(x,y;\theta)}[k(x,y)|x]\quad\text{Variance.}\end{aligned} \tag{7}$$

We will also assume there exists some prior distribution on parameter θ

$$\mathbb{P}(\theta)=\frac{1}{Z}\exp\left(-\frac{\|\theta\|_{\mathscr{H}}^2}{2\sigma^2}\right),$$

where Z is a normalizing constant.[1] In this case, the posterior density can be written as $\mathbb{P}(\theta|x)\propto\mathbb{P}(x|\theta)\mathbb{P}(\theta)$. In this paper only the estimation of parameter θ using the maximum a posteriori principle (MAP) together with the normal prior will be taken into account.

Note that other priors could have been used such as the conjugate prior. Also other regularisation terms could have been considered by optimizing a penalized likelihood. Relation between the choice of the regularization term and the nature of the estimates is still an open problem.

3.3. Natural parameter space

The exponential family is not defined for any parameter θ but only in a natural parameter space defined as follows:

Definition 4 (*natural parameter space*). The natural parameter space of an exponential family is the set of θ where it is defined, i.e. such that

$$\int_{\mathscr{X}}\exp(\langle\theta(.),k(x,.)\rangle_{\mathscr{H}})\,\mathrm{d}\mu(x)<\infty.$$

The question is now to find out the structure of the natural parameter space. To do so we can define admissible kernels for a given exponential family.

Definition 5 (*admissible kernel*). A kernel k is said to be admissible for a measure μ if

$$\int_{\mathscr{X}}\exp^{k(x,x)^{1/2}}\,\mathrm{d}\mu(x)<\infty.$$

[1] Note that Z need not exist on the entire function space but rather only on the linear subspace of points where $\mathbb{P}(x;\theta)$ is evaluated. The extension to entire function spaces requires tools from functional analysis, namely Radon derivatives with respect to the Gaussian measure imposed by the prior itself.

This allow us to state the following proposition:

Proposition 6. *For any admissible kernel k, the associated natural parameter space includes the unit ball of the associated RKHS.*

Proof. By Cauchy-Schwartz:

$$\int_{\mathscr{X}} \exp(\langle \theta(.), k(x,.)\rangle_{\mathscr{H}})\,\mathrm{d}\mu(x) \leqslant \int_{\mathscr{X}} \exp(\|\theta\|_{\mathscr{H}} \|k(x,.)\|_{\mathscr{H}})\,\mathrm{d}\mu(x)$$

thus, $\|k(x,.)\|^2_{\mathscr{H}} = k(x,x)$ and in the unit ball of the RKHS $\|\theta\|_{\mathscr{H}} \leqslant 1$. □

Some simple examples of nice behavior can be given. If the domain $\mathscr{X}$ is bounded, then any kernel is admissible. For unbounded domain and the Gaussian kernel, there exists some measure μ that makes the kernel admissible. As we will see below, the knowledge of the reference measure is not needed. Only its existence is required.

4. Nonparametric exponential families for learning

4.1. Kernel logistic regression and Gaussian process

Assume we observe some training data (x_i, y_i), $i = 1, n$. The binary classification problem arises when $y_i \in \{-1, +1\}$. In this case we can use the conditional exponential family to model $\mathbb{P}(Y = y|x)$. The estimation of its parameter θ using the maximum a posteriori (MAP) principle aims at minimizing the following cost function:

$$-\log \mathbb{P}(\theta|\text{data}) = -\sum_{i=1}^{n} \langle \theta(.), k(x_i, y_i, .)\rangle_{\mathscr{H}} + g(\theta, x_i) + \langle \theta(.), \theta(.)\rangle_{\mathscr{H}}/2\sigma^2 + C, \qquad (8)$$

where C is some constant term. Note that this can also be seen as a penalized likelihood cost function and thus connected to minimum description length principle and regularized risk minimization.

Proposition 7 (*Feature map for binary classification*). *Without loss of generality the kernel for binary classification can be written as*

$$\langle k(x,y,\cdot), k(x',y',\cdot)\rangle_{\mathscr{H}} = yy'k(x,x'). \qquad (9)$$

Proof. Assume that instead of $k(x,y,\cdot)$ we have $k(x,y,\cdot) + f_0$ where f_0 is a function of x only. In this case $g(\theta)$ is transformed into $g(\theta|x) + f_0(x)$ (and likewise $\langle k(x,y,\cdot), \theta(\cdot)\rangle_{\mathscr{H}}$ into $\theta(x,y) + f_0(x)$). Hence the conditional density remains unchanged. This implies that we can find an offset f_0 such that $\sum_y k(x,y,\cdot) = 0$. The consequence for binary classification is that

$$k(x,1,\cdot) + k(x,-1,\cdot) = 0$$

and therefore

$$k(x,y,\cdot) = yk_0(x,\cdot) \qquad (10)$$

for some k_0, such as $k_0(x,\cdot) = k(x,1,\cdot)$. Taking inner products proves the claim. □

Using the reproducing property Eq. (2) ($\theta(x_i) = \langle \theta(.), k(x_i,.)\rangle_{\mathscr{H}}$) we have

$$g(\theta, x_i) = \log(\exp \theta(x_i) + \exp -\theta(x_i)).$$

Then after some algebra the MAP estimator can be found by minimizing

$$\sum_{i=1}^{n} \log(1 + \exp(-2\theta(x_i)y_i)) + \frac{1}{2\sigma^2}\|\theta\|^2_{\mathscr{H}}.$$

On this minimization problem, the representer theorem (see [10] for more details) gives us

$$\theta(.) = \sum_{i=1}^{n} y_i\alpha_i k(x_i,.).$$

The associated optimization problem can be rewritten in terms of α

$$\min_{\alpha\in\mathbb{R}^n} \sum_{i=1}^{n} \log\left(1 + \exp\left(-2\sum_{j=1}^{n} y_j\alpha_j k(x_i,x_j)\right)\right) + \frac{1}{2\sigma^2}\sum_{i=1}^{n}\sum_{j=1}^{n} y_j y_i \alpha_i \alpha_j k(x_i,x_j).$$

It is non-linear and can be solved using Newton method. The connection is made with the kernel logistic regression since in our framework we have

$$\log \frac{\mathbb{P}(Y = 1|x)}{\mathbb{P}(Y = -1|x)} = 2\sum_{i=1}^{n} y_i\alpha_i k(x_i,x)$$

and thus the decision of classifying a new data x only depends on the sign of the kernel term through the following decision rule:

$$\text{decide class 1} \quad \text{if} \sum_{i=1}^{n} y_i\alpha_i k(x_i,x) > b,$$

where b is a decision threshold to be set according to a given risk. This is closely related with kernel-based receiver as treated in [1].

Note that the multiclass problem can be solved by using the same kind of derivations assuming that $k(x_i,y_i,x,y) = k(x_i,x)\delta_{y_iy}$.

4.2. Two-class support vector machines

We now define the margin of a classifier (binary or not) as the most pessimistic log-likelihood ratio for classification. That is, we define

$$r(x,y,\theta) := \log \mathbb{P}(y|x,\theta) - \max_{\tilde{y}\neq y} \log \mathbb{P}(\tilde{y}|x,\theta). \qquad (11)$$

Clearly, whenever $r(x,y,\theta) > 0$, we classify correctly. Moreover, its magnitude gives the logarithm of the ratio between the correct class and the most dominant incorrect class. In other words, large values of r imply that we are very confident about classifying x as y. In terms of classification accuracy this is a more useful proposition than the log-likelihood, as the latter does not provide necessary and sufficient conditions for correct classification.

It is easy to see that, for binary classification, this yields

$$r(x,y,\theta) = \langle \theta, yk(x,\cdot)\rangle - \langle \theta, -yk(x,\cdot)\rangle = 2y\theta(x) \tag{12}$$

which is the standard definition of the margin. Instead of the MAP estimate we optimize with respect to a reference margin ρ, i.e. we minimize in the soft case margin $1/2\|\theta\|^2 + C\sum_{i=1}^n \xi_i$ with $r_i \geqslant \rho - \xi_i$ and $\xi_i \geqslant 0$ leading to

$$\min_\theta \sum_{i=1}^n \max(0, \rho - r(x_i, y_i, \theta)) + \frac{1}{\sigma^2}\|\theta\|^2_{\mathscr{H}}. \tag{13}$$

Together with the exponential family model, the minimization of this criterion leads to the maximum margin classifier. Here again this can be easily generalized to the multiclass problem.

4.3. One-class support vector machines

The one-class SVM algorithm has been designed to estimate some quantile from sample data. This is closely related but simpler than estimating the whole density. Recently it has been proven that it is a consistent estimate of the density level set [12] and of the minimum measure set [5] (i.e. roughly, for a given $p_0 \in [0,1]$, the subset $C \subseteq \mathscr{X}$ minimizing some volume and such that $\mathbb{P}(C) = p_0$).

One-class SVMs are also more relevant when the target application is novelty detection. As a matter of fact, any point not belonging to the support of a density can be seen as a novel one.

Back with our exponential family model for $\mathbb{P}(x)$, a robust approximation of maximum a posteriori (MAP) estimator for θ is the one maximizing

$$\max_{\theta\in\mathscr{H}} \prod_{i=1}^n \min\left(\frac{\mathbb{P}_0(x_i|\theta)}{p_0}, 1\right)\mathbb{P}(\theta)$$

with $p_0 = \exp(\rho - g(\theta))$. After some tedious algebra, this problem can be rewritten as

$$\min_{\theta\in\mathscr{H}} \sum_{i=1}^n \max(\rho - \langle \theta(.), k(x_i,.)\rangle_{\mathscr{H}}, 0) + \frac{1}{2\sigma^2}\|\theta\|^2_{\mathscr{H}}. \tag{14}$$

On this problem again the representer theorem gives us the existence of some coefficient α_i such that

$$\theta(.) = \sum_{i=1}^n \alpha_i k(x_i,.)$$

and thus the estimator has the following form:

$$\widehat{\mathbb{P}}(x) = \mu(x)\exp\left(\sum_{i=1}^n \alpha_i k(x_i,.) - b\right),$$

where coefficients α are determined by solving the one-class SVM problem (14). Parameter b represents the value of the log partition function and thus the normalization factor. It can be hard to compute it but it is possible to do without it in our applications. The nature and the choice of the reference measure μ depends on the application (see [7] for a non parametric estimation of the reference measure together with the exponential family).

Here again the one-class SVM algorithm can be derived using the exponential family on an RKHS and a relevant cost function to be minimized.

4.4. Regression

It is possible to see the problem as a generalization of the classification case to continuous y. But in this case, a generalized version of the representer theorem shows that parameters α are no longer scalar but functions, leading to intractable optimization problems. Some additional hypotheses have to be made about the nature of the unknown distribution. One way to do it is to use the conditional gaussian representation with its natural parameters

$$\mathbb{P}(y|x) = \exp(y\theta_1(x) + y^2\theta_2(x) - g(\theta_1(x), \theta_2(x)))$$

with $\theta_1(x) = m(x)/\sigma^2(x)$ and $\theta_2(x) = -1/2\sigma^2(x)$ where $m(x)$ is the conditional expectation of y given x and $\sigma^2(x)$ its conditional variance. The associated kernel can be written as follows:

$$k(x_i, y_i, x, y) = k_1(x_i, x)yy_i + k_2(x_i, x)y^2y_i^2,$$

where k_1 and k_2 are two positive kernels. In this case the application of the represented theorem gives a heteroscedastic gaussian process (with non-constant variance) as the model of the data, associated with a convex optimization problem (see [8] for details). Convexity is the main interest of the framework. Such a model directly using functions $m(x)$ and $\sigma(x)$ as unknown have been proposed before. But the associated optimization problem is non convex while the use of the exponential family representation leads to convex optimization problem.

5. Application to novelty detection

Let $X_i, i = 1, 2, \ldots, 2t$ be a sequence of random variables distributed according to some distribution $\mathbb{P}_i$. We are interested in finding whether or not a change has occurred at time t. To begin with a simple framework we will assume the sequence to be stationary from 1 to t and from $t+1$ to $2t$, i.e. there exist some distributions $\mathbb{P}_0$ and $\mathbb{P}_1$ such that $P_i = P_0$ for $i \in [1,t]$ and $P_i = P_1$ for $i \in [t+1, 2t]$. The question we are addressing can be seen as determining if $\mathbb{P}_0 = \mathbb{P}_1$ (no change has occurred) or else $\mathbb{P}_0 \neq \mathbb{P}_1$ (some change have occurred). This can be restated as the following statistical test:

$$\begin{cases} \mathscr{H}_0\colon\ \mathbb{P}_0 = \mathbb{P}_1, \\ \mathscr{H}_1\colon\ \mathbb{P}_0 \neq \mathbb{P}_1. \end{cases}$$

In this case the likelihood ratio is the following:

$$\Lambda_t(x_1,\dots,x_{2t}) = \frac{\prod_{i=1}^{t}\mathbb{P}_0(x_i)\prod_{i=t+1}^{2t}\mathbb{P}_1(x_i)}{\prod_{i=1}^{2t}\mathbb{P}_0(x_i)} = \prod_{i=t+1}^{2t}\frac{\mathbb{P}_1(x_i)}{\mathbb{P}_0(x_i)}$$

since both densities are unknown the generalized likelihood ratio (GLR) has to be used

$$\Lambda(x_1,\dots,x_{2t}) = \prod_{i=t+1}^{2t}\frac{\widehat{\mathbb{P}}_1(x_i)}{\widehat{\mathbb{P}}_0(x_i)},$$

where $\widehat{\mathbb{P}}_0$ and $\widehat{\mathbb{P}}_1$ are the maximum likelihood estimates of the densities.

Because we want our detection method to be universal, we want it to work for any possible density. Thus some approximations have to be done to clarify our framework. First, assuming both densities $\mathbb{P}_0$ and $\mathbb{P}_1$ belong to the generalized exponential family, there exists a reproducing kernel Hilbert space $\mathscr{H}$ embedded with the dot product $\langle .,. \rangle_{\mathscr{H}}$ and with a reproducing kernel k such that

$$\mathbb{P}_0(x) = \mu(x)\exp\langle\theta_0(.), k(x,.)\rangle_{\mathscr{H}} - g(\theta_0) \tag{15a}$$

and

$$\mathbb{P}_1(x) = \mu(x)\exp\langle\theta_1(.), k(x,.)\rangle_{\mathscr{H}} - g(\theta_1), \tag{15b}$$

where $g(\theta)$ is the so-called log-partition function and μ some reference measure. Note that this factorisation of the unknown density in two terms (μ and the exponential) makes it possible to focus on a relevant quantity for the task to be performed. This quantity (the exponential part) converges towards the exponential of the minimum measure set and it is sufficient to make relevant decisions regarding the detection of abrupt changes. The analysis of this factorisation has to be investigated in more details.

Second hypothesis, the functional parameter θ_0 and θ_1 of these densities will be estimated on the data of respectively the first and the second half of the sample by using the one-class SVM algorithm. By doing so we are following our initial assumption that before time t we know the distribution is constant and equal to some $\mathbb{P}_0$. The one-class SVM algorithm provides us with a good estimator of this density. The situation of $\widehat{\mathbb{P}}_1(x)$ is more simple. It is clearly a robust approximation of the maximum likelihood estimator. Using one-class SVM algorithm and the exponential family model, both estimates can be written as

$$\widehat{\mathbb{P}}_0(x) = \mu(x)\exp\left(\sum_{i=1}^{t}\alpha_i^{(0)}k(x,x_i) - g(\theta_0)\right), \tag{16a}$$

$$\widehat{\mathbb{P}}_1(x) = \mu(x)\exp\left(\sum_{i=t+1}^{2t}\alpha_i^{(1)}k(x,x_i) - g(\theta_1)\right), \tag{16b}$$

where $\alpha_i^{(0)}$ is determined by solving the one-class SVM problem on the first half of the data (x_1 to x_t). while $\alpha_i^{(1)}$ is given by solving the one-class SVM problem on the second half of the data (x_{t+1} to x_{2t}). Using these hypotheses, the generalized likelihood ratio test is approximated as follows:

$$\begin{aligned}
&\Lambda(x_1,\dots,x_{2t}) > s\\
&\Leftrightarrow \prod_{j=t+1}^{2t}\frac{\exp\sum_{i=t+1}^{2t}\alpha_i^{(1)}k(x_j,x_i) - g(\theta_1)}{\exp\sum_{i=1}^{t}\alpha_i^{(0)}k(x_j,x_i) - g(\theta_0)} > s\\
&\Leftrightarrow \sum_{j=t+1}^{2t}\left(\sum_{i=1}^{t}\alpha_i^{(0)}k(x_j,x_i) - \sum_{i=t+1}^{2t}\alpha_i^{(1)}k(x_j,x_i)\right) < s',
\end{aligned}$$

where s' is a threshold to be fixed to have a given risk of the first kind a such that

$$\mathbb{P}_0\left(\sum_{j=t+1}^{2t}\left(\sum_{i=1}^{t}\alpha_i^{(0)}k(x_j,x_i) - \sum_{i=t+1}^{2t}\alpha_i^{(1)}k(x_j,x_i)\right) < s'\right) = a.$$

It turns out that the variation of $\sum_{i=t+1}^{2t}\alpha_i^{(1)}k(x_j,x_i)$ is very small in comparison to that of $\sum_{i=1}^{t}\alpha_i^{(0)}k(x_j,x_i)$. Thus $\widehat{\mathbb{P}}_1(x)$ can be assumed to be constant, simplifying computations. In this case the test can be performed by only considering

$$\sum_{j=t+1}^{2t}\left(\sum_{i=1}^{t}\alpha_i^{(0)}k(x_j,x_i)\right) < s. \tag{17}$$

This is exactly the novelty detection algorithm as proposed by Schölkopf et al. [9] and adapted for change detection in [4] and continued in [6]. To sum up, we showed how to derive the heuristic described Eq. (17) as a statistical test approximating a generalized likelihood ratio test, optimal under some condition in the Neyman Pearson framework.

This framework can be easily extended to do sequential hypothesis testing through a Wald sequential probability ratio test using almost the same likelihood ratio. The associated decision function is the following:

$$\begin{array}{lll}
\text{if} & \sum_{j=t+1}^{t+\tau}\left(\sum_{i=1}^{t}\alpha_i^{(0)}k(x_j,x_i)\right) - \tau s > A & \text{decide}\mathscr{H}_0\\
\text{elseif} & \sum_{j=t+1}^{t+\tau}\left(\sum_{i=1}^{t}\alpha_i^{(0)}k(x_j,x_i)\right) - \tau s < B & \text{decide}\mathscr{H}_1\\
\text{elseif} & \text{pick one more sample,} &
\end{array}$$

where A and B are two thresholds to be set according to some prefixed error rates. The goal here is to minimize detection delay.

Note that for practical use the distribution of the test statistic under the null-hypothesis is required to define the threshold levels for a given level of significance. To do so, resampling techniques can be used.

6. Conclusion

In this paper we have illustrated how powerful the link made between kernel algorithms, Reproducing Kernel Hilbert Space and the exponential family is. A lot of learning algorithms can be revisited using this framework. Here we have discussed the logistic kernel regression, the SVM, the gaussian process for regression and the novelty detection using the one-class SVM. This framework is applicable to many different cases and other derivations are possible: exponential family in a RKHS can be used to recover sequence annotation (via Conditional Random Fields) or boosting, to name just a few. The exponential family framework is powerful because it makes it possible to connect learning algorithm with usual statistical tools such as posterior densities and likelihood ratio, and to do so with almost no loss of generality. These links between statistics and learning have been detailed in the case of novelty detection restated as a quasi optimal statistical test based on a robust approximation of the generalized likelihood. Further works on this field regard the application of sequential analysis tools such as the CUSUM algorithm for real-time novelty detection minimizing the expectation of the detection delay.

Acknowledgements

National ICT Australia is funded through the Australian Government's *Backing Australia's Ability* initiative, in part through the Australian Research Council. This work was supported by grants of the ARC and by the IST Programme of the European Community, under the Pascal Network of Excellence, IST-2002-506778.

References

[1] F. Abdallah, C. Richard, R. Lengellé, An improved training algorithm for nonlinear kernel discriminants, IEEE Trans. Signal Process. 52 (2004) 2798–2806.

[2] N. Aronszajn, La théorie générale des noyaux réproduisants et ses applications, Proc. Cambridge Philos. Soc. 39 (1944) 133–153.

[3] S. Canu, X. Mary, A. Rakotomamonjy, Advances in Learning Theory: Methods, Models and Applications NATO Science Series III: Computer and Systems Sciences, Functional Learning Through Kernel, vol. 190, IOS Press, Amsterdam, 2003, pp. 89–110.

[4] F. Desobry, M. Davy, Support vector-based online detection of abrupt changes, in: IEEE ICASSP, 2003.

[5] F. Desobry, M. Davy, S. Canu, Kernel methods for minimum measure sets estimation and application, Technical report, Cambridge University, 2005.

[6] F. Desobry, M. Davy, C. Doncarli, An online kernel change detection algorithm, IEEE Trans. Signal Process. 54 (2005) 115–256.

[7] A. Gous, Adaptive estimation of distributions using exponential subfamilies, J. Comput. Graphical Statist. 7 (3) (1998) 388–396.

[8] Q.V. Le, A.J. Smola, S. Canu, Heteroscedastic gaussian process regression, in: ICML, 2005.

[9] B. Schölkopf, J. Platt, J. Shawe-Taylor, A.J. Smola, R.C. Williamson, Estimating the support of a high-dimensional distribution, Neural Comput. 13 (7) (2001).

[10] B. Schölkopf, A.J. Smola, Learning with Kernels, MIT Press, Cambridge, MA, 2001.

[11] L. Schwartz, Sous espaces hilbertiens d'espaces vectoriels topologiques et noyaux associés, J. Anal. Math. (1964) 115–256.

[12] R. Vert, J.-P. Vert, Consistency and convergence rate of one-class svm and related algorithms, Technical report, LRI, 2005.

[13] G. Wahba, Spline Models for Observational Data, Series in Applied Mathematics, vol. 59, SIAM, Philadelphia, PA, 1990.

Pr. Canu received his Ph.D. degree in System Command from the Comiegne University of Technology in 1986. He joined the faculty department of Computer Science at the Compiegne University of Technology in 1987. He received his French habilitation degree from the Paris 6 University. In 1997, he joined the Rouen Applied Sciences National Institute (INSA) as a Full Professor, where he created the Information Engineering Department. He has been the Director of this department until 2001 where he was named the Director of the computer service team. In 2003, he joined for one sabbatical year the machine learning group at the ANU/NICTA (Canberra).

Now he is responsible of the machine learning group of the PSI laboratory.

In the last 5 years, he has published approximately 30 papers in refereed conference proceedings or journals in the areas of theory, applications and forecasting using kernel machines learning algorithm and other flexible regression methods. His research interests are kernels and frames machines, machine learning, pattern classification, and learing for context aware applications. He has been involved in three European research Esprit projects (Neufodi, EMS and EM2S) and the network of excellence Pascal.

Dr Smola received his Doctoral Degree in Computer Science in 1998 at the University of Technology, Berlin. Until 1999, he was a researcher at the IDA Group of the GMD Institute for Software Engineering and Computer Architecture in Berlin (now part of the Fraunhofer Gesellschaft). After that he joined the Australian National University and worked from 2002 to 2004 as a leader of the machine learning group. Since 2004, he has been Program Leader of the Statistical Machine Learning program of National ICT Australia. Dr. Smola obtained his Ph.D. in Computer Science at the University of Technology in Berlin in 1998.

Dr. Smola research interests are nonparametric methods for estimation, such as kernels (Support Vector Machines and Gaussian Processes); inference on discrete objects (graphs, strings, automata); structured estimation, optimisation, and numerical analysis (Interior Point methods, matrix factorisation); and learning theory (uniform convergence bounds, design of priors, etc.). Dr. Smola has an adjunct position at the Research School of Information Sciences and Engineering (RSISE) at the Australian National University.