Handling learning samples uncertainties in SVM : application to MRI-based prostate cancer Computer-Aided Diagnosis

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

emilie.niaf@creatis.insa-lyon.fr

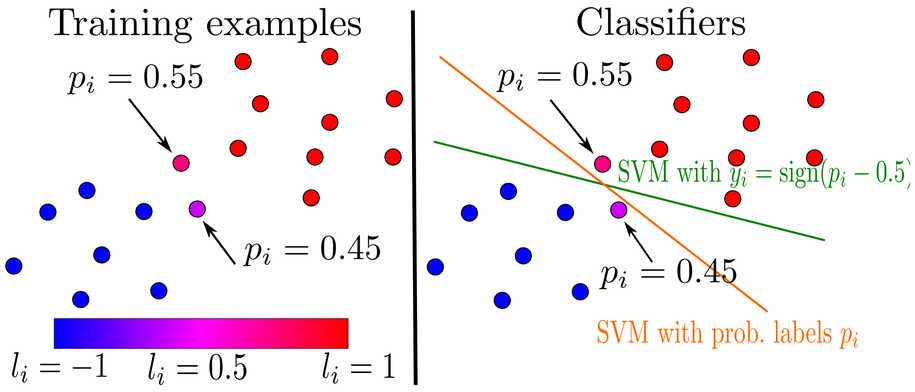
I. Motivation

Building an accurate training database is challenging in supervised classification. Radiologists often delineate malignant and benign tissues without access to the ground truth, thus leading to uncertain datasets. **We propose to deal with this uncertainty by introducing probabilistic labels in the learning stage** [1]. We introduce a probabilistic support vector machine (P-SVM) inspired from the regular C-SVM formulation allowing to consider class labels through a hinge loss and probability estimates using ε-insensitive cost function together with a minimum norm (maximum margin) objective. Solution is used for both decision and posterior probability estimation.

II. Problem formulation

Let $(x_i, l_i)_{i=1\dots m}$ be the learning dataset of input vectors $(x_i)_{i=1\dots m} \in X$ (the feature space) along with their labels $(l_i)_{i=1\dots m}$, such that

- **class labels**: $l_i{=}y_i \in \{-1, +1\}$ for $i = 1\dots n$ (certain labels),
- **real values**: $l_i{=}p_i = \mathbb{P}(Y_i = 1 \mid X_i = x_i) \in [0,1]$ for $i = n+1\dots m$ (uncertain labels).



III. Problem solution

Let k be a positive kernel satisfying Mercer's condition and H the associated reproducing kernel hilbert space. We propose the P-SVM pattern recognition problem [1]

$$\min_{f,b,\xi,\xi^-,\xi^+} \frac{1}{2}\|f\|_H^2 + C\sum_{i=1}^{n}\xi_i + \tilde{C}\sum_{i=n+1}^{m}(\xi_i^- + \xi_i^+)$$

subject to

$$\begin{cases} y_i(f(x_i)+b) \geq 1-\xi_i, & i{=}1\dots n \\ z_i^- - \xi_i^- \leq f(x_i)+b \leq z_i^+ + \xi_i^+, & i{=}n{+}1\dots m \\ 0 \leq \xi_i, & i{=}1\dots n \\ 0 \leq \xi_i^- \text{ and } 0 \leq \xi_i^+ & i{=}n{+}1\dots m \end{cases}$$

Following the idea of soft margin introduced in regular C-SVM, slack variables ξ_i measure the degree of misclassification of the datum x_i. C and $\tilde{C} \in \mathbb{R}^*$ control the relative weighting of classification and regression performances. Let ε be the labelling precision, δ be the confidence in the labelling and $\eta = \varepsilon + \delta$. The regression problem consists in finding optimal f such that

$$\left| \frac{1}{1+e^{-a(f(x_i)+b)}} - p_i \right| < \eta \ ,$$

thus constraining the probability prediction for point x_i to remain around to $\frac{1}{1+e^{-a(f(x_i)+b)}}$ within distance η [2, 3, 4]. This leads to $z_i^- {=} -\frac{1}{a}\ln(\frac{1}{p_i-\eta} - 1)$ and $z_i^+ {=} -\frac{1}{a}\ln(\frac{1}{p_i+\eta} - 1)$

Note that regular C-SVM is often associated with Platt's scaling algorithm [5] to estimate class probability membership whereas P-SVM makes it possible to directly estimate probabilities as $P(y=1|x) = \frac{1}{1+e^{-a(f(x)+b)}}$.

DUAL FORMULATION

Lagrange multipliers allow to rewrite the problem in its dual form

$$\begin{cases} \min\limits_{\Gamma} \ \frac{1}{2}\Gamma^\top G\Gamma - \tilde{e}^\top\Gamma \\ f^\top\Gamma = 0 \end{cases}, \text{ with}$$

$$f^\top = [y^\top, \underbrace{-1\cdots-1}_{\text{n-m times}}, \underbrace{1\dots1}_{\text{n-m times}}] \ ,$$

$$\tilde{e} = [\underbrace{1\dots1}_{\text{n times}} \underbrace{-z_{n+1}^+\cdots-z_m^+}_{\text{n-m times}} \underbrace{z_{n+1}^-\dots z_m^-}_{\text{n-m times}}] \ ,$$

$$0 \leq \Gamma \leq [\underbrace{C\dots C}_{\text{n times}} \underbrace{\tilde{C}\dots\tilde{C}}_{\text{n-m times}} \underbrace{\tilde{C}\dots\tilde{C}}_{\text{n-m times}}]^\top$$

$$G = \begin{pmatrix} K_1 & -K_2 & K_2 \\ -K_2^\top & K_3 & -K_3 \\ K_2^\top & -K_3 & K_3 \end{pmatrix}, \text{ with}$$

$$K_1 = (y_i y_j k(x_i, x_j))_{i,j=1\dots n}$$
$$K_2 = (k(x_i, x_j)y_i)_{i=1\dots n, j=n+1\dots m}$$
$$K_3 = (k(x_i, x_j))_{i,j=n+1\dots m}$$

The dual formulation is in the classical SVM form and is thus easy to implement.

V. Toy example

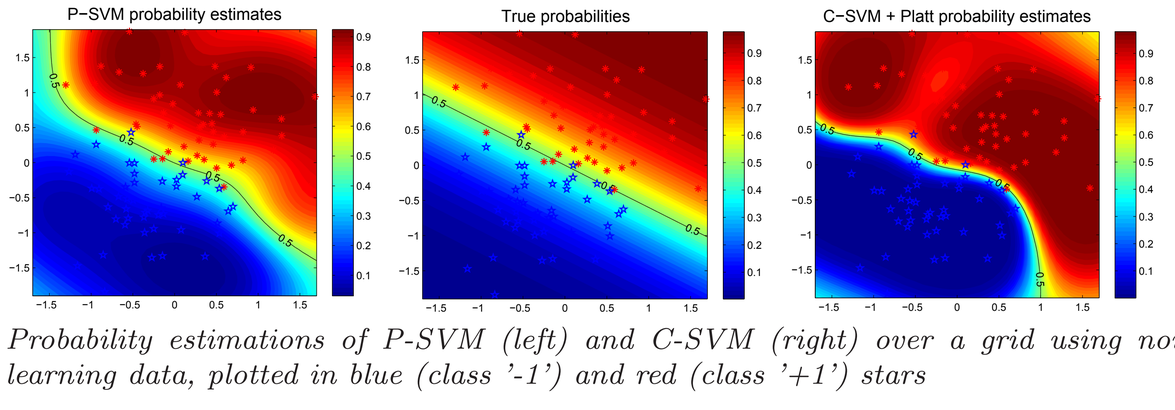
C-SVM versus P-SVM : performances are evaluated by computing the Accuracy (Acc), Kullback-Leibler distance (KL), Alignement (Align) and Mean Cross-Entropy (MCE). Implementation uses the SVM-KM Toolbox [6].

PROBABILITY ESTIMATION AND NOISE ROBUSTNESS

We generate two $\mathcal{N}(\mu, \sigma)$ 2D datasets, labelled '+1' and '-1' and compute the true '+1' class membership probability $P(y_i = 1|x_i)$ for each x_i of the learning data set (n=100). To simulate classification error, we add a uniform noise (amplitude 0.1) to probabilities, such that, for $i = 1\dots n$, $\hat{P}(y_i = 1|x_i) = P(y_i + 1|x_i) + \delta_i$. Learning data are labelled in two ways :

1) Dataset $(x_i, y_i)_{i=1\dots n}$, used to train C-SVM. For $i = 1\dots n$,
- if $\hat{P}(y_i{=}1|x_i){>}0.5$, then $y_i{=}1$,
- if $\hat{P}(y_i{=}1|x_i){\leq}0.5$, then $y_i{=}-1$

2) Dataset $(x_i, \breve{y}_i)_{i=1\dots n}$, used to train P-SVM. For $i = 1\dots n$,
- if $\hat{P}(y_i{=}1|x_i){>}1{-}\eta$, then $\breve{y}_i{=}1$,
- if $\hat{P}(y_i{=}1|x_i){<}\eta$, then $\breve{y}_i{=}-1$,
- otherwise $\breve{y}_i{=}\hat{P}(y_i{=}1|x_i)$.



Probability estimations of P-SVM (left) and C-SVM (right) over a grid using noisy learning data, plotted in blue (class '-1') and red (class '+1') stars

P-SVM classification and probability estimations obtained for 1000 test points are clearly more alike the ground truth ($\text{Acc}_{\text{P-SVM}}$=99%, $\text{KL}_{\text{P-SVM}}$=3.6) than C-SVM ($\text{Acc}_{\text{C-SVM}}$=95%, $\text{KL}_{\text{C-SVM}}$=95). C-SVM is sensitive to classification noise (no more convergence to the Bayes rule).

Conclusion

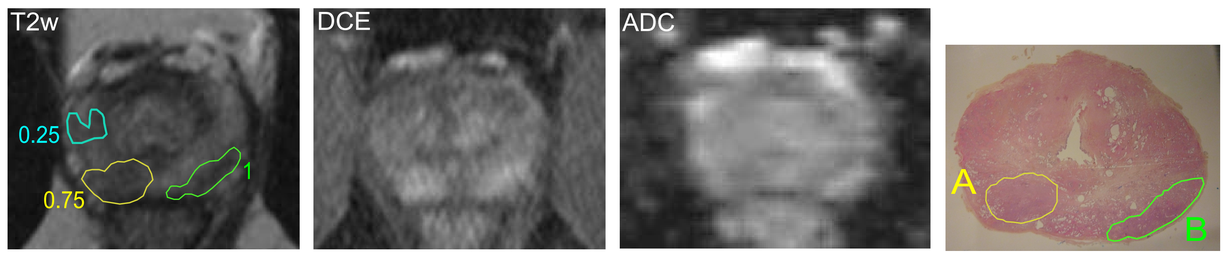
Training data used for computer-aided systems design often rely on expert's annotations, considered as the ground truth. Expert's uncertainty is rarely considered. We show that including these uncertainties into the learning step via P-SVM balances their influence and allows better predictions than those achieved with C-SVM.

References

[1] Niaf, *SSP*, 2011
[2] Rüping, *LWA* , 2004
[3] Grandvalet, *NIPS*, 2005
[4] Rueping, *ICML*, 2010
[5] Platt, *MIT Press*, 1999
[6] Canu, SVM-KM toolbox, 2005
[7] Niaf, *PMB Press*, 2012

VI. Clinical data

Database :
- Multiparametric MR images of the prostate acquired on 49 patients
- 350 regions of interest delineated and scored by experts using a 5-level scale of confidence from 1 = definitely malignant to 0 = definitely benign.
- Gold standard = Prostatectomy specimens (analysed *a posteriori*).



Prostate MRI : axial T2-weighted, DCE (after Gd-injection) and ADC MR images together with the corresponding histology slice. Histologically assessed cancers (A and B) were outlined on MR images (score 0.75 and 1 respectively) as well as a suspicious tissue (scored 0.25)

Objective : We compare the performances obtained using a P-SVM trained on expert's scores to those obtained using regular C-SVM trained on the same binarized database (i.e. score $> 0.5 \Rightarrow$ malignant, benign otherwise).

Evaluation : Performance achieved by both C-SVM and P-SVM using :

1) the expert's scores both as the training labelling and testing reference, thus assuming that the histologic ground truth is unknown.

2) the expert's scores as the training labelling and histology as the testing reference, thus evaluating if an expert's score-based database is accurate enough to predict true data class, thus possibly avoiding the tedious histology analysis.

3) the histology gold standard both as training labelling and test reference.

Leave-One-Patient-Out cross-validation using optimal parameters in [7].

Evaluation	on expert'scores				on ground truth			
	AUC	KL	Alignt	MCE	AUC	KL	Alignt	MCE
P-SVM	**.89**	**41**	**.25**	**.31**	**.86**	**73**	**.32**	**.33**
C-SVM	.85	76	.31	.38	.82	118	.38	.43
P-SVM/C-SVM learning on ground truth					.86	77	.31	.35

P-SVM systematically outperforms the classical C-SVM approach whatever training and testing database is used