Kernel-Based Learning from Both Qualitative and Quantitative Labels: Application to Prostate Cancer Diagnosis Based on Multiparametric MR Imaging

Émilie Niaf, Rémi Flamary, Olivier Rouvière, Carole Lartizien, and Stéphane Canu

***Abstract*— Building an accurate training database is challenging in supervised classification. For instance, in medical imaging, radiologists often delineate malignant and benign tissues without access to the histological ground truth, leading to uncertain data sets. This paper addresses the pattern classification problem arising when available target data include some uncertainty information. Target data considered here are both qualitative (a class label) or quantitative (an estimation of the posterior probability). In this context, usual discriminative methods, such as the support vector machine (SVM), fail either to learn a robust classifier or to predict accurate probability estimates. We generalize the regular SVM by introducing a new formulation of the learning problem to take into account class labels as well as class probability estimates. This original reformulation into a probabilistic SVM (P-SVM) can be efficiently solved by adapting existing flexible SVM solvers. Furthermore, this framework allows deriving a unique learned prediction function for both decision and posterior probability estimation providing qualitative and quantitative predictions. The method is first tested on synthetic data sets to evaluate its properties as compared with the classical SVM and fuzzy-SVM. It is then evaluated on a clinical data set of multiparametric prostate magnetic resonance images to assess its performances in discriminating benign from malignant tissues. P-SVM is shown to outperform classical SVM as well as the fuzzy-SVM in terms of probability predictions and classification performances, and demonstrates its potential for the design of an efficient computer-aided decision system for prostate cancer diagnosis based on multiparametric magnetic resonance (MR) imaging.**

***Index Terms*— Computer-assisted decision system, machine learning, support vector machines, maximal margin algorithm, uncertain labels, medical imaging, multiparametric magnetic resonance imaging.**

Manuscript received April 25, 2013; revised October 2, 2013; accepted December 2, 2013. Date of publication December 20, 2013; date of current version January 20, 2014. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Jong C. Ye.

É. Niaf and C. Lartizien are with the Université de Lyon; CREATIS; CNRS UMR5220; INSERM U1044; INSA-Lyon; Université Lyon 1; Lyon, France (e-mail: emilie.niaf@creatis.insa-lyon.fr; carole.lartizien@creatis.insa-lyon.fr).

R. Flamary is with Laboratoire Lagrange, UMR 7293, Université de Nice Sophia-Antipolis, CNRS, Observatoire de la Côte d'Azur, Nice 06108, France (e-mail: remi.flamary@unice.fr).

O. Rouvière is with INSERM, U1032, LabTau, Lyon F-69003, France, with the Université de Lyon, Lyon F-69003, France, with the Université Lyon 1, Lyon, F-69003, France, and with Hospices Civils de Lyon, Department of Urinary and Vascular Imaging, Hôpital Edouard Herriot, Lyon, F-69003, France (e-mail: olivier.rouviere@chu-lyon.fr).

S. Canu is with the LITIS, EA 4108, INSA de Rouen and Université de Normandie, Saint-Etienne-du-Rouvray 76801, France (e-mail: stephane.canu@insa-rouen.fr).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TIP.2013.2295759

I. Introduction

IMAGE classification remains a major challenge to the computer vision community in various application domains, including biomedical imaging [1], web image and video search [2], industrial visual inspection, biometry, and remote sensing [3], [4] to name but a few. Image classification aims at assigning to each pixel, or to each region of interest (ROI) extracted from the image, a label associated with a class of object that can possibly be present in the analyzed scene. More particularly, supervised pattern recognition approaches consist in learning a classification model based on a training dataset for which the class belonging, usually referred to as the ground truth, is known and can thus be used to infer discriminative rules. However, in most of real problems, training datasets are doomed to contain classification errors or uncertainties even when data is labeled by experts. For instance, in medical imaging, radiologists often have to outline what they think are malignant tissues over medical images without access to the reference histopathologic information (E.g. for prostate cancer imaging, [5]–[8]). It is thus a common use in clinical practice to affect a malignancy suspicion score to the outlined targets, using different scales such as the Likert scale [9]. These scores, however, are usually converted into binary class variables prior to the classification step by setting an arbitrary threshold. This conversion has two negative effects on the data: first, uncertainty information that may be of interest for probability estimation is lost, second, classification noise (badly labeled examples) is added to the data leading to non robust classifiers [10].

We propose to deal with these learning samples uncertainties by directly using probabilistic labels in the learning stage so as to avoid discarding uncertain data while constructing a robust classifier.

This study extends the widely used support vector machine (SVM) two-class classification problem [11]. This supervised classification algorithm is based on the maximum margin principle. It has good generalization ability, is very effective in high dimensional feature space and the learning phase associated with the minimization of a convex cost function guarantees the uniqueness of the solution. In this framework, relatively little work has discussed means for considering uncertain or probabilistic dataset as learning samples. Some authors have proposed to mix uncertainties together with classification through a weighting scheme [11], [12]. Other studies

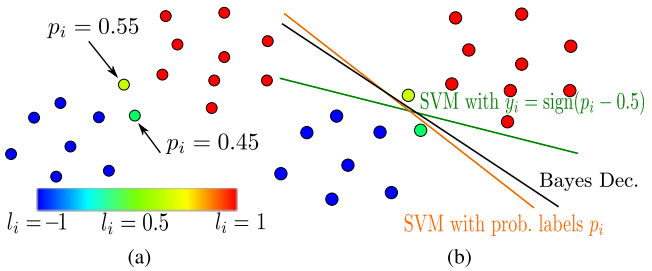

Fig. 1. Examples of discrimination functions learned on probabilistic datasets via SVM and P-SVM. (a) training dataset spatial distribution along with the associated class probability $p_i = \mathbb{P}(l_i = 1|X = \mathbf{x}_i)$. (b) learned functions for the regular SVM (in green) and the P-SVM (in orange) while considering binary and probabilistic class labels l_i respectively. The regular SVM separating hyperplane is impacted by the yellow ($p_i = 0.55$) and green ($p_i = 0.45$) points which are considered as '+1' and '−1' examples respectively; the P-SVM algorithm accounts for the low class probability of these two examples ($p_i \simeq 0.5$), thus resulting in a more robust discriminative hyperplan, close to the Bayes decision (in black).

focused on learning probabilities after discrimination using an additional mapping algorithm [13], [14]. These mappings are indeed performed *a posteriori* on a discriminant prediction function. We believe that including them directly in a unique learning process will improve the classification performances as well as the posterior probability prediction.

Our main contribution is a SVM inspired formulation of the learning problem allowing to take into account class labels through a hinge loss cost function as well as class probability estimates using ε-insensitive cost function together with a minimum norm (maximum margin) objective. This formulation, referred to as Probabilistic SVM (P-SVM) in the following, shows a dual form leading to a quadratic problem similar to the classical SVM formulation, and allows the use of a representer theorem and associated kernel. It hence can be efficiently solved by adapting existing flexible SVM solvers. As illustrated in Fig. 1, training data points with uncertain class probability p_i ($0.4 < p_i < 0.6$ for instance) can have a dramatic impact on the decision function (green line) if used as certain after thresholding (green and yellow samples assigned to class '+1' and '−1' respectively). A classifier that takes this uncertainty into account will allow the discriminant hyperplane (orange line) to pass near those uncertain samples ($p_i \simeq 0.5$), thus leading to a decision closer to the Bayes decision (black line).

This paper details the basis of the P-SVM formalism that was first introduced in our preliminary work [15], and generalize the optimization problem by introducing new parameters; we also propose an extensive evaluation study against state-of-the-art methods on synthetic datasets and evaluate the relevance of the proposed algorithm on a clinical dataset, using radiologists' scores as probabilistic inputs. The paper proceeds as follows. In section II, we briefly review some basics on SVM and present a state-of-the-art of related studies focusing on posterior class probability prediction and accounting for data uncertainty within this framework. In sections III, we define the P-SVM formalism, our contribution to handle both binary and probabilistic labels simultaneously, and discuss the associated optimization problem. In section IV, we compare the performances of the P-SVM to two state-of-the-art methods over different simulated datasets. Finally, in section V, we demonstrate the potential of this method in the clinical context of prostate cancer diagnosis based on multiparametric MR (mpMR) images.

II. Background

In this section we shortly review some basics on support vector machine as well as some previous works on incorporating class probability or uncertainty in this framework.

A. Support Vector Machines for Classification

Suppose that we are given a training dataset of n samples $\{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)\} \subset \mathcal{X} \times \mathcal{Y}$, where $\mathcal{X}$ denotes the feature space (in the following practical examples, $\mathcal{X} = \mathbb{R}^d$ where d is the number of features per sample) and $\mathcal{Y}$ represents the two-class labelling, $\mathcal{Y} = \{-1, +1\}$. The SVM, introduced by Vapnik [16], aims at constructing a separating hyperplan, of the form:

$$\{\mathbf{x} \in \mathcal{X} | \mathbf{w}^\top \mathbf{x} + b = 0\}, \tag{1}$$

maximizing the margin between the data of the two classes. The associated pattern recognition problem is defined as:

$$\begin{cases} \min\limits_{\mathbf{w}\in\mathbb{R}^d, b, \xi_i \in \mathbb{R}} \dfrac{1}{2}\|\mathbf{w}\|^2 + C\sum\limits_{i=1}^{n} \xi_i, & \text{(2a)} \\ \text{subject to} \\ \quad y_i(\mathbf{w}^\top \mathbf{x}_i + b) \geq 1 - \xi_i, \qquad i = 1, \ldots, n & \text{(2b)} \\ \quad 0 \leq \xi_i, \qquad\qquad\qquad\quad i = 1, \ldots, n & \text{(2c)} \end{cases}$$

which combines a minimum norm (maximum margin) objective function (2a) and good classification constraints (2b). Slack variables $\xi_i \geq 0$ (2c) correspond to the distance to the margin of possibly misclassified samples $\mathbf{x}_i$. Parameter C (2a) is the associated cost coefficient that weights the classification error.

This optimization problem can be expressed in its dual form, resulting in the following expression:

$$\begin{cases} \max\limits_{\alpha\in\mathbb{R}} -\dfrac{1}{2}\sum\limits_{i=1}^{n}\sum\limits_{j=1}^{n} \alpha_i \alpha_j y_i y_j \mathbf{x}_i^\top \mathbf{x}_j + \sum\limits_{i=1}^{n} \alpha_i \\ \text{subject to} \\ \quad 0 \leq \alpha_i \leq C \qquad i = 1, \ldots, n \\ \quad \sum\limits_{i=1}^{n} \alpha_i y_i = 0 \end{cases} \tag{3}$$

where coefficients α_i are the Lagrange multipliers.

The classification of a test vector $\mathbf{x} \in \mathcal{X}$ is then performed by computing $sign(f(\mathbf{x}))$ where $f(\mathbf{x})$ represents the signed distance to the margin and can be expressed as:

$$f(\mathbf{x}) = \mathbf{w}^\top \mathbf{x} + b = \sum_{i=1\ldots n} \alpha_i y_i \mathbf{x}_i^\top \mathbf{x} + b. \tag{4}$$

Note that examples located on or inside the constructed margins, for which $\alpha \neq 0$, are called "support vectors".

When the classification problem is nonlinear, SVM can be easily adapted thanks to the "kernel trick". The basic idea is

to introduce a nonlinear mapping function $\phi : \mathbf{x} \longrightarrow \phi(\mathbf{x})$ that maps the data to a higher dimensional space $\mathcal{H}$, where the data is linearly separable. Then, expression (4) becomes:

$$f(\mathbf{x}) = \sum_{i=1\ldots n} \alpha_i y_i \langle \phi(\mathbf{x}_i), \phi(\mathbf{x}) \rangle_{\mathcal{H}} + b. \quad (5)$$

Note that when the problem is solved in its dual form, only the scalar product $\langle \phi(\mathbf{x}_i), \phi(\mathbf{x}_j) \rangle_{\mathcal{H}}$ has to be computed. In practice, we do not need to define explicitly the mapping function ϕ but only the kernel function defined as $k(\mathbf{x}_i, \mathbf{x}_j) = \langle \phi(\mathbf{x}_i), \phi(\mathbf{x}_j) \rangle_{\mathcal{H}}$. The kernel k must be a positive definite function satisfying Mercer's condition and $\mathcal{H}$ is the associated Reproducing Kernel Hilbert Space (RKHS). This is particularly interesting as it allows the use of closed form kernel such as the Gaussian kernel. When using kernels the prediction function (5) can be rewritten as:

$$f(\mathbf{x}) = \sum_{i=1\ldots n} \alpha_i y_i k(\mathbf{x}_i, \mathbf{x}) + b. \quad (6)$$

B. Related Works

1) From Discrimination to Probability Estimation: Classical SVM aims at constructing a discriminative uncalibrated function f predicting a signed distance to the margin, and its associated decision function $sign(f)$. Nevertheless, in many applications, it is desirable to get an estimate of the posterior probabilities for each class in addition to the classification. This can be useful indeed to evaluate the confidence associated to the prediction. Given a measurement $\mathbf{x}$ the goal is to estimate:

$$\mathbb{P}(\text{class}|\text{input}) = \mathbb{P}(Y = y_i | X = \mathbf{x}) = p, \text{ where } y_i = \pm 1. \quad (7)$$

For a recent survey on this issue see for instance [17] with included references. We just give some details on the methods related with our work.

The approach proposed by Platt [13] to estimate posterior class probabilities using SVM is the most popular method. It is a parametric approach to retrospectively transform the SVM scores f to probabilities by using a sigmoid approximation function $\varphi : \mathbb{R} \mapsto [0, 1]$ such that:

$$\mathbb{P}(Y = +1 | \mathbf{X} = \mathbf{x}) = \varphi_A(f(\mathbf{x})) = \frac{1}{1 + \exp(-Af(\mathbf{x}))}, \quad (8)$$

where parameter A is obtained by minimizing the negative log-likelihood function (cross-entropy) on a training or validation set (gradient descent). Supplemental offset parameter B can also be added in the exponential term and adjusted. But this approach remains heuristic.

To provide a formal framework, Grandvalet et al. [18], have shown that an interval-valued mapping from scores to probabilities has to be used, thus providing a set of probabilities compatible with each SVM score:

$$\varphi(f(\mathbf{x})) - \epsilon^-(\mathbf{x}) \leq \mathbb{P}(Y = +1 | X = \mathbf{x}) \leq \varphi(f(\mathbf{x})) + \epsilon^+(\mathbf{x}) \quad (9)$$

where φ is the sigmoid function and ϵ^+ and ϵ^- define the tolerated interval. This interval mapping allows a better understanding of the links between likelihood maximization and SVM and has shown interest particularly on unbalanced datasets.

Yet, this approach has been introduced as a post processing to SVM outputs. We'll see in our work how to include it right from the beginning into the SVM optimization framework to deal with both known labels and probabilities.

2) Incorporating Labelling Uncertainty Into the SVM Training Step: Lin and Wang [12] were the first to define a *Fuzzy SVM* (F-SVM), ie. a binary classifier capable of integrating labelling uncertainty into the learning phase. Given a labeled training data set of points $(\mathbf{x}_i, y_i)_{i=1\ldots n}$ and the associated fuzzy membership measures $0 \leq m_i \leq 1, i = 1 \ldots n$, they proposed to rewrite the optimization problem as:

$$\min_{f \in \mathcal{H}, \xi \in \mathbb{R}} \frac{1}{2} ||f||^2 + C \sum_i m_i \xi_i, \quad (10)$$

with the same constraints as in equation (2), thus ponderating the misclassification cost with class uncertainty. Note that [3] extended the previous approach by using both m_i and $1 - m_i$ in equation (10) for both $+1$ and -1 classes. A very similar weighting procedure, taking advantage of the probabilistic labelling, has been proposed, as an exercise, by Scholkopf and Smola [11] (p. 223), with $m_i = |2p_i - 1|$.
This approach allows ponderating the influence of uncertain examples on the margin construction, but they do not aim at constructing a soft probabilistic output even less a probabilistic prediction function.

III. Probabilistic Problem Formulation

We present a new formulation derived from the classical SVM two-class classification problem, which allows accounting for uncertain labels during the training step while constructing an accurate probabilistic discrimination function. This approach is referred to as P-SVM in the following where P stands for "probabilistic". We introduce the problem in a linear context, but we can easily extend the formulation to non-linearly separable datasets using kernels.
Let $\mathcal{X}$ be a feature space. We define $(\mathbf{x}_i, l_i)_{i=1\ldots m}$ the training dataset of input vectors $(\mathbf{x}_i)_{i=1\ldots m} \in \mathcal{X}$ along with their corresponding labels $(l_i)_{i=1\ldots m}$, the latter of which being:

- class labels: $l_i = y_i \in \{-1, +1\}$ with $i = 1 \ldots n$ for the n training samples that are assigned to any of the two classes with certainty,
- real values: $l_i = p_i \in [0, 1]$ with $i = n + 1 \ldots m$ for the $m - n$ training samples for which the class labelling is more ambiguous.

Probability p_i, associated to point $\mathbf{x}_i$, is an estimated conditional probability for class +1: $p_i = p(\mathbf{x}_i) = \mathbb{P}(Y_i = 1 \mid X_i = \mathbf{x}_i)$.

We aim at constructing the optimal maximal margin hyperplane

$$\{\mathbf{x} \in \mathcal{X} | f(\mathbf{x}) = \mathbf{w}^\top \mathbf{x} + b = 0\}$$

so as to efficiently predict:

- class membership for binary labelled data $(\mathbf{x}_i, y_i)_{i=1\ldots n}$ (in classification),
- conditional class probability for uncertain labelled data $(\mathbf{x}_i, p_i)_{i=n+1\ldots m}$ (in regression).

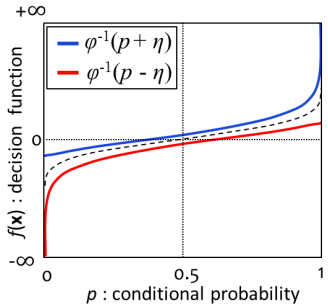

Fig. 2. Localisation constraints representation depending on p. Localisation constraints for prediction $f(\mathbf{x})$ are defined by boundaries z^+ et z^-. They aim at maintaining predictions between defined limits depending on p. The nearest to 0 or 1 (up to minimum distance η) label p is, the softest the localisation constraint on $f(\mathbf{x})$ is ($\rightarrow \infty$).

Let η be an estimate of the uncertainty in the probabilistic labelling. To avoid dealing with undefined cases, we constrain labels $\{p_i\}_{i=n+1\ldots m}$ of samples $\{\mathbf{x}_i\}_{i=n+1\ldots m}$, to belong to $[\eta, 1\text{-}\eta]$, even if it means re-labelling data such that:

$$l_i = \begin{cases} y_i = -1 \text{ if } p_i - \eta \leq 0 \\ y_i = +1 \text{ if } p_i + \eta \geq 1 \\ p_i \text{ otherwise.} \end{cases}$$

Let $\mathbf{x}_i$ be a sample of conditional class probability p_i. Following Platt's formulation, we are looking for probability predictions of the form $\varphi(f(\mathbf{x}))$, where φ is the sigmoid function (see (8)).
This additional regression problem on uncertain examples consists in finding optimal f such that:

$$\begin{aligned} &\mid \varphi(f(\mathbf{x}_i)) - p_i \mid < \eta, \quad \text{for} \quad i = n+1 \ldots m \\ &\text{where} \quad \varphi(z) = \frac{1}{1+e^{-Az}}. \end{aligned} \tag{11}$$

This is aimed at constraining the posterior class probability prediction for point $\mathbf{x}_i$ to remain within distance η of $\varphi(f(\mathbf{x}_i))$, where parameter η represents the labelling precision [13], [18].

Condition (11) can be equivalently rewritten:

$$\begin{aligned} & p_i - \eta \leq \varphi(f(\mathbf{x}_i)) \leq p_i + \eta, \\ \Longleftrightarrow \quad & a.z_i^- \leq \quad f(\mathbf{x}_i) \quad \leq a.z_i^+, \end{aligned} \tag{12}$$

with:

$$z_i^{\pm} = \frac{1}{a}\varphi^{-1}(p_i \pm \eta) = \ln\left(\frac{1}{p_i \pm \eta} - 1\right) \text{ and } a = -\frac{1}{A}. \tag{13}$$

Fig. 2 represents these probability prediction constraints.

A reasonable hypothesis following inequality (12) is to consider that the boundaries of the tube defining probability prediction constraints correspond to extreme cases where $f(\mathbf{x}_i) = \mathbf{w}^\top \mathbf{x}_i + b = \pm 1$. We then get: $\varphi(f(\mathbf{x}_i)) = \eta$ for the $\mathbf{x}_i$ such that $\mathbf{w}^\top \mathbf{x}_i + b = -1$ and $\varphi(f(\mathbf{x}_i)) = 1 - \eta$ for the $\mathbf{x}_i$ such that $\mathbf{w}^\top \mathbf{x}_i + b = +1$, which, in both cases, leads to:

$$A = \ln\left(\frac{1}{\eta} - 1\right). \tag{14}$$

This hypothesis allows to set parameter A *a priori* but can be judged arbitrary. In the following general problem, we thus consider A (or equivalently a) as a scale parameter to learn. We then can chose to learn A or to set it *a priori*.

We define the associated pattern recognition problem as:

$$\begin{cases} \min\limits_{\substack{\mathbf{w}\in\mathbb{R}^d, \\ b,a,\xi_i,\xi_i^-,\xi_i^+\in\mathbb{R}}} \dfrac{1}{2}\|\mathbf{w}\|^2 + C\sum\limits_{i=1}^{n}\xi_i & \text{(15a)} \\ \qquad\qquad + \tilde{C}\sum\limits_{i=n+1}^{m}(\xi_i^- + \xi_i^+), & \\ \text{subject to} & \\ y_i(\mathbf{w}^\top \mathbf{x}_i + b) \geq 1 - \xi_i, \quad i = 1\ldots n & \text{(15b)} \\ az_i^- - \xi_i^- \leq \mathbf{w}^\top \mathbf{x}_i + b \leq az_i^+ + \xi_i^+, \; i = n+1\ldots m & \text{(15c)} \\ 0 \leq \xi_i, \quad i = 1\ldots n & \text{(15d)} \\ 0 \leq \xi_i^- \text{ and } 0 \leq \xi_i^+, \quad i = n+1\ldots m & \text{(15e)} \end{cases}$$

This formulation consists in minimizing the complexity of the model (15a) while forcing good classification (15b) and good probability estimation (close to p_i) (15c).
ξ_i, ξ_i^-, ξ_i^+ (15c, 15d) are slack variables measuring the degree of misclassification/misprediction of the datum $\mathbf{x}_i$. Parameters C and $\tilde{C}$ are predefined positive real numbers controlling the relative weighting of classification and regression performances. Obviously, if $n = m$, the problem boils down to the classical SVM.

We can rewrite the primal problem (eq. 15) in its dual form as described in Appendix A. Reformulated as a quadratic problem, it can be solved using standard SVM solvers software packages. The P-SVM code was implemented within the SVM-KM Toolbox [19]. It is open-source and can be downloaded on the project homepage http://remi.flamary.com/soft/soft-svmuncertain.html.

Moreover, as introduced in section II-A, the primal and dual formulations of the P-SVM problem can be easily generalized to non-linearly separable data by introducing kernel functions. This is detailed in Appendix B.

IV. Experiments With Synthetic Data Sets

In order to experimentally evaluate the proposed method for handling uncertain labels in SVM classification, we simulate different datasets described below. We compare the classification performances and probabilistic predictions of the SVM, F-SVM and P-SVM approaches.

In the standard SVM and the F-SVM approaches, prediction outputs are distances to the margin (which are unbounded), while the P-SVM formulation directly generates class probabilities. In order to perform a fair comparison, we thus decided to estimate the posterior class probabilities of the classical SVM and F-SVM outputs using Platt's scaling algorithm [13]. Classification performances are evaluated by computing the area under the ROC curve (AUC) and the accuracy (Acc). Probability prediction performances are evaluated by computing the Kullback-Leibler distance (KL) (or relative entropy) and the alignment error (Err_{Al}). Definitions of these different metrics are given in Appendix C.
The performance of P-SVM are compared to that of the standard SVM (+ Platt) and F-SVM (+ Platt) based on paired non-parametric sign tests performed on a collection of 100-bootstrap resamples drawn from the test points population.

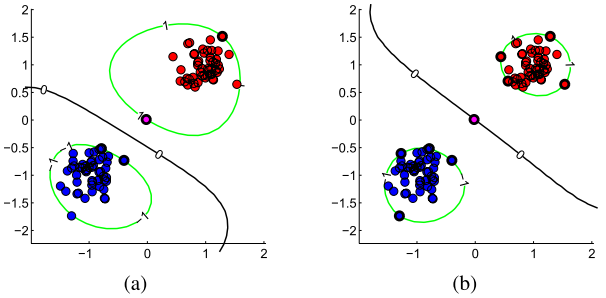

Fig. 3. Introduction of one outlier. We simulate two gaussian datasets labelled '−1' (blue) and '+1' (red), to visualize the position of the constructed P-SVM frontier (in black) and the margins (in green) depending on label l of the outlier (in pink): (a) $l = +1$, (b) $l = \mathbb{P}(Y = 1 \mid X = \mathbf{x}) = 0.51$, with $\eta = 0.1$.

In these numerical examples, $C = \tilde{C} = 100$ and a gaussian radial basis function kernel of the form:

$$k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$$
$$(u, v) \mapsto k(u, v) = \exp\left(-\frac{||u - v||^2}{2\sigma^2}\right),$$

is used, where parameter σ is set to 1.

A. Impact of One Outlier Sample

We simulate two gaussian data sets (n^l =100 learning points) labelled '+1' and '−1' and arbitrarily generate a unique outlier $\mathbf{x}$ located at quasi-equal distance from both centers, such that $\mathbb{P}(Y = 1 \mid \mathbf{X} = \mathbf{x}) = 0.51$.
We evaluate the impact of this point on the decision frontier depending on its label l:

- class label: $y = 1$, since $\mathbb{P}(Y = 1 \mid \mathbf{X} = \mathbf{x}) > 0.5$,
- probabilistic label: $p = 0.51$.

We train the P-SVM classifier on both datasets. Note that in the first case, we are brought back to classical SVM with binary labelled training samples.
Results are presented on Fig. 3. When label l equals '+1' (binary training dataset), the frontier is constructed so as to minimize classification error and maximize the margin: the outlier $\mathbf{x}$ then becomes a support vector that impacts the position of frontier [Fig. 3(a)] which is largely deviated to the dataset of class '−1'. We thus loose the generalization power of SVM. On the contrary, the P-SVM approach takes advantage of the probabilistic information learnt from probabilistic label $l = p = 0.51$ of the outlier. This very uncertain sample ($p \simeq 0.5$) lies on the separating frontier [Fig. 3(b)] while binary labelled data are, from a "maximum margin" point of view, separated in an optimal way.

B. Probability Estimation

We generate two unidimensional datasets, labelled '+1' and '−1', from normal distributions of variances $\sigma_{-1}^2 = \sigma_1^2 = 0.3$ and means $\mu_{-1} = -0.5$ and $\mu_1 = +0.5$ (see Fig. 4(a)). Let $(\mathbf{x}_i^l)_{i=1\ldots n^l}$ denotes the learning dataset ($n^l = 100$) and $(\mathbf{x}_i^t)_{i=1\ldots n^t}$ the test set (n^t=1000). We compute, for each point

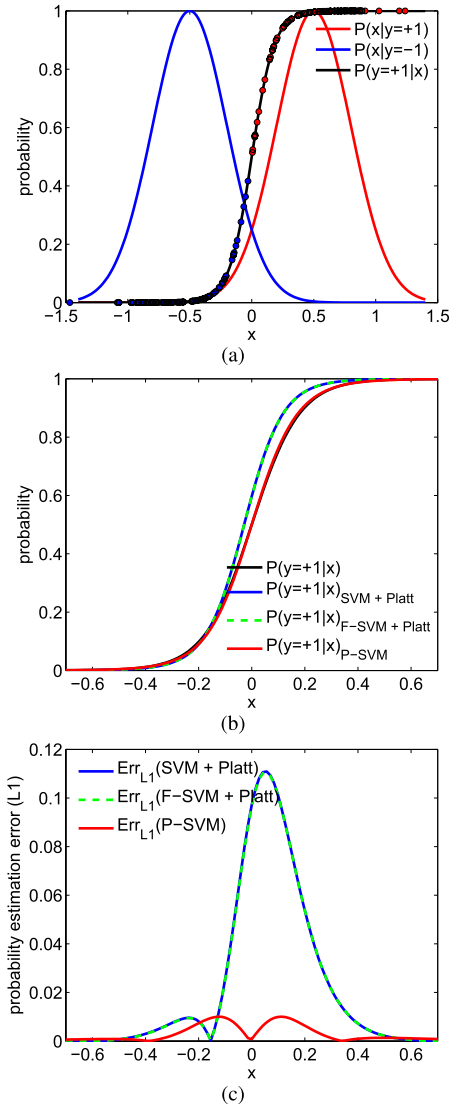

Fig. 4. Probability estimations comparison. Top plot (a) shows the true probability distributions for classes '−1' (blue) and '+1' (red) (ground truth); the overlaying circles represent the n^t learning examples. Middle plot (b) shows the true posterior probability (black) with SVM+Platt (blue), F-SVM+Platt (green) and P-SVM (red) estimations overlaying. Lower plot (c) shows the distance between true probabilities and estimations.

$\mathbf{x}_i$, $i = 1 \ldots n^l$, its true probability $\mathbb{P}(Y_i = +1|\mathbf{x}_i)$ to belong to class '+1':

$$\mathbb{P}(Y_i = +1|\mathbf{x}_i) = \frac{\mathbb{P}(\mathbf{x}_i|Y_i = +1)}{\mathbb{P}(\mathbf{x}_i|Y_i = +1) + \mathbb{P}(\mathbf{x}_i|Y_i = -1)}$$
$$\text{where } \mathbb{P}(\mathbf{x}_i|Y_i{=}+1) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{\mathbf{x}_i - \mu_1}{\sigma}\right)^2\right)$$
$$\text{and } \mathbb{P}(\mathbf{x}_i|Y_i{=}-1) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{\mathbf{x}_i - \mu_{-1}}{\sigma}\right)^2\right). \quad (16)$$

From here on, learning data are labelled in three ways, as follows:

- The regular SVM training dataset is obtained by setting a threshold of 0.5 on the true probability $p_i^l = \mathbb{P}(Y_i^l = +1|\mathbf{x}_i^l)$ for assigning class label y_i^l associated to point $\mathbf{x}_i^l$, for $i = 1 \ldots n^l$. This is what would be done in

TABLE I
COMPARISON OF P-SVM, SVM+PLATT AND F-SVM+PLATT CLASSIFICATION AND PREDICTION PERFORMANCES FOR NOISELESS PROBABILITY ESTIMATION PROBLEM

Algorithm	P-SVM	SVM	*(p-value)*	F-SVM	*(p-value)*
AUC	**1**	**1**	(0.29)	**1**	(0.07)
Acc	**1***	0.99	($< 10^{-3}$)	0.99	($< 10^{-3}$)
KL	**0.4***	11	($< 10^{-3}$)	11	($< 10^{-3}$)
Err_{Al}	$\mathbf{1.10^{-5}}$*	6.10^{-4}	($< 10^{-3}$)	6.10^{-4}	($< 10^{-3}$)

practical cases when the data contains class membership probabilities:

$$\begin{array}{l} \text{if } \; p_i^l > 0.5, \text{ then } \; y_i^l = +1, \\ \text{if } \; p_i^l \leq 0.5, \text{ then } \; y_i^l = -1. \end{array} \tag{17}$$

This dataset $(x_i^l, y_i^l)_{i=1\ldots n^l}$ is used to train the classical SVM classifier. Following SVM classification, Platt's algorithm is used to transform SVM output distances to class probability estimations.

- The P-SVM training dataset $(\mathbf{x}_i^l, \hat{y}_i^l)_{i=1\ldots n^l}$ is obtained as follow. For $i = 1 \ldots n^l$,

$$\begin{array}{ll} \text{if } \; p_i^l > 1-\eta, & \text{then } \hat{y}_i^l = +1, \\ \text{if } \; p_i^l < \eta, & \text{then } \hat{y}_i^l = -1, \\ \hat{y}_i^l = p_i^l & \text{otherwise.} \end{array} \tag{18}$$

  If the probability values are sufficiently close to 0 or 1 (within a confidence/precision interval η=0.01), we admit that they belong respectively to class "-1" or "+1". This probabilistic dataset $(\mathbf{x}_i^l, \hat{y}_i^l)_{i=1\ldots n^l}$ is used to train the P-SVM algorithm.
- Finally, the dataset used to the train F-SVM algorithm is of the form $(\mathbf{x}_i^l, y_i^l, m_i^l)_{i=1\ldots n^l}$ where $m_i^l = |2p_i^l - 1|$ is a weighting factor in the misclassification cost, as described in section II-B.2.

These three approaches are compared based on the test set $(\mathbf{x}_i^t)_{i=1\ldots n^t}$ and using the true probabilities $\left(\mathbb{P}(Y_i^t = +1|\mathbf{x}_i^t)\right)_{i=1\ldots n^t}$ to estimate probability prediction errors.
Fig. 4(b) shows the probability predictions achieved by the regular SVM and F-SVM coupled with Platt's algorithm, as well as those achieved by P-SVM. Performances are improved with the P-SVM classifier. The true probabilities (black) and P-SVM estimations (red) are indeed quasi-superimposed (KL = 0.4) whereas Platt's estimations (blue and green) are less accurate (KL = 11). This is also illustrated by the prediction error (L_1) computed on Fig. 4(c).

Table I summarizes the results of the quantitative evaluation on the $n^t = 1000$ random test points. The best performances are highlighted in bold for each of the four evaluation metrics, and the corresponding p-values for the comparison P-SVM versus others are given in parenthesis. As an example, the SVM and P-SVM classifiers perform equally based on the AUC metric (p = 0.29) while the P-SVM significantly outperform the SVM prediction probability performance based on the KL metric (p<0.001). Note that a superscript of star indicates when P-SVM performs statistically better than others at the 5% significance level. Classification performances (AUC and Acc) are either better or equivalent for P-SVM than for SVM and F-SVM while probability prediction errors (KL, Err_{Al}) are systematically lower for P-SVM than those achieved for the classical SVM and F-SVM (p<0.001).

C. Noise Robustness

We generate two 2D datasets, labelled '+1' and '−1', from normal distributions of variances $\sigma_{-1}^2 = \sigma_1^2 = 0.7$ and means $\mu_{-1} = (-0.3, 0.5)$ and $\mu_1 = (+0.3, +0.5)$. As in the previous experiment, we compute the probability of class '+1' membership for each point x^l of the learning data set. We simulate classification error by artificially adding a centered uniform noise of amplitude δ to the probabilities, such that for $i = 1 \ldots n^l$,

$$\hat{\mathbb{P}}(Y_i = +1|x_i) = \mathbb{P}(Y_i = 1|x_i) + \delta_i.$$

The learning data are then labelled following the same scheme as described in (17) and (18).
The three classifiers are trained on $n^l = 100$ noisy sample points ($\delta = 0.15$) and evaluated on $n^t = 1000$ random test points. Fig. 5 shows the margin location and probabilities estimations derived from using the three algorithms over a grid of values. The estimated classification performances and prediction errors are reported in Table II.

The P-SVM classification and probability estimations better correlate with the ground truth ($\text{Acc}_{\text{P-SVM}} = 0.98$, $\text{KL}_{\text{P-SVM}} = 23$) than SVM ($\text{Acc}_{\text{SVM}} = 0.93$, $\text{KL}_{\text{SVM}} = 255$; p<0.001). Contrary to P-SVM which, by combining both classification and regression, predicts good probabilities, SVM is more sensitive to the classification noise of the input training samples and does not converge any more to the Bayes rule as seen in [10]. This is confirmed by the estimated classification performances and prediction errors reported in Table II. As for F-SVM, the discrimination frontier is better located than for SVM when compared to the ground truth (as indicated by the classification accuracy $\text{Acc}_{\text{F-SVM}} = 0.96$), the noisy points located near the frontier are weighted based on their low class probability and thus have less impact on the separating hyperplan. But there remains a major scaling divergence in the predicted probability maps ($\text{KL}_{\text{F-SVM}} = 166$; p<0.001). Contrary to P-SVM, F-SVM doesn't use the probabilistic labels to construct a probability prediction function. These uncertain labels are only used as weighted factor in the classification boundary construction. Thus, the predicted probabilities are less accurate than for P-SVM. This is a major difference with P-SVM since the posterior class probability conveys information about the uncertainty measure regarding the class prediction. Note that far from the n^l learning data points (top left, bottom right corners of Fig. 5), every probability estimations are less accurate, this being directly linked to the choice of a gaussian kernel.

Fig. 6 shows the impact of noise amplitude on the SVM, F-SVM and P-SVM classification [Fig. 6(a)] and probability prediction [Fig. 6(b)] performances. Values were averaged over 100 random repeated simulations. Even when noise increases, classification performances (respectively

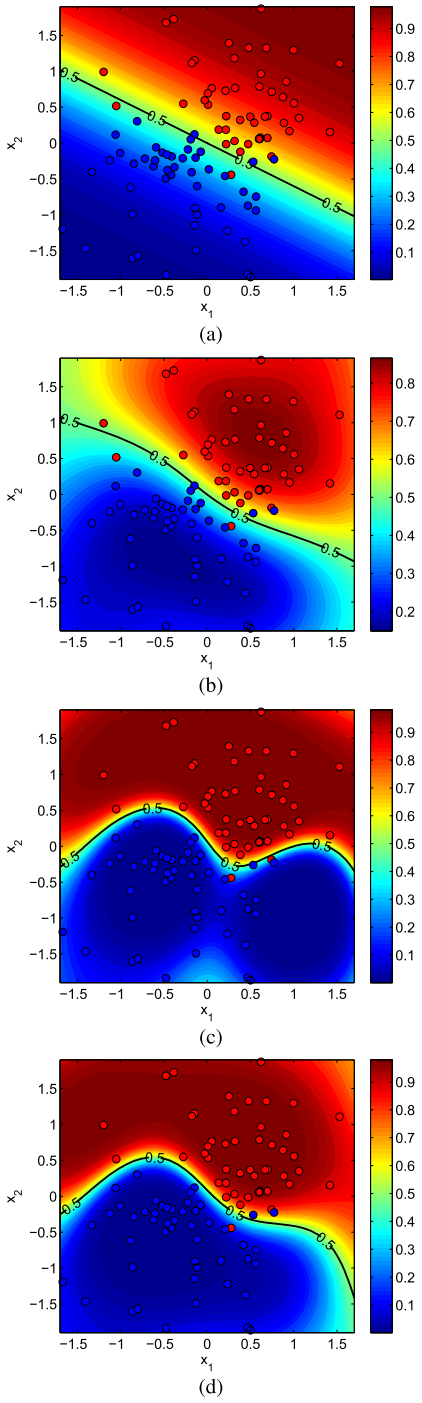

Fig. 5. Comparison of P-SVM, regular SVM+Platt and F-SVM+Platt robustness to labelling noise. (a) True probability distribution together with noisy learning data points, plotted in blue (class '−1') and red (class '+1') circles. Probability estimations of (b) P-SVM, (c) SVM+Platt and (d) F-SVM+Platt over a grid when trained on the noisy data points.

probability predictions errors) of the P-SVM remain significantly higher (respectively lower) than those of the classical SVM and F-SVM.

Finally, we evaluate the impact of the training dataset probabilistic labels proportion on the prediction performances of the P-SVM. Fig. 7 shows the evolution of the classification and probability predictions performance measures with increasing

TABLE II
COMPARISON OF P-SVM, SVM+PLATT AND F-SVM+PLATT CLASSIFICATION AND PREDICTION PERFORMANCES FOR NOISY PROBABILISTIC ESTIMATES

Algorithm	P-SVM	SVM	*(p-value)*	F-SVM	*(p-value)*
AUC	**1***	0.97	$(< 10^{-3})$	0.99	$(< 10^{-3})$
Acc	**0.98***	0.93	$(< 10^{-3})$	0.96	$(< 10^{-3})$
KL	**23***	255	$(< 10^{-3})$	166	$(< 10^{-3})$
Err_{Al}	**0.015***	0.061	$(< 10^{-3})$	0.043	$(< 10^{-3})$

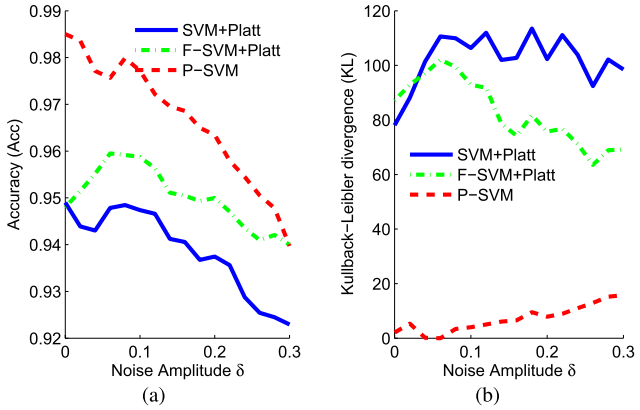

Fig. 6. Evolution of P-SVM, SVM+Platt and F-SVM+Platt classification and prediction performances depending on noise amplitude δ ($n^l = 100$, $n^t = 1000$, drawing repeated 100 times). (a) Classification performances. (b) Probability estimations performances.

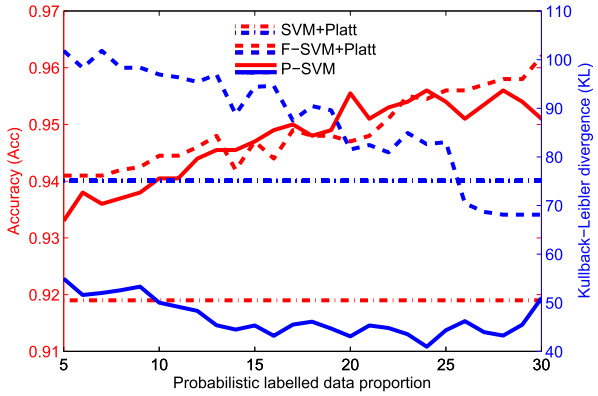

Fig. 7. Classification and probability prediction performances of the P-SVM and F-SVM+Platt depending on the proportion of probabilistic labels. Comparison to classical SVM+Platt performances.

number of probabilistic labels. Increasing the proportion of probabilistic training samples improves both classification and prediction performances.

V. APPLICATION TO A CLINICAL DATA SET

This section reports an experimental evaluation of our proposed P-SVM algorithm over a clinical dataset. The targeted task consists in discriminating benign from malignant regions on mpMR images of prostate cancer patients. The same metrics as in Section IV are used to compare P-SVM to the regular SVM and the F-SVM.

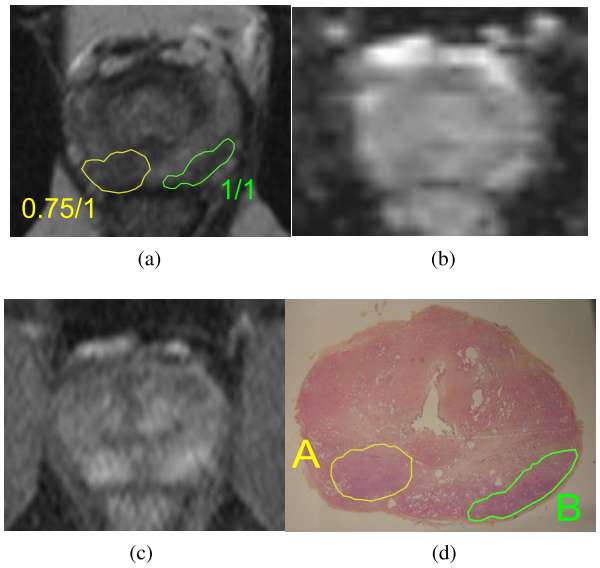

Fig. 8. Prostate MRI: (a) axial T2-weighted, (b) Apparent Diffusion Coefficient and (c) Dynamic Contrast-Enhanced (after Gd-injection) MR images together with the corresponding (d) histology slice. Histologically assessed cancers (A and B) were outlined on MR images (scored 0.75 and 1 respectively) during the blinded *a priori* MR analysis.

A. Data set Description

The dataset consists in a series of mpMR images of 49 patients, including T2-weighted, dynamic contrast-enhanced and diffusion-weighted imaging as illustrated on Fig. 8. A total of $n = 350$ regions of interest (ROI) were delineated on the images and scored by four experts using a five-point ordinal scale of confidence ranging from $0 =$ surely benign to $1 =$ surely malignant [9], thus with a discrete sampling interval of ± 0.125. All patients of this database underwent a prostatectomy after the MR exams. The prostatectomy specimens were analyzed *a posteriori* by an anatomopathologist thus providing the histological ground truth as seen on Fig. 8(d).

B. Experiments

In a previous study [20], we presented a computer-aided diagnosis (CAD) scheme based on the combination of a regular SVM algorithm and a set of discriminant statistical, structural (gradients and Haralick's attributes) and functional features derived from the mpMR images. This CAD system was trained on the series of 284 benign and 66 cancer regions of interest that were labelled '-1' or '$+1$' based on the histological ground truth. Learning and testing on this binary dataset led to AUC $= 0.855$ and Acc $= 0.871$.

In this study, we hypothesize that we do not have access to the histological ground truth, since it is rarely available in practice. The learning process is thus only based on the qualitative degree of confidence returned by the radiologists when analyzing the MR images. We test whether integrating this scoring information to the P-SVM allows achieving more reliable predictions than when using a F-SVM or a regular SVM trained on the thresholded scores.

We propose to use the confidence scores of the expert with the highest level of expertise to construct three distinct datasets:

- The first one, $(\mathbf{x}_i, y_i)_{i=1\ldots n}$, is the standard binary labelled dataset obtained from setting a threshold value of 0.5 on the confidence scores of the expert;
- The second one, $(\mathbf{x}_i, \hat{y}_i)_{i=1\ldots n}$, is a mixed dataset of binary labels and probability estimations constructed from the expert's confidence scores $(p_i)_{i=1\ldots n}$, defined as in (18);
- The third one, $(\mathbf{x}_i, y_i, m_i)_{i=1\ldots n}$, associates a class membership $m_i = |2p_i - 1|$ to each of the binary labelled example, where p_i is the radiologist's score.

These three different learning datasets are used to train the SVM, P-SVM and F-SVM classifiers respectively.
We then test the three predictive models obtained on two different testing datasets:

- The first one corresponds to the set of 350 ROIs scored by the expert and considering a threshold value of 0.5 for labelling ROIs as cancer or benign;
- The second test set corresponds to the set of 350 ROIs labelled as cancer or benign samples based on the histological ground truth.

In this analysis, the labelling precision η introduced in (12) is set to the confidence interval value of 0.125.
We finally propose, in a second experiment, a refinement of the formalism described in (12) by considering an adaptative parameter η_i which reflects the scoring "uncertainty" (or the labelling precision) for each target i. This uncertainty is derived from the standard deviation of the four experts' scores. We choose $\eta_i = \sigma_i$, where σ_i is the standard deviation of the scores attributed to target i. We estimate the conditional probabilities p_i as the average score values over all experts for each target i. If experts agree, then η_i is small, whereas if inter-reader variability is high, η_i is large to account for uncertainty in the learning stage.
Given the limited size of the patient database, classification performance was estimated using a leave-one-patient-out cross-validation approach [20], thus avoiding training and testing on the same data. As for the synthetic experiments, we perform a bootstrap resampling based on 100 resamples to compute one-sided sign tests of the difference in performances of the P-SVM versus the F-SVM and the standard SVM.

C. Results

Based on the results obtained in [20], we set parameters $\sigma = 2^5$ and $C = \tilde{C} = 2^{13}$. Tables III–VI report the performances for the discrimination of benign and malignant tissues in mpMR imaging of prostate cancer patients. The best performances are highlighted in bold. A superscript star indicates when P-SVM performs statistically better than the two others at the 5% significance level.

Table III evaluates the performance achieved using the radiologist of higher level of expertise analysis to label both the learning and test datasets, thus assuming that the histological reference is unknown. P-SVM achieves better classification

TABLE III
P-SVM, SVM+PLATT AND F-SVM+PLATT PERFORMANCES WHEN TRAINED AND TESTED ON SCORES OF THE RADIOLOGIST WITH THE HIGHER LEVEL OF EXPERTISE

Algorithm	P-SVM	SVM	*(p-value)*	F-SVM	*(p-value)*
AUC	**0.889***	0.845	$(< 10^{-3})$	0.847	$(< 10^{-3})$
Acc	**0.909**	0.905	(0.3)	0.900	$(< 10^{-3})$
KL	**43.4***	75.7	$(< 10^{-3})$	74.6	$(< 10^{-3})$
Err_{Al}	**0.256***	0.306	$(< 10^{-3})$	0.318	$(< 10^{-3})$

TABLE IV
P-SVM, SVM+PLATT AND F-SVM+PLATT PERFORMANCES WHEN TRAINED ON SCORES OF THE RADIOLOGIST WITH THE HIGHER LEVEL OF EXPERTISE, AND TESTED WITH RESPECT TO THE HISTOLOGY

Algorithm	P-SVM	SVM	*(p-value)*	F-SVM	*(p-value)*
AUC	**0.857***	0.817	$(< 10^{-3})$	0.832	$(< 10^{-3})$
Acc	**0.863***	0.857	$(< 10^{-3})$	0.854	$(< 10^{-3})$

TABLE V
P-SVM, SVM+PLATT AND F-SVM+PLATT PERFORMANCES WHEN TRAINED AND TESTED ON THE FOUR EXPERTS' SCORES (AVERAGE), USING ADAPTATIVE η (STANDARD DEVIATION OF THE SCORES)

Algorithm	P-SVM	SVM	*(p-value)*	F-SVM	*(p-value)*
AUC	**0.888***	0.861	$(< 10^{-3})$	0.876	$(< 10^{-3})$
Acc	**0.883**	0.868	$(< 10^{-3})$	0.880	(0.002)
KL	**30.8***	59.6	$(< 10^{-3})$	56.5	$(< 10^{-3})$
Err_{Al}	**0.189***	0.226	$(< 10^{-3})$	0.213	$(< 10^{-3})$

TABLE VI
P-SVM, SVM+PLATT AND F-SVM+PLATT PERFORMANCES WHEN TRAINED ON THE FOUR EXPERTS' SCORES (AVERAGE), USING ADAPTATIVE η (STANDARD DEVIATION OF THE SCORES), AND TESTED WITH RESPECT TO THE HISTOLOGY

Algorithm	P-SVM	SVM	*(p-value)*	F-SVM	*(p-value)*
AUC	**0.862**	0.847	$(< 10^{-3})$	0.861	(0.18)
Acc	**0.863**	0.857	$(< 10^{-3})$	0.860	(0.07)

(AUC = 0.89) and probability estimation (KL = 43) performances than SVM (AUC = 0.85 and KL = 76) and F-SVM (AUC = 0.85 and KL = 75). These first results show that P-SVM better reproduces the diagnosis of the expert radiologist than do both the classical SVM and the F-SVM. This implies that P-SVM might be a useful tool for a CAD system that would be dedicated to training junior radiologists for instance.

Table IV reports the performance achieved using the expert's scores for learning and the histological ground truth for testing. This shows that by including the expert's uncertainty into the learning step, P-SVM balances the influence of uncertain data and allows achieving statistically better classification performances with respect to the histological ground truth (AUC = 0.86) than those achieved with classical SVM (AUC = 0.82; p <0.001). The P-SVM classification performance is also better than that achieved with the F-SVM (AUC = 83; p <0.001), thus demonstrating that P-SVM better handles uncertain data than F-SVM in this clinical context.

Another important result of this comparison is that classification performance in the ideal case where the histological ground truth is available for training and testing a regular SVM (AUC = 0.86 in [20]) and those obtained by training P-SVM on the expert's scores (AUC = 0.86 in Table IV) are equivalent. This suggests that the radiologist expertise could be sufficient to construct a reliable classifier. This result however is strongly dependant on the radiologist diagnostic expertise.

The tissues discrimination performances obtained by combining the scores from the four experts and using an adaptative parameters η_i for each training ROI i depending on the inter-experts' variability are reported on Tables V and VI. For each target i, we define p_i as the average score values over all experts and $\eta_i = \sigma_i$, the standard deviation of the scores. P-SVM, F-SVM and SVM are trained on these newly defined scores. Evaluation is performed either by considering the average score over the 4 readers (Table V) or the histology (Table VI) as the ground truth. As we can see, classification performances reported in Table VI are slightly improved compared to those in Table IV but the gain is not significant for this specific application. Nevertheless, this approach allows to obtain better probability estimates with KL = 30.8 instead of KL = 43 using a unique value of η.

VI. DISCUSSION AND PERSPECTIVES

The proposed method aims at learning a prediction function that will both discriminate samples and be able to predict class probabilities. The approach can be easily extended to other kind of prediction (non-probabilistic regression, multi-class prediction). We proposed a generic approach to handle heterogeneous labeled datasets, containing both quantitative and qualitative labels. In this sense, we proposed a multitask method to jointly address a task of classification and a task of regression [21]. Note that the information between the different tasks is shared in our framework through a common prediction function whose prediction score is used for both tasks. It seems interesting in future works to investigate other multitask approaches where one function is learned per prediction task and the information is shared through regularization [22].

The basic idea of P-SVM has been introduced in our preliminary work [15]. In this paper, we generalized the formulation for the optimization problem by introducing the scaling parameter a, which can be either learnt or set *a priori*. We proposed to use an adaptative η parameter depending on the sample uncertainty rather than a global measure depending on the labelling scale. Besides, performances are slightly improved with the use of an adaptative η, set depending on reader's variability, when compared to a constant η value, set depending on the labelling precision. Further study is required to better understand the impact of both a and η. In this paper, we also discussed a lot more comparative

synthetic experiments: we studied the impact of outliers on the prediction function, the impact of the number of probabilistic labels introduced in the training step, the influence of labelling noise and introduced more evaluation criteria. We also evaluated the P-SVM algorithm in a challenging clinical context using radiologists scores as probabilistic inputs. At last, we compared the P-SVM to two other state-of-the-art methods.

With both the synthetic and the clinical datasets, we showed that the P-SVM classification and probability prediction performances compared favorably with that of the standard SVM and Fuzzy-SVM.

The synthetic examples allowed to show that the proposed P-SVM behaves efficiently in presence of outliers or labelling noise when compared to others methods. On the contrary, the regular SVM (combined with Platt's algorithm) can't integrate any measure of uncertainty. It is thus very sensitive to the presence of outliers, which strongly affect the position of the classification frontier, and is not robust to labelling noise. The F-SVM formalism allows to weight the misclassification cost using labelling uncertainty. The classification frontier is thus less impacted with labelling noise but the predicted posterior class probability is less accurate.

In the clinical data example, it is interesting to notice that the classification performances reached when training a classical SVM on the binary histological ground truth are quasi-equivalent to those obtained when training a P-SVM on the experts' scores. If these preliminary results are confirmed on larger databases, this could open new perspectives for CAD systems design, especially in medical imaging. Indeed, constructing a training database based on the histological ground truth is expensive, time consuming, fastidious and requires anatomopathologists and radiologists to work in consensus to register histological slices onto MR images. On the contrary, using radiologists blind analysis (scores) would be much easier, thus allowing to construct larger training databases.

The development of CAD systems that can assist radiologists in their diagnostic task of prostate cancer screening on MR images has gained interest in the past few years ([1], [23]–[26]). Most of these prototypes rely on training datasets of patients for which the histological ground truth has been annotated following radical prostatectomy to guarantee a certain and binary labelling. Consequently, these datasets are bound to be of limited size. Moroever, the study population is limited to patients who underwent prostatectomy which restrains the representativeness of the population under study. Including the data of patients who underwent a radiological exam but for whom no surgical treatment have been performed could enhance this representativeness. We can thus consider combining datasets for which the ground truth is either defined with respect to the expertise of radiologists or to the histological analysis of anatomopathologists (when the prostatectomy specimen analysis is performed).

Note that for the synthetic experiments, we have only showed the results of the algorithms when tested on continuously labelled datasets whereas our clinical dataset is discrete since it uses the Likert scoring system [9]. We have also tested and compared the different algorithms on discrete synthetic datasets (data not shown), this leading to the same conclusions. The idea of the paper was to show that the proposed algorithm can adapt to a large span of annotation scales and labelling noises. In particular, the five-points Likert scale used in this study is not the only one, others prefer to use a continuous $0-100\%$ scale (e.g. [27]).

Note that Doyle *et al* [28] tackle the problem of time and effort dedicated to the construction of a training database relying on the histological ground truth with another perspective. They introduce an intelligent labelling strategy which aims at selecting only informative examples for annotation, ie. those which would increase accuracy of the resulting trained classifier. It would be an interesting idea to combine both approaches to limit the effort dedicated to the annotation of a training database by preselecting the samples of interest while taking into account the uncertainty of some labels (scores) into the learning process.

We can also note that some studies tackle the related problem of learning a SVM classifier when the input vectors $\{\mathbf{x}_i\}_i$ rather than the class labels are noisy. Bi and Zhang [29], for instance, introduced an additive noise Δ, such that $\mathbf{x}'_i = \mathbf{x}_i + \Delta\mathbf{x}_i$ where noise $\Delta\mathbf{x}_i$ is constrained to be bounded by δ_i; similarly, Yang and Gunn [30] modelled input data $\{\mathbf{x}_i\}_i$ with gaussian distributions.

VII. Conclusion

We present a new way to take into account both qualitative and quantitative target data by shrewdly combining both SVM classification and regression loss. Experimental results show that our formulation can perform very well on simulated data for discrimination as well as posterior probability estimation. We have also tested our approach on a clinical dataset thus allowing to assess its usefulness in designing a computer-assisted diagnosis system for prostate cancer. These results are promising since they suggest that we could construct larger and less expensive training database for our classifier by combining samples labelled with respect to the histological ground truth or the radiologic expertise, taking into account the radiologist's uncertainty instead of discarding it. Note that the proposed framework, initially designed for probabilistic labels, can be generalized to other problems involving both qualitative and quantitative labels such as censored data.

Appendix

A. P-SVM Dual Formulation

The primal formulation introduced in III can be rewritten into its dual form by introducing Lagrange multipliers $\boldsymbol{\alpha}$, $\boldsymbol{\mu}^+$, $\boldsymbol{\mu}^-$, $\boldsymbol{\gamma}^+$, $\boldsymbol{\gamma}^-$. We are looking for a stationary point for the Lagrange function $\mathcal{L}$:

$$\max_{\alpha,\beta,\mu^+,\mu^-,\gamma^+,\gamma^-} \min_{\mathbf{w},b,a,\xi,\xi^+,\xi^-} \mathcal{L} \tag{19}$$

where

$$\begin{aligned}\mathcal{L}(\mathbf{w},b,a,\xi,\alpha,\beta,\xi^-,\xi^+,\mu^-,\mu^+,\gamma^-,\gamma^+)\\ =\frac{1}{2}\|\mathbf{w}\|^2+C\sum_{i=1}^{n}\xi_i+\tilde{C}\sum_{i=n+1}^{m}(\xi_i^-+\xi_i^+)\\ -\sum_{i=1}^{n}\alpha_i(y_i(\mathbf{w}^\top\mathbf{x}_i+b)-(1-\xi_i))-\sum_{i=1}^{n}\beta_i\xi_i\\ -\sum_{i=n+1}^{m}\mu_i^-((\mathbf{w}^\top\mathbf{x}_i+b)-(az_i^--\xi_i^-))-\sum_{i=n+1}^{m}\gamma_i^-\xi_i^-\\ -\sum_{i=n+1}^{m}\mu_i^+((az_i^++\xi_i^+)-(\mathbf{w}^\top\mathbf{x}_i+b))-\sum_{i=n+1}^{m}\gamma_i^+\nu_i^+\end{aligned}$$

with $\alpha\geq 0$, $\beta\geq 0$, $\mu^+\geq 0$, $\mu^-\geq 0$, $\gamma^+\geq 0$ and $\gamma^-\geq 0$.

Computing the derivatives of $\mathcal{L}$ with respect to primal parameters $\mathbf{w}, b, \xi, \xi^-$ and ξ^+ leads to the following optimality conditions:

$$\mathbf{w}=\sum_{i=1}^{n}\alpha_i y_i\mathbf{x}_i-\sum_{i=n+1}^{m}(\mu_i^+-\mu_i^-)\mathbf{x}_i \quad (20a)$$

$$\sum_{i=1}^{n}\alpha_i y_i=\sum_{i=n+1}^{m}(\mu_i^+-\mu_i^-) \quad (20b)$$

$$\sum_{i=n+1}^{m}\mu_i^-z_i^-=\sum_{i=n+1}^{m}\mu_i^+z_i^+ \quad (20c)$$

$$Ce_1=\alpha+\beta \quad (20d)$$

$$\tilde{C}e_2=\mu^-+\gamma^-=\mu^++\gamma^+. \quad (20e)$$

where

$$e_1=[\underbrace{1\ldots1}_{n\text{ times}}\ \underbrace{0\ldots0}_{(m-n)\text{ times}}\ ]^\top \text{ and } e_2=[\underbrace{0\ldots0}_{n\text{ times}}\ \underbrace{1\ldots1}_{(m-n)\text{ times}}\ ]^\top.$$

Calculations simplifications then lead to:

$$\begin{aligned}L(\mathbf{w},b,\xi,\xi^-,\xi^+,\alpha,\beta,\mu,\gamma^+,\gamma^-)\\ =-\frac{1}{2}\mathbf{w}^\top\mathbf{w}+\sum_{i=1}^{n}\alpha_i+\sum_{i=n+1}^{m}\mu_i^-z_i^--\sum_{i=n+1}^{m}\mu_i^+z_i^+. \quad (21)\end{aligned}$$

Knowing that $\beta\geq 0$, $\gamma^+\geq 0$, $\gamma^-\geq 0$, conditions (20d) and (20e) become:

$$\begin{cases}0\leq\alpha_i\leq C, i=1...n\\ 0\leq\mu_i^+,\mu_i^-\leq\tilde{C}, i=n+1...m.\end{cases}$$

Finally, let $\Gamma=[\alpha_1\ldots\alpha_n\ \mu_{n+1}^+\ldots\mu_m^+\ \mu_{n+1}^-\ldots\mu_m^-]^\top$ be a vector of dimension $2m-n$. Then:

$$\mathbf{w}^\top\mathbf{w}=\Gamma^\top G\Gamma$$

where

$$G=\begin{pmatrix}\mathbf{K}_1 & -\mathbf{K}_2 & \mathbf{K}_2\\ -\mathbf{K}_2^\top & \mathbf{K}_3 & -\mathbf{K}_3\\ \mathbf{K}_2^\top & -\mathbf{K}_3 & \mathbf{K}_3\end{pmatrix}$$

with

$$\begin{aligned}\mathbf{K}_1&=(y_iy_j\mathbf{x}_i^\top\mathbf{x}_j)_{i,j=1...n},\\ \mathbf{K}_2&=(\mathbf{x}_i^\top\mathbf{x}_jy_i)_{i=1...n,j=n+1...m},\\ \mathbf{K}_3&=(\mathbf{x}_i^\top\mathbf{x}_j)_{i,j=n+1...m}.\end{aligned}$$

The dual formulation of the optimization problem becomes:

$$\begin{cases}\min\limits_{\Gamma} & \frac{1}{2}\Gamma^\top G\Gamma-\tilde{e}^\top\Gamma,\\ & \mathbf{f}^\top\Gamma=0\\ & \mathbf{g}^\top\Gamma=0\\ \text{and} & 0\leq\Gamma\leq[\underbrace{C\ldots C}_{n\text{ times}}\ \underbrace{\tilde{C}\ldots\tilde{C}}_{m-n\text{ times}}\ \underbrace{\tilde{C}\ldots\tilde{C}}_{m-n\text{ times}}]^\top\\ \text{with} & \Gamma=[\alpha_1\ldots\alpha_n,\mu_{n+1}^+\ldots\mu_m^+,\mu_{n+1}^-\ldots\mu_m^-]^\top\\ & \tilde{\mathbf{e}}=[\underbrace{1\ldots1}_{n\text{ times}}\ \underbrace{-z_{n+1}^+\cdots-z_m^+}_{m-n\text{ times}}\ \underbrace{z_{n+1}^-\ldots z_m^-}_{m-n\text{ times}}]\\ & \mathbf{f}^\top=[\mathbf{y}^\top,\underbrace{-1\cdots-1}_{m-n\text{ times}},\ \underbrace{1\ldots1}_{m-n\text{ times}}]\\ & \mathbf{g}^\top=[\underbrace{0\ldots0}_{n\text{ times}},\underbrace{-z_{n+1}^+\cdots-z_m^+}_{m-n\text{ times}}\ \underbrace{z_{n+1}^-\ldots z_m^-}_{m-n\text{ times}}].\end{cases} \quad (22)$$

B. Kernelization

The primal and dual formulations of the P-SVM problem (Eqs. 15 and 22) can be easily generalized to non-linearly separable data by introducing kernel functions. Let k be a positive kernel satisfying Mercer's condition and $\mathcal{H}$ the associated Reproducing Kernel Hilbert Space (RKHS). Within this framework, the primal formulation becomes

$$\begin{cases}\min\limits_{\substack{f\in\mathcal{H},\\ b,a,\xi,\xi^-,\xi^+\in\mathbb{R}}} & \frac{1}{2}\|f\|_{\mathcal{H}}^2+C\sum_{i=1}^{n}\xi_i+\tilde{C}\sum_{i=n+1}^{m}(\xi_i^-+\xi_i^+)\\ \text{subject to} & \\ & y_if(\mathbf{x}_i)\geq 1-\xi_i, & i=1...n\\ & az_i^--\xi_i^-\leq f(\mathbf{x}_i)\leq az_i^++\xi_i^+, & i=n+1...m\\ & 0\leq\xi_i, & i=1...n\\ & 0\leq\xi_i^- \text{ and } 0\leq\xi_i^+ & i=n+1...m\end{cases} \quad (23)$$

The dual formulation remains identical, with

$$\begin{aligned}\mathbf{K}_1&=(y_iy_jk(\mathbf{x}_i,\mathbf{x}_j))_{i,j=1...n},\\ \mathbf{K}_2&=(k(\mathbf{x}_i,\mathbf{x}_j)y_i)_{i=1...n,j=n+1...m},\\ \mathbf{K}_3&=(k(\mathbf{x}_i,\mathbf{x}_j))_{i,j=n+1...m},\end{aligned} \quad (24)$$

C. Metrics

We present the metrics introduced in IV to evaluate the prediction (class label and class probability) performances.

Classification performances are evaluated by computing:

- The area under the ROC curve (AUC), which is a global performance measure expressing the compromise between sensitivity and specificity,
- The accuracy (Acc), which represents the good classification rate when a specific threshold is applied to the classifier output,

$$Acc=\frac{\text{True positives}+\text{True negatives}}{\text{Total number of samples}}.$$

These two metrics range within the [0, 1] interval with higher values indicating better classification performance.

Probability estimation performances are evaluated by computing two metrics that express a dissimilarity measure between two probability distributions P (ground truth) and Q (estimated):

- The Kullback Leibler distance (KL) (or relative entropy), which is a measure of the information lost when Q is used to approximate P,

$$KL(P||Q) = \sum_{i=1}^{n} P(Y_i = 1|\mathbf{x}_i) \log\left(\frac{P(Y_i = 1|\mathbf{x}_i)}{Q(Y_i = 1|\mathbf{x}_i)}\right),$$

- The alignment error (Err_{Al}), which can be assimilated to the inner product and measures how the distributions are misaligned,

$$Err_{Al} = 1 - \frac{\sum_{i=1}^{n} P(Y_i = 1|\mathbf{x}_i) Q(Y_i = 1|\mathbf{x}_i)}{\sqrt{\sum_i P(Y_i = 1|\mathbf{x}_i)^2}\sqrt{\sum_i Q(Y_i = 1|\mathbf{x}_i)^2}}.$$

These metrics are unbounded, with small values indicating good estimation of the reference distribution.

REFERENCES

[1] Y. Artan, M. Haider, D. Langer, T. van der Kwast, A. J. Evans, Y. Yang, *et al.*, "Prostate cancer localization with multispectral MRI using cost-sensitive support vector machines and conditional random fields," *IEEE Trans. Image Process.*, vol. 19, no. 9, pp. 2444–2455, Sep. 2010.

[2] J. Zhang and L. Ye, "Content based image retrieval using unclean positive examples," *IEEE Trans. Image Process.*, vol. 18, no. 10, pp. 2370–2375, Oct. 2009.

[3] F. Bovolo, L. Bruzzone, and L. Carlin, "A novel technique for subpixel image classification based on support vector machine," *IEEE Trans. Image Process.*, vol. 19, no. 11, pp. 2983–2999, Nov. 2010.

[4] G. Camps-Valls and L. Bruzzone, *Kernel Methods for Remote Sensing Data Analysis*. New York, NY, USA: Wiley, 2009.

[5] I. Chan, W. Wells, R. V. Mulkern, S. Haker, J. Zhang, K. H. Zou, *et al.*, "Detection of prostate cancer by integration of line-scan diffusion, T2-mapping and T2-weighted magnetic resonance imaging; a multichannel statistical classifier," *Med. Phys.*, vol. 30, no. 9, pp. 2390–2398, Sep. 2003.

[6] P. Puech, N. Betrouni, N. Makni, A. Dewalle, A. Villers, and L. Lemaitre, "Computer-assisted diagnosis of prostate cancer using DCE-MRI data: Design, implementation and preliminary results," *Int. J. Comput. Assist. Radiol. Surgery*, vol. 4, no. 1, pp. 1–10, Jan. 2009.

[7] S. Mohamed, M. Salama, M. Kamek, and K. Rizkalla, "Region of interest based prostate tissue characterization using least square support vector machine LS-SVM," in *Image Analysis and Recognition* (Lecture Notes in Computer Science). New York, NY, USA: Springer-Verlag, 2004, pp. 51–58.

[8] P. Tiwari, M. Rosen, and A. Madabushi, "A hierarchical spectral clustering and nonlinear dimensionality reduction scheme for detection of prostate cancer from magnetic resonance spectroscopy (MRS)," *Med. Phys.*, vol. 36, no. 9, pp. 3927–3939, Sep. 2009.

[9] L. Dickinson, H. U. Ahmed, C. Allen, J. O. Barentsz, B. Carey, J. J. Futterer, *et al.*, "Magnetic resonance imaging for the detection, localisation, and characterisation of prostate cancer: Recommendations from a European consensus meeting," *Eur. Urol.*, vol. 59, no. 4, pp. 477–494, Apr. 2011.

[10] G. Stempfel and L. Ralaivola, "Learning SVMs from sloppily labeled data," in *Artificial Neural Networks-ICANN*. New York, NY, USA: Springer-Verlag, 2009, pp. 884–893.

[11] B. Schölkopf and A. J. Smola, *Learning With Kernels: Support Vector Machines, Regularization, Optimization, and Beyond* (Adaptive Computation and Machine Learning), 1st ed. Cambridge, MA, USA: MIT Press, Dec. 2001.

[12] C. Lin and S. Wang, "Fuzzy support vector machine," *IEEE Trans. Neural Netw.*, vol. 13, no. 2, pp. 464–471, Mar. 2002.

[13] J. C. Platt, "Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods," in *Advances in Large Margin Classifiers*. Cambridge, MA, USA: MIT Press, Mar. 1999, pp. 61–74.

[14] P. Sollich, "Probabilistic methods for support vector machines," in *Advances in Neural Information Processing Systems*. Cambridge, MA, USA: MIT Press, 2000, pp. 349–355.

[15] E. Niaf, R. Flamary, C. Lartizien, and S. Canu, "Handling uncertainties in SVM classification," in *Proc. IEEE Workshop Statist. Signal Process.*, Jun. 2011, pp. 757–760.

[16] V. N. Vapnik, *Statistical Learning Theory*, 1st ed. New York, NY, USA: Wiley, Sep. 1998.

[17] V. Franc, A. Zien, and B. Scholkopf, "Support vector machines as probabilistic models," in *Proc. 28th ICML*, 2011, pp. 665–672.

[18] Y. Grandvalet, J. Mariéthoz, and S. Bengio, "A probabilistic interpretation of SVMs with an application to unbalanced classification," in *Proc. Adv. NIPS*, Nov. 2006, pp. 1–11.

[19] S. Canu, Y. Grandvalet, V. Guigue, and A. Rakotomamonjy, "SVM and kernel methods MATLAB toolbox," Perception, Syst. et Inf., INSA de Rouen, Rouen, France, 2005.

[20] E. Niaf, O. Rouvière, F. Mège-Lechevallier, F. Bratan, and C. Lartizien, "Computer-aided diagnosis of prostate cancer in the peripheral zone using multiparametric MRI," *Phys. Med. Biol.*, vol. 57, no. 12, pp. 3833–3851, Jun. 2012.

[21] S. Ben-David and R. Schuller, "Exploiting task relatedness for multiple task learning," in *Learning Theory and Kernel Machines* (Lecture Notes in Computer Science). New York, NY, USA: Springer-Verlag, 2003, pp. 567–580.

[22] X. Yang, S. Kim, and E. Xing, "Heterogeneous multitask learning with joint sparsity constraints," in *Proc. Adv. Neural Inf. Process. Syst.*, 2009, pp. 2151–2159.

[23] S. Viswanath, B. Bloch, E. Genega, N. Rofsky, R. Lenkinski, J. Chappelow, *et al.*, "A comprehensive segmentation, registration, and cancer detection scheme on 3 Tesla in vivo prostate DCE-MRI," in *Proc. Int. Conf. MICCAI*, Jan. 2008, pp. 662–669.

[24] P. Vos, J. Barentsz, N. Karssemeijer, and H. Huisman, "Automatic computer-aided detection of prostate cancer based on multiparametric magnetic resonance image analysis," *Phys. Med. Biol.*, vol. 57, no. 6, pp. 1527–1542, Mar. 2012.

[25] S. Ozer, D. Langer, X. Liu, M. Haider, T. van der Kwast, A. Evans, *et al.*, "Supervised and unsupervised methods for prostate cancer segmentation with multispectral MRI," *Med. Phys.*, vol. 37, no. 4, pp. 1873–1883, Apr. 2010.

[26] R. Lopes, A. Ayache, N. Makni, P. Puech, A. Villers, S. Mordon, *et al.*, "Prostate cancer characterization on MR images using fractal features," *Med. Phys.*, vol. 38, no. 1, pp. 83–95, Jan. 2011.

[27] T. Hambrock, P. Vos, C. H. van de Kaa, J. Barentsz, and H. Huisman, "Prostate cancer: Computer-aided diagnosis with multiparametric 3-T MR imaging—Effect on observer performance," *Radiology*, vol. 266, no. 2, pp. 521–30, Feb. 2013.

[28] S. Doyle, J. Monaco, M. Feldman, J. Tomaszewski, and A. Madabhushi, "An active learning based classification strategy for the minority class problem: Appplication to histopathology annotation," *BMC Informat.*, vol. 12, no. 1, pp. 1–14, Oct. 2011.

[29] J. Bi and T. Zhang, "Support vector classification with input data uncertainty," in *Proc. Adv. Neural Inf. Process. Syst.*, 2005, pp. 161–169.

[30] J. Yang and S. Gunn, "Exploiting uncertain data in support vector classification," in *Proc. LNAI*, Sep. 2007, pp. 148–155.

Émilie Niaf received the Dipl.-Ing. degree in computer science and applied mathematics from Ecole Nationale Supérieur d'Informatique et Mathématiques Appliquées of Grenoble, France, and the M.Sc. degree in image processing from the University Paris VI and Télécom ParisTech, Paris, France, in 2008 and 2009, respectively, and the Ph.D. degree in image processing from Claude Bernard University, Lyon, France, in 2012. She is currently a Post-Doctoral Researcher with the Medical Imaging Research Center CREATIS, Lyon, France. Her current research interests include biomedical imaging with a focus on MR images, image processing, and machine learning applied to computer-assisted diagnosis systems.

Rémi Flamary is an Assistant Professor with Université de Nice Sophia-Antipolis and has been a member of the Lagrange Laboratory/Observatoire de la Côte d'Azur since 2012. He received the Dipl.-Ing. degree in electrical engineering and the M.S. degree in image processing from Institut National de Sciences Appliquées de Lyon in 2008 and the Ph.D. degree from the University of Rouen in 2011. His current research interest involve signal processing, machine learning, and image processing.

Carole Lartizien received the bachelor's degree in nuclear engineering from the National Polytechnic Institute, Grenoble, France, in 1996, and the master's degree in biomedical engineering and the Ph.D. degree in medical imaging from the University Paris XI, France, in 1997 and 2001, respectively. She is a Research Associate with CNRS and the University of Lyon, France, and is conducting research at the CREATIS laboratory, whose aim is to develop image processing methods for medical imaging. Her research interests include supervised methods for medical image analysis, kernel learning approaches (e. g. SVM) to classification problems, and the prototyping of computer-aided diagnosis system for cancer and neuro-imaging.

Olivier Rouvière is a Professor of radiology and the Head of the Department of Imaging, Edouard Herriot University Hospital, Lyon, France. He has specialized in Genitourinary and Vascular diagnostic and interventional imaging. His main field of research is prostate cancer imaging using ultrasound based approaches and multiparametric MRI.

Stéphane Canu is a Professor of the LITIS Research Laboratory and the Information Technology Department, National Institute of Applied Science, Rouen. He received the Ph.D. degree in system command from the Compiègne University of Technology in 1986. He joined the faculty of the Department of Computer Science, Compiègne University of Technology, in 1987. He received the French Habilitation degree from Paris 6 University. In 1997, he joined the Rouen Applied Sciences National Institute as a Full Professor, where he created the Information Engineering Department. He has been the Dean of this department since 2002, when he was named director of the computing service and facilities unit. In 2004, he joined the Machine Learning Group, ANU/NICTA, Canberra, with A. Smola and B. Williamson. He has published over 30 papers in refereed conference proceedings or journals in the areas of theory, algorithms and applications using kernel machines learning algorithm, and other flexible regression methods. His research interests includes kernels machines, regularization, machine learning applied to signal processing, pattern classification, factorization for recommander systems, and learning for context aware applications.