HANDLING UNCERTAINTIES IN SVM CLASSIFICATION

Émilie NIAF[1,2]*, Rémi FLAMARY*[3]*, Carole LARTIZIEN*[2]*, Stéphane CANU*[3]

[1]INSERM U556, Lyon, 69424, France
[2]CREATIS, UMR CNRS 5220; INSERM U1044; INSA-Lyon; UCBL, Villeurbanne, 69621, France
[3]LITIS EA 4108, INSA-Universite de Rouen, Saint-Etienne-du-Rouvray, 76801, France

ABSTRACT

This paper addresses the pattern classification problem arising when available target data include some uncertainty information. Target data considered here is either qualitative (a class label) or quantitative (an estimation of the posterior probability). Our main contribution is a SVM inspired formulation of this problem allowing to take into account class label through a hinge loss as well as probability estimates using ε-insensitive cost function together with a minimum norm (maximum margin) objective. This formulation shows a dual form leading to a quadratic problem and allows the use of a representer theorem and associated kernel. The solution provided can be used for both decision and posterior probability estimation. Based on empirical evidence our method outperforms regular SVM in terms of probability predictions and classification performances.

Index Terms— support vector machines, maximal margin algorithm, uncertain labels.

1. INTRODUCTION

In the mainstream supervised classification scheme, an expert is required for labelling a set of data used then as inputs for training the classifier. However, even for an expert, this labeling task is likely to be difficult in many applications. In the end the training data set may contain inaccurate classes for some examples, which leads to non robust classifiers[1]. For instance, this is often the case in medical imaging where radiologists have to outline what they think are malignant tissues over medical images without access to the reference histopatologic information. We propose to deal with these uncertainties by introducing probabilistic labels in the learning stage so as to: 1. stick to the real life annotation problem, 2. avoid discarding uncertain data, 3. balance the influence of uncertain data in the classification process.

Our study focuses on the widely used Support Vector Machines (SVM) two-class classification problem [2]. This method aims a finding the separating hyperplane maximizing the margin between the examples of both classes. Several mappings from SVM scores to class membership probabilities have been proposed in the literature [3, 4]. In our approach, we propose to use both labels and probabilities as input thus learning simultaneously a classifier and a probabilistic output. Note that the output of our classifier may be transformed to probability estimations without using any mapping algorithm.

In section 2 we define our new SVM problem formulation (referred to as P-SVM) to deal with certain and probabilistic labels simultaneously. Section 3 describes the whole framework of P-SVM and presents the associated quadratic problem. Finally, in section 5 we compare its performances to the classical SVM formulation (C-SVM) over different data sets to demonstrate its potential.

2. PROBLEM FORMULATION

We present below a new formulation for the two-class classification problem dealing with uncertain labels. Let X be a feature space. We define $(x_i, l_i)_{i=1\dots m}$ the learning dataset of input vectors $(x_i)_{i=1\dots m} \in X$ along with their corresponding labels $(l_i)_{i=1\dots m}$, the latter of which being

- class labels: $l_i = y_i \in \{-1, +1\}$ for $i = 1 \dots n$ (in classification),
- real values: $l_i = p_i \in [0, 1]$ for $i = n+1 \dots m$ (in regression).

p_i, associated to point x_i allows to consider uncertainties about point x_i's class. We define it as the posterior probability for class 1.

$$p_i = p(x_i) = \mathbb{P}(Y_i = 1 \mid X_i = x_i).$$

We define the associated pattern recognition problem as

$$\min_{w} \quad \tfrac{1}{2}\|w\|^2 \qquad (1)$$

$$\text{subject to} \quad \begin{cases} y_i(w^\top x_i + b) \geq 1, & i = 1...n \\ z_i^- \leq w^\top x_i + b \leq z_i^+, & i = n+1...m \end{cases}$$

Where boundaries z_i^-, z_i^+ directly depend on p_i. This formulation consists in minimizing the complexity of the model while forcing good classification and good probability estimation (close to p_i). Obviously, if $n = m$, we are brought back to the classical SVM problem formulation.

Following the idea of soft margin introduced in regular SVM to deal with the case of inseparable data, we introduce

slack variables ξ_i. This measure the degree of misclassification of the datum x_i thus relaxing hard constraints of the initial optimization problem which becomes

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

subject to

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

Parameters C and $\tilde{C}$ are predefined positive real numbers controlling the relative weighting of classification and regression performances.
Let ε be the labelling precision and δ the confidence we have in the labelling. Let's define $\eta = \varepsilon + \delta$. Then, the regression problem consists in finding optimal parameters w and b such that

$$\mid \frac{1}{1+e^{-a(w^\top x_i+b)}} - p_i \mid < \eta\,,$$

Thus constraining the probability prediction for point x_i to remain around to $\frac{1}{1+e^{-a(w^\top x_i+b)}}$ within distance η [5, 6, 7]. The boundaries (where $w^\top x_i + b = \pm 1$), define parameter a as:

$$a = \ln(\tfrac{1}{\eta} - 1)$$

Finally:

$$\begin{array}{ccccc} \max(0, p_i - \eta) & \leq & \frac{1}{1+e^{-a(w^\top x_i+b)}} & < & \min(p_i + \eta, 1), \\ \Longleftrightarrow z_i^- & \leq & w^\top x_i + b & < & z_i^+, \end{array}$$

where $z_i^- = -\frac{1}{a}\ln(\frac{1}{p_i-\eta} - 1)$ and $z_i^+ = -\frac{1}{a}\ln(\frac{1}{p_i+\eta} - 1)$.

3. DUAL FORMULATION

We can rewrite the problem in its dual form, introducing Lagrange multipliers. We are looking for a stationary point for the Lagrange function L defined as

$$\begin{aligned} &L(w, b, \xi, \xi^-, \xi^+, \alpha, \beta, \mu^+, \mu^-, \gamma^+, \gamma^-) = \\ &\tfrac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}\xi_i + \tilde{C}\sum_{i=n+1}^{m}(\xi_i^- + \xi_i^+) \\ &- \sum_{i=1}^{n}\alpha_i(y_i(w^\top x_i + b) - (1 - \xi_i)) - \sum_{i=1}^{n}\beta_i\xi_i \\ &- \sum_{i=n+1}^{m}\mu_i^-((w^\top x_i + b) - (z_i^- - \xi_i^-)) - \sum_{i=n+1}^{m}\gamma_i^-\xi_i^- \\ &- \sum_{i=n+1}^{m}\mu_i^+((z_i^+ + \xi_i^+) - (w^\top x_i + b)) - \sum_{i=n+1}^{m}\gamma_i^+\xi_i^+ \end{aligned}$$

with $\alpha \geq 0$, $\beta \geq 0$, $\mu^+ \geq 0$, $\mu^- \geq 0$,$\gamma^+ \geq 0$ and $\gamma^- \geq 0$
Computing the derivatives of L with respect to w, b, ξ, ξ^- and ξ^+ leads to the following optimality conditions:

$$\begin{cases} 0 \leq \alpha_i \leq C, & i = 1...n \\ 0 \leq \mu_i^+ \leq \tilde{C}, & i = n+1...m \\ 0 \leq \mu_i^- \leq \tilde{C}, & i = n+1...m \\ w = \sum_{i=1}^{n}\alpha_i y_i x_i - \sum_{i=n+1}^{m}(\mu_i^+ - \mu_i^-)x_i & \\ y^\top\alpha = \sum_{i=n+1}^{m}(\mu_i^+ - \mu_i^-) & \end{cases}$$

where $e_1 = [\underbrace{1\ldots1}_{\text{n times}}\ \underbrace{0\ldots0}_{\text{(m-n) times}}\ ]^\top$ and $e_2 = [\underbrace{0\ldots0}_{\text{n times}}\ \underbrace{1\ldots1}_{\text{(m-n) times}}\ ]^\top$.
Calculations simplifications then lead to
$L(w, b, \xi, \xi^-, \xi^+, \alpha, \beta, \mu, \gamma^+, \gamma^-) =$

$$-\tfrac{1}{2}w^\top w + \sum_{i=1}^{n}\alpha_i + \sum_{i=n+1}^{m}\mu_i^- z_i^- - \sum_{i=n+1}^{m}\mu_i^+ z_i^+$$

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

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

where

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

with

$$\begin{aligned} K_1 &= (y_i y_j x_i^\top x_j)_{i,j=1...n}, \\ K_2 &= (x_i^\top x_j y_i)_{i=1...n, j=n+1...m}, \\ K_3 &= (x_i^\top x_j)_{i,j=n+1...m}, \end{aligned}$$

The dual formulation becomes

$$\begin{cases} \min\limits_{\Gamma} & \tfrac{1}{2}\Gamma^\top G\Gamma - \tilde{e}^\top\Gamma, \\ & f^\top\Gamma = 0 \\ \text{with} & \tilde{e} = [\underbrace{1\ldots1}_{\text{n times}}\ \underbrace{-z_{n+1}^+ \cdots - z_m^+}_{\text{n-m times}}\ \underbrace{z_{n+1}^- \ldots z_m^-}_{\text{n-m times}}] \\ \text{with} & f^\top = [y^\top, \underbrace{-1\cdots-1}_{\text{n-m times}}, \underbrace{1\ldots1}_{\text{n-m times}}\ ] \\ \text{and} & 0 \leq \Gamma \leq [\underbrace{C\ldots C}_{\text{n times}}\ \underbrace{\tilde{C}\ldots\tilde{C}}_{\text{n-m times}}\ \underbrace{\tilde{C}\ldots\tilde{C}}_{\text{n-m times}}\ ]^\top \end{cases} \quad (3)$$

4. KERNELIZATION

Formulations (2) and (3) can be easily generalized by introducing kernel functions. Let k be a positive kernel satisfying Mercer's condition and H the associated Reproducing Kernel Hilbert Space (RKHS). Within this framework equation (2) becomes

$$\min_{f,b,\xi,\xi^-,\xi^+} \frac{1}{2}\|f\|_H^2 + C\sum_{i=1}^{n}\xi_i + \tilde{C}\sum_{i=n+1}^{m}(\xi_i^- + \xi_i^+) \quad (4)$$

subject to

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

Formulation (3) remains identical, with

$$\begin{aligned} K_1 &= (y_i y_j k(x_i,x_j))_{i,j=1...n}, \\ K_2 &= (k(x_i,x_j)y_i)_{i=1...n,j=n+1...m}, \\ K_3 &= (k(x_i,x_j))_{i,j=n+1...m}, \end{aligned}$$

5. EXAMPLES

In order to experimentally evaluate the proposed method for handling uncertain labels in SVM classification, we have simulated different data sets described below. In these numerical examples, a RBF kernel $k(u,v) = e^{-\|u-v\|^2/2\sigma^2}$ is used and $C = \tilde{C} = 100$. We implemented our method using the SVM-KM Toolbox [8]. We compare the classification performances and probabilistic predictions of the C-SVM and P-SVM approaches. In the first case, probabilities are estimated by using Platt's scaling algorithm [3] while in the second case, probabilities are directly estimated via the formula defined in (2): $P(y=1|x) = \frac{1}{1+e^{-a(w^\top x+b)}}$. Performances are evaluated by computing

- Accuracy (Acc)
  Proportion of well predicted examples in the test set (for evaluating classification).
- Kullback Leibler distance (KL)
  $$D_{KL}(P||Q) = \sum_{i=1}^{n} P(y_i = 1|x_i) \log(\frac{P(y_i=1|x_i)}{Q(y_i=1|x_i)})$$
  for probability distributions P and Q (for evaluating probability estimation).

5.1. Probability estimation

We generate two unidimensional datasets, labelled '+1' and '-1', from normal distributions of variances σ^2_{-1}= σ^2_1=0.3 and means μ_{-1}=-0.5 and μ_1=+0.5. Let's $(x_i^l)_{i=1...n^l}$ denote the learning data set (n^l=200) and $(x_i^t)_{i=1...n^t}$ the test set (n^t=1000). We compute, for each point x_i, its true probability $P(y_i = +1|x_i)$ to belong to class '+1'. From here on, learning data are labelled in two ways, as follows

a) For $i = 1 \ldots n^l$, we get the regular SVM dataset by simply using a probability of 0.5 as the threshold for assigning class labels y_i associated to point x_i. This is what would be done in practical cases when the data contains class membership probabilities and a SVM classifier is used.

$$\begin{array}{llllll} \text{if} & P(y_i^l=1|x_i^l) & > & 0.5, & \text{then} & y_i^l = 1, \\ \text{if} & P(y_i^l=1|x_i^l) & \leq & 0.5, & \text{then} & y_i^l = -1 \end{array} \quad (5)$$

This dataset $(x_i^l, y_i^l)_{i=1...n^l}$ is used to train the C-SVM classifier.

b) We define another data set $(x_i^l, \hat{y}_i^l)_{i=1...n^l}$ such that, for $i = 1 \ldots n$,

$$\begin{array}{llllll} \text{if} & P(y_i^l=1|x_i^l) & > & 1-\eta, & \text{then} & \hat{y}_i^l = 1, \\ \text{if} & P(y_i^l=1|x_i^l) & < & \eta, & \text{then} & \hat{y}_i^l = -1, \\ & \hat{y}_i^l & = & P(y_i^l=1|x_i^l) & \text{otherwise.} & \end{array} \quad (6)$$

If the probability values are sufficiently close to 0 or 1 (closeness being defined by the precision and confidence), we admit that they belong respectively to class -1 or 1. This probabilistic dataset $(x_i^l, \hat{y}_i^l)_{i=1...n^l}$ is used to train the P-SVM algorithm.

We compare our two approaches using the test set $(x_i^t)_{i=1...n^t}$. As we know the true probabilities $(P(y_i^t=1|x_i^t))_{i=1...n^t}$, we can estimate the probability prediction error (KL). Figure 1 shows the probability predictions performances improvement shown by the P-SVM: the true probabilities (black) and P-SVM estimations (red) are quasi-superimposed (KL=0.2) whereas Platt's estimations are less accurate (KL=11.3).

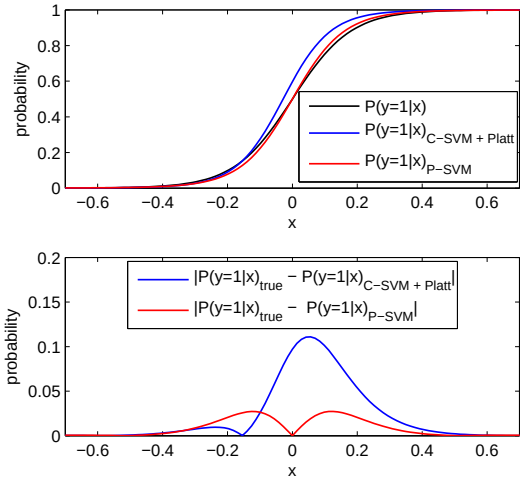

Fig. 1: Probability estimations comparison. Top plot shows the true posterior probabilities with C-SVM and P-SVM estimations overlaying. Lower plot shows the distance between true probabilities and estimations.

5.2. Noise robustness

We generate two 2D datasets, labelled '+1' and '-1', from normal distributions of variances σ^2_{-1}=σ^2_1=0.7 and means μ_{-1} = (-0.3, -0.5) and μ_1=(+0.3, +0.5). As in the previous experiment, we compute class '1' membership probability for each point x^l of the learning data set. We simulate classification error by artificially adding a centered uniform noise (δ of amplitude 0.1), to the probabilities, such that for $i = 1 \ldots n$,

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

We then label learning data following the same scheme as described in (5) and (6). Figure 2 shows the margin location and probabilities estimations using the two methods over a grid of values. Far from learning data points, both probability estimations are less accurate, this being directly linked to

the choice of a gaussian kernel. However, P-SVM classification and probability estimations obtained for 1000 test points, are clearly more alike the ground truth ($Acc_{P\text{-}SVM}$ = 99% , $KL_{P\text{-}SVM}$ = 3.6) than C-SVM ($Acc_{C\text{-}SVM}$ = 95%, $KL_{C\text{-}SVM}$ = 95). Contrary to P-SVM which, by combining both classification and regression, predicts good probabilities, C-SVM is sensitive to classification noise and is no more converging to the Bayes rule as seen in [1].

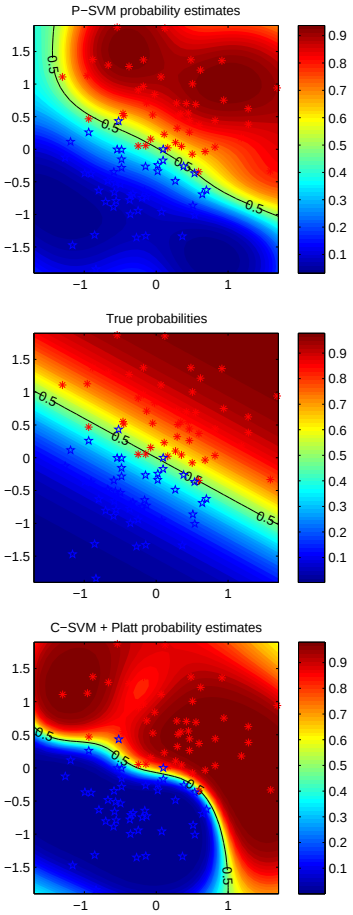

Fig. 2: Probability estimations of C-SVM and P-SVM over a grid using noisy learning data (uniform noise, amplitude 0.1). Noisy learning data are plotted in blue (class '-1') and red (class '1') stars.

Figure 3 shows the impact of noise amplitude on classifiers performances (values are averaged over 30 random simulations). Even if noise increases, classifications and probability predictions performances of the P-SVM remain significantly higher than those of C-SVM.

6. CONCLUSION

This paper has presented 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. This approach will soon be applied on clinical data thus allowing to assess its usefulness in computer assisted diagnosis for prostate cancer. Note that this framework initially designed for probabilistic labels can also be generalized to other dataset involving quantitative data as it can be used for instance to estimate a conditional cumulative distribution function.

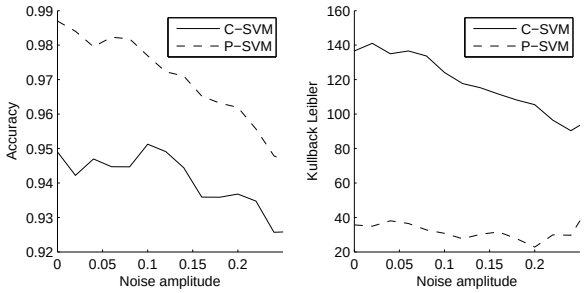

Fig. 3: Noise impact on P-SVM and C-SVM classification performances

7. REFERENCES

[1] G. Stempfel and L. Ralaivola, "Learning SVMs from Sloppily Labeled Data," *Artificial Neural Networks–ICANN 2009*, pp. 884–893, 2009.

[2] Bernhard Schölkopf and Alexander J. Smola, *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond (Adaptive Computation and Machine Learning)*, The MIT Press, 1st edition, December 2001.

[3] John C. Platt, "Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods," in *Advances in large margin classifiers*. 1999, pp. 61–74, MIT Press.

[4] Peter Sollich, "Probabilistic methods for support vector machines," in *Advances in Neural Information Processing Systems 12*. 2000, pp. 349–355, MIT Press.

[5] S. Rüping, "A Simple Method For Estimating Conditional Probabilities For SVMs," in *LWA 2004*, Bickel S. Brefeld U. Drost I. Henze N. Herden O. Minor M. Scheffer T. Stojanovic L. Abecker, A. and S. Weibelza hl, Eds., 2004.

[6] Y. Grandvalet, J. Mariéthoz, and S. Bengio, "A probabilistic interpretation of SVMs with an application to unbalanced classification," *Advances in Neural Information Processing Systems*, vol. 18, pp. 467, 2006.

[7] S. Rueping, "SVM Classifier Estimation from Group Probabilities," in *ICML 2010*.

[8] S. Canu, Y. Grandvalet, V. Guigue, and A. Rakotomamonjy, "Svm and kernel methods matlab toolbox," Perception Systèmes et Information, INSA de Rouen, Rouen, France, 2005.