SyMIL: MinMax Latent SVM for Weakly Labeled Data

Thibaut Durand, Nicolas Thome, Matthieu Cord

***Abstract*—Designing powerful models able to handle weakly labeled data is a crucial problem in machine learning. In this paper, we propose a new Multiple Instance Learning (MIL) framework. Examples are represented as bags of instances, but we depart from standard MIL assumptions by introducing a symmetric strategy (SyMIL) that seeks discriminative instances in positive and negative bags. The idea is to use the instance the most distant from the hyper-plan to classify the bag. We provide a theoretical analysis featuring the generalization properties of our model. We derive a large margin formulation of our problem, which is cast as a difference of convex functions, and optimized using CCCP. We provide a primal version optimizing with stochastic sub-gradient descent and a dual version optimizing with one-slack cutting-plane. Successful experimental results are reported on standard MIL and weakly-supervised object detection datasets: SyMIL significantly outperforms competitive methods (mi/MI/Latent-SVM), and gives very competitive performance compared to state-of-the-art works. We also analyze the selected instances of symmetric and asymmetric approaches on weakly-supervised object detection and text classification tasks. Finally we show complementarity of SyMIL with recent works on learning with label proportions on standard MIL datasets.**

***Index Terms*—Weakly Supervised Learning, Multiple Instance Learning, Latent SVM, Image Categorization and Pattern Recognition**

I. Introduction

LEARNING from weakly labeled data is a very important problem that covers several theoretical and practical aspects towards the development of powerful learning machines. On the one hand, relaxing the requirement of expensive manual and accurate annotations of training data offers the possibility to build large scale databases at reasonable cost. For example, in the computer vision field, annotating images with a global label makes it possible to build databases containing several millions of examples, whereas annotations at the pixel level (*i.e.* segmentation) are much more expensive which explains that only moderate-size datasets (around thousands of images) are available. On the other hand, handling weakly labeled data generally requires to expand the representation space with latent variables to model hidden factors and compensate for the weak supervision.

The literature in learning from weakly labeled data is very abundant. In this paper, we focus on the Multiple Instance Learning (MIL) paradigm: each example is represented as a bag of instances, and the weak supervision consists in providing a single label for each bag. This issue has been extensively studied during the last 15 years in several contexts: drug activity recognition in the seminal work of [1], text classification [2], content-based image retrieval [3], *etc*. The main MIL assumption is related to the relationship between bag and instance labels: a bag is positive if it contains at least one positive instance, and negative if it contains only negative instances. A classical toy example consists in viewing a bag as set of keys: a bag is labeled positive if it contains a key able to open the door, and negative if none of the keys can. The MIL approaches can be classified into two categories: bag [3]–[6] vs instance [2], [7]–[10] approaches. Bag approaches embed each bag into a feature space, where standard supervised learning techniques are used, whereas instance approaches learn a classification function in the instance space.

Another recent paradigm related to this instance-to-bag label issue is the Learning with Label Proportion (LLP) framework [11]–[14], which generalizes MIL. The authors show that LLP outperforms baseline MIL methods, by relaxing the common negative instances in negative bags assumption: a large portion of instances in a positive bag should be positive, whereas few instances in the negative bags may be positive. In LLP, only label ratios between $\oplus/\ominus$ instances in bags are provided during training.

Unlike the standard MIL framework and with a similar idea to that of LLP, we propose in this paper to model positive and negative bags in a symmetric manner (SyMIL): any bag, either positive or negative, must contain at least one correct instance. The proposed method casts the weakly supervised learning problem as an optimization scheme dedicated to identifying the most discriminative instances in each bag: discriminative instances correspond to maximum scoring values for positive bags, and to minimum scoring values for negative bags. To this end, we propose a novel learning framework based on a symmetric Latent SVM. We show that this novel MIL framework significantly improves predictive accuracy over state-of-the-art MIL methods in a variety of applications, from image to text and molecule classification.

II. State-of-the-art

The first MIL approach is certainly the work of [1], where the positive instances are iteratively estimated in the feature space using the hypothesis class of axis-parallel rectangles. Basically, as highlighted above, we can classify MIL approaches between bag and instance learning schemes.

An important class of MIL approaches correspond to methods that embed each bag into a feature space, where standard supervised learning techniques (*e.g.* SVM) can be applied. In

T. Durand and M. Cord are with the Sorbonne Université, CNRS, Laboratoire d'Informatique de Paris 6, 75005 Paris, France. E-mail: {thibaut.durand, matthieu.cord}@lip6.fr.

N. Thome is with the CEDRIC - Conservatoire National des Arts et Métiers, 292 rue St Martin, 75003 Paris, France. E-mail: nicolas.thome@cnam.fr

This research was supported by a DGA-MRIS scholarship.

this context, Diverse Density (DD) [3], [4] or set kernels [5] handle the weak labeling by designing a proper embedding or distance/similarity that captures the structure of the problem. Recently, eMIL [6] was introduced to represent a bag as an ellipsoid, and an algorithm specifically tailored for the MIL paradigm is introduced. These methods treat the instances as *i.i.d* samples, but in [15], the authors proposed to go beyond the *i.i.d* assumption by introducing two graph kernels mi/MI-graph that model the correlations between instances. Despite the appealing ability to model correlation between instances, these embedding methods somehow lack locality in the similarity and are prone to noise. More generally, they are not designed to seek discriminative instances in each bag. Note that a recent work [16] proposes to combine the embedding techniques to instance-based methods to further improve performances.

Because embedding methods are not designed to seek discriminative instances in each bag, other approaches directly learn a classification function in the instance space. In this context, a major issue is to decide how to treat bag instances during training and prediction. In the reference paper [2], two variants adapting the soft-margin SVM formulation to the MIL problem are proposed: mi-SVM and MI-SVM, that are formulated as a mixed-integer programs. In mi-SVM, the problem consists in labeling each instance in positive bags as positive or negative, whereas MI-SVM selects a single instance (denoted as "witness") in each positive bag. Both mi-SVM and MI-SVM result in non-convex problems due to the negative max function for positive examples. mi-SVM and MI-SVM are thus two heuristics to solve this complex problem, and they basically consist in alternating between solving a standard SVM problem for fixed labeled instances, and re-labeling instances for positive bags. They are strong baselines that have been extensively studied for the MIL problem. The MI-SVM inspired pioneer works for weakly labeled object detection in the computer vision community. Specifically, the latent SVM (LSVM) [17] solves a "MI-SVM-like" problem, where the instances correspond to sub-part positions of the putative object position. It is worth mentioning that LSVM slightly differs from MI-SVM in the optimization scheme, since only the maximum output latent variable is used for negative examples to solve LSVM optimization problem, whereas all negative instances are used for MI-SVM. Figure 1 shows the instances used during training, and the position of the hyperplane for the standard MIL approaches. For example, mi-SVM uses all instances for each bag, whereas LSVM uses only one instance per bag. Interesting adaptations of these SVM-like MIL algorithms have been proposed: a solution dedicated to sparse positive bags [7], using deterministic annealing to continuously approximate the problem [8], a convex relaxation with the soft-max loss function [10], modeling instance dependencies as in MI-CRF [9], or modeling the ambiguity over latent variables as in max-margin min-entropy models (M3E) [18].

Another recent paradigm related to this instance-to-bag label issue is the Learning with Label Proportion (LLP) framework [11]–[13]. In LLP, only label ratios between $\oplus/\ominus$ instances in bags are provided during training. Different methods have been developed to estimate the label proportion of each bag. [11] proposed a theoretically sound method to estimate the mean of each class using the mean of each bag and the label proportions. [12] proposed treating the mean of each bag as a "super-instance", which was assumed to have a soft label corresponding to the label proportion. [13] proposes a method that explicitly models the latent unknown instance labels together with the known label proportions in a large-margin framework. In [14] the LLP method of [13] is explicitly applied to MIL problems, in the context of video event detection. LLP can be regarded as a generalization of MIL, and is shown to outperform baseline methods (mi/MI-SVM), especially by its capacity to relax the assumption that all negative instances in negative bags are negatives.

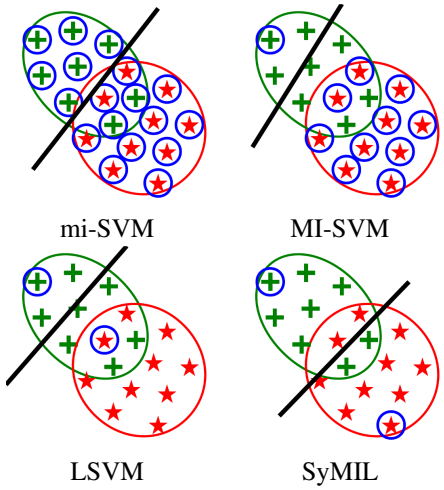


Fig. 1. Our approach seeks discriminative instances in positive (green) and negative (red) bags, whereas state-of-the-art methods consider all negative instances for learning(mi/MI-SVM), which may contain noisy features, or select the max scoring negative instance (LSVM), which may not be discriminative for the class of interest. Selected instances for training are surrounded in blue.

In this paper, we introduce a new method for solving MIL problems, with the following main contributions:

- We propose a new MIL model following LLP ideas, where the label proportion in positive and negative bags is set up in a symmetric manner (SyMIL). SyMIL requires having at least one correct instance in any positive or negative bag. SyMIL is represented with a latent variable model seeking the most discriminative instances, *i.e.* it seeks the instance which is the most distant from the hyperplane (see Figure 1). We also provide a theoretical analysis for SyMIL, highlighting its robustness to outliers.
- We derive an optimization based on concave-convex procedure (CCCP). We propose two different methods to solve the optimization problem in the primal with stochastic sub-gradient descent, and another to solve it in the dual with one-slack cutting-plane.
- We provide an experimental validation to assess the relevance of this symmetric modeling on standard MIL datasets. We analyze the selected instances for weakly supervised object detection and text classification. We also show the complementary between the local information in our symmetric modeling and the global bag statistics

in LLP.

III. SyMIL Model

First, we introduce the notations and the SyMIL model, then we propose a learning formulation scheme, and finally we provide a theoretical analysis.

Notations. We consider the problem of learning with weak supervision in a binary classification context. Training data are composed of n labeled bags $\mathcal{A}_n = \{(b_1, y_1), \ldots, (b_n, y_n)\} \in (\mathcal{X} \times \mathcal{Y})^n$ with input space $\mathcal{X}$ and $\mathcal{Y} = \{-1, +1\}$. Let us denote the set of positive bags as $\mathcal{A}_n^+ = \{(b_i, y_i), y_i = +1\}$ (with $n^+ = |\mathcal{A}_n^+|$), and the set of negative bags as $\mathcal{A}_n^- = \{(b_i, y_i), y_i = -1\}$ (with $n^- = |\mathcal{A}_n^-|$). Each bag b_i is itself a set of m_i instances, which are represented using latent variables $h \in \mathcal{H}_i$, with $|\mathcal{H}_i| = m_i$[1], and represented with a joint feature vector $\Psi(b_i, h) \in \mathbb{R}^d$. Using latent variables makes the definition of bag instances more general, including bags with an infinite number of instances or even continuous latent spaces.

A. Prediction function

Given an unlabeled bag b_i, we want to design a discriminant function $f_w : \mathcal{X} \to \mathbb{R}$, parametrized by w, such that $g(b) = \text{sign}\,[f_w(b_i)]$ gives predicted label of b_i: $f_w(b_i) > 0$ classifies the example as positive, and negative otherwise.

The main novelty of the SyMIL model is based on the definition of the latent variables h_i^+ and h_i^-:

$$h_i^+ = \arg\max_{h \in \mathcal{H}} \langle w, \Psi(b_i, h)\rangle \qquad h_i^- = \arg\min_{h \in \mathcal{H}} \langle w, \Psi(b_i, h)\rangle$$

h_i^+ (resp. h_i^-) is the maximum (resp. minimum) scoring latent value for the linear model $\langle w, \Psi(b, h)\rangle$. Using h_i^+ and h_i^-, we define the following prediction function:

$$f_w(b_i) = \begin{cases} \langle w, \Psi(b_i, h_i^+)\rangle & \text{if } \langle w, \Psi(b_i, h_i^+)\rangle \geq -\langle w, \Psi(b_i, h_i^-)\rangle \\ \langle w, \Psi(b_i, h_i^-)\rangle & \text{otherwise} \end{cases} \tag{1}$$

Model intuition & discussion. The rationale of the function $f_w(b_i)$ in Eq. (1) is to compare the score of the most '$\oplus$-like' instance (*i.e.* $\langle w, \Psi(b_i, h_i^+)\rangle$) to the score of the most '$\ominus$-like' instance (*i.e.* $-\langle w, \Psi(b_i, h_i^-)\rangle$). h_i^+ (resp. h_i^-) represents the most discriminative latent value for class $\oplus$ (resp. $\ominus$). During training, we aim at using h_i^+ (resp. h_i^-) for positive (resp. negative) bags, and so learning the most discriminative model.

From the LLP perspective, SyMIL corresponds to a symmetric prior for the label proportion: for a positive bag $(b_i, +1)$ (resp. negative bag $(b_i, -1)$), the proportion of positive (resp. negative) instances is $p_\oplus(b_i, +1) \geq \frac{1}{m_i}$ (resp. $p_\ominus(b_i, -1) \geq \frac{1}{m_i}$). This departs from state-of-the-art SVM-like MIL algorithms, *e.g.* mi/MI-SVM/ LSVM [2], [17]), where the prediction function takes the form $f_w(b_i) = \max_h \langle w, \Psi(b_i, h)\rangle$, corresponding to $p_\oplus(b_i, +1) \geq \frac{1}{m_i}$ but $p_\ominus(b_i, -1) = 1$. In this asymmetric modeling, $\oplus$ instances represent patterns that are discriminative for the $\oplus$ class, whereas $\ominus$ instances are implicitly regarded as background (*i.e.* everything different from $\oplus$ instances in the feature space). In contrast, SyMIL uses symmetric selection of instances in both positive and negative bags, which is supposed to be beneficial for classification, because instances shared by positive and negative classes (*i.e.* background) are ignored during training.

An illustrative comparison between symmetric and asymmetric MIL modeling is provided in Figure 2, for an image classification task. Here, bags represent images, and instances are rectangular image regions, in a simple two-class case (bison $\oplus$ *vs* llama $\ominus$). Basically, asymmetric models tend to learn a function discriminating bison patches from the most difficult patches in negative images, *i.e.* background patches of llama images. In contrast, for a bison (resp. llama) bag, the symmetric SyMIL model seeks regions that are statistically discriminant for bison (resp. llama) class, *i.e.* the instance the most distant from the hyperplane. SyMIL model tends to ignore background regions, *i.e.* those shared between $\oplus$ and $\ominus$ images. We validate this intuition of our model with toys experiments in Section V-A.

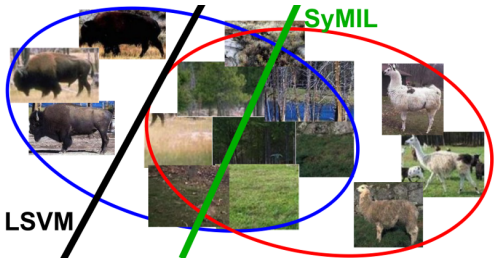


Fig. 2. SyMIL model motivation: symmetric *vs* asymmetric modeling between $\oplus$(blue)/$\ominus$(red) bags. Asymmetric model seeks discriminative regions for positive bags only, whereas symmetric model seeks discriminative regions for both positive and negative bags

B. Learning formulation

During the training step, we want to satisfy the following constraints:

$$\begin{aligned} &\forall i \in \mathcal{A}_n^+ \quad \langle w, \Psi(b_i, h_i^+)\rangle \geq 1 \\ &\forall i \in \mathcal{A}_n^- \quad \langle w, \Psi(b_i, h_i^-)\rangle \leq -1 \\ &\forall i \in \mathcal{A}_n \quad y_i \left[\langle w, \Psi(b_i, h_i^+) + \Psi(b_i, h_i^-)\rangle\right] \geq 1 \end{aligned} \tag{2}$$

The constraints given in Eq. (2) are interpreted as follows:

1) The first constraint $\langle w, \Psi(b_i, h_i^+)\rangle \geq 1$ enforces that the bag $b_i \in \mathcal{A}_n^+$ is properly classified in the class $\oplus$, using the latent value h_i^+, with a safety margin of 1. This is satisfied for the green positive bag in Figure 3.
2) The second constraint $\langle w, \Psi(b_i, h_i^-)\rangle \leq -1$ enforces that the bag $b_i \in \mathcal{A}_n^-$ is properly classified in the class $\ominus$, using the latent value h_i^-, with a safety margin of 1. This is satisfied for the red negative bag in Figure 3.
3) $y_i \left[\langle w, \Psi(b_i, h_i^+) + \Psi(b_i, h_i^-)\rangle\right] \geq 1$ enforces that each positive (resp. negative) bag is represented by h_i^+ (resp. h_i^-). For example, for $y_i = 1$, it translates into $\langle w, \Psi(b_i, h_i^+)\rangle \geq -\langle w, \Psi(b_i, h_i^-)\rangle + 1$, so that h_i^+ is preferred over h_i^- to represent b_i with a safety margin of 1, and $f_w(b_i) = \langle w, \Psi(b_i, h_i^+)\rangle$. In Figure 3, this constraint is satisfied for the positive green bag since $\Delta = \left(\langle w, \Psi(b_i, h_i^+) + \Psi(b_i, h_i^-)\rangle\right) \geq 1$. In a similar

[1]We ignore the dependence in i for $\mathcal{H}_i$ in the following.

fashion, this constraint is satisfied in Figure 3 for the negative red bag with $\Delta \leq -1$.

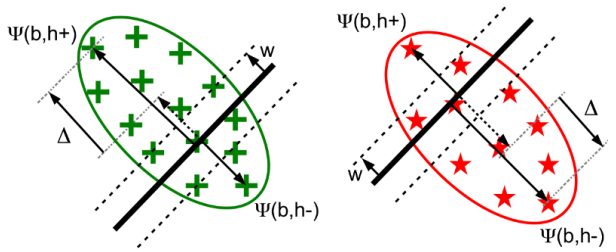


Fig. 3. Illustration of Eq. (2) constraints enforced during training for a positive (green) and a negative (red) bag. Dashed lines represent the safety margin of 1. $\Delta = \langle w, \Psi(b,h^+) + \Psi(b,h^-)\rangle$.

To optimize w over all the constraints of Eq. (2), the following primal regularized loss function $\mathcal{P}(w)$ is minimized:

$$\mathcal{P}(w) = \frac{1}{2}\|w\|^2 + \frac{C}{n}\left(\frac{n}{n^+}\sum_{i\in\mathcal{A}_n^+}\left[1-\max_{h\in\mathcal{H}}(\langle w,\Psi(b_i,h)\rangle)\right]_+ + \frac{n}{n^-}\sum_{i\in\mathcal{A}_n^-}\left[1+\min_{h\in\mathcal{H}}(\langle w,\Psi(b_i,h)\rangle)\right]_+ + \lambda\sum_{i\in\mathcal{A}_n}\left[1-y_i\left(\max_{h\in\mathcal{H}}(\langle w,\Psi(b_i,h)\rangle)+\min_{h\in\mathcal{H}}(\langle w,\Psi(b_i,h)\rangle)\right)\right]_+\right) \quad (3)$$

$\mathcal{P}(w)$ contains the standard max margin regularization term $\|w\|^2$ and a data-dependent term $E_l(w)$ penalizing the violation of Eq. (2) constraints using a hinge loss function $[b]_+ = \max(0,b)$. We derive the following property:

Lemma 1. *The loss function $E_l(w)$ in Eq. (3) defined over constraints $1-3$ in Eq. (2) is a surrogate of the $0/1$ loss on the prediction function $g(b) = sign\,[f_w(b)]$. (Proof in Appendix A).*

Connection to LSSVM [19]. Our prediction function $g(b) = \text{sign}\,[f_w(b)] = \text{sign}[\langle w,\Psi(b,h^+) + \Psi(b,h^-)\rangle]$ in Eq. (1) may be seen as an instantiation of the LSSVM prediction function. Indeed, with $\mathcal{Y} = \{-1;1\}$ and $\Phi(b,y,h) = y\cdot\Psi(b,h)$, we have $g(b) = \arg\max_{y\in\mathcal{Y}} \max_{h\in\mathcal{H}} \langle w,\Phi(b,y,h)\rangle$. Interestingly, our prediction function $g(b)$ is actually the natural instantiation of LSSVM to the binary classification case, which is not the case for competitive algorithms, *e.g.* mi/MI-SVM or LSVM. However, regarding learning formulation, $\mathcal{P}(w)$ in Eq. (3) differs from the LSSVM objective with the previous instantiation of $g(b)$, which would correspond to only incorporating the third constraint in Eq. (2). We add the constraints 1 & 2 of Eq. (2), because they correspond to the ultimate goal of the weakly supervised classifier: properly classifying (beyond the margin) training bags. We show in the experiments that adding these two constraints indeed favorably impacts classification performances.

Theoretical Analysis. We provide a bound of the average Rademacher complexity ($\mathcal{R}_n$) of SyMIL model. We note $\mathcal{F}$ the hypothesis class for instances and $\bar{\mathcal{F}}$ the hypothesis class for bags and we assume that the instances are in the hyper-sphere of radius B. To bound the average Rademacher complexity, we use the Theorem 20 of [20]. The bound in the general case is:

$$\mathcal{R}_n(\bar{\mathcal{F}},\mathcal{D}) \leq \frac{4+10\log(4ea_1^2a_2^2B^2rn^2)(N+\tau)}{\sqrt{n}} \quad (4)$$

where $\tau = \frac{a_1a_2}{\beta+1}K\ln^{\beta+1}(16a_1^2a_2^2n)$, n is the number of training examples, a_1 (resp. a_2) is the Lipschitz constant of bag-labeling (resp. loss) function, and N is a constant. The constant K and β must satisfy an inequality which depends on the worst-case Rademacher complexity over instances: $\mathcal{R}_n^{sup}(\mathcal{F}) \leq \frac{K\ln^\beta(n)}{\sqrt{n}}$. SyMIL learns a classification function in the instance space, so that the worst-case Rademacher complexity over instances is the same than SVM, *i.e.* $\mathcal{R}_n^{sup}(\mathcal{F}) \leq \frac{W}{\sqrt{n}}$[2] (proof in [21]), corresponding to $\beta = 0$ and $K = W$. Note that this bound is the same for LSVM, since both models use the same classification model over instances.

As mentioned when drawing the connection with LSSVM, SyMIL prediction function in Eq. (1) is equivalent to $g(b) = \text{sign}[\langle w,\Psi(b,h^+) + \Psi(b,h^-)\rangle]$. Therefore, the SyMIL bag-labeling function is 2-Lipschitz with respect to the infinity norm, because max and min are 1-Lipschitz with respect to the infinity norm. The loss function in Eq. (3) is $(1+2\lambda)$-Lipschitz. Therefore, by substituting a_1, a_2, β and K values, we get $\tau_{SyMIL} = 2(1+2\lambda)W\ln(64(1+2\lambda)^2n)$, leading to the following bound of the average Rademacher complexity:

$$\mathcal{R}_n(\bar{\mathcal{F}},\mathcal{D}) \leq \frac{4+10\log(16eB^2rn^2(1+2\lambda)^2)(N+\tau_{SyMIL})}{\sqrt{n}} \quad (5)$$

The resulting bound indicates that there is a poly-logarithmic dependence of the sample complexity on the average bag size. By comparing SyMIL bound with the LSVM one provided by [20], both bounds are similar and have the same order of magnitude[3]. Despite selecting the maximum or minimum instance, which introduces a non-linearity to the hypothesis class, this bound enables a control of the model complexity. It shows that the SyMIL and max prediction models have a bound with the same asymptotic behavior ($\frac{\ln(n)}{\sqrt{n}}$). In spite of the use of max and min instances, our model has similar robustness to outliers that max prediction models.

IV. Solving the Optimization Problem

Like competitive MIL algorithms (mi-SVM, MI-SVM, LSSVM), $\mathcal{P}$ in Eq. (3) is not a convex function of w. Without the first and the second constraints, our model is an instantiation of the LSSVM model [19]. As previously mentioned, these constraints are however important for optimal performances. In this section, we introduce our own solver to optimize Eq. (3).

A. Difference of Convex Functions

First, we show that $\mathcal{P}$ in Eq. (3) can be written as $\mathcal{P}(w) = u(w) - v(w)$, where u and v are convex functions. Rewriting

[2]In the SVM case, the class of functions is the set of linear separators with a bounded norm $\{x \mapsto \langle w,\Phi(x)\rangle : \|w\| \leq W\}$, for some $W > 0$.

[3]The difference with LSVM is that bag and loss functions are 1-Lipschitz

$\mathcal{P}(w)$ as a difference of convex functions is not straightforward given the form of Eq. (3). To demonstrate that $\mathcal{P}(w)$ (Eq. (3)) can be written as a difference of convex functions, we use the property:

$$\max(0, a-b) = \max(a, b) - b \tag{6}$$

where a, b are convex functions. We also use the properties that the maximum of a linear functions is a convex function, and the minimum of a linear functions is a concave function. Next, we will show that each hinge loss can be rewritten as a difference of convex functions. It is not straight-forward because each loss is neither a concave nor a convex function. For example, the first loss is the maximum of a concave function and a constant function, so it is neither a concave nor a convex function. But with the property (6), this loss can be written as a difference of convex functions. Each loss can be written as the difference of convex functions, so we can rewrite the global optimization problem $\mathcal{P}(w)$ as a difference of convex functions: $\mathcal{P}(w) = u(w) - v(w)$ where:

$$\begin{aligned} u(w) = & \frac{1}{2}\|w\|^2 + \frac{C}{n}\Bigg(\sum_{i \in \mathcal{A}_n^+} \Big[\frac{n}{n^+} \max(0, \max_{h \in \mathcal{H}}(\langle w, \Psi(b_i, h)\rangle)) - 1) \\ & + \lambda \ \max(1 - \min_{h \in \mathcal{H}}(\langle w, \Psi(b_i, h)\rangle), \max_{h \in \mathcal{H}} \langle w, \Psi(b_i, h)\rangle) \Big] \\ & + \sum_{i \in \mathcal{A}_n^-} \Big[\frac{n}{n^-} \max(0, -\min_{h \in \mathcal{H}}(\langle w, \Psi(b_i, h)\rangle) - 1) \\ & + \lambda \max(1 + \max_{h \in \mathcal{H}}(\langle w, \Psi(b_i, h)\rangle), -\min_{h \in \mathcal{H}} \langle w, \Psi(b_i, h)\rangle) \Big] \Bigg) \end{aligned} \tag{7}$$

$$\begin{aligned} v(w) = & \frac{C}{n}\Bigg(\sum_{i \in \mathcal{A}_n^+} \Big[(\frac{n}{n^+} + \lambda) \max_{h \in \mathcal{H}}(\langle w, \Psi(b_i, h)\rangle) - \frac{n}{n^+} \Big] \\ & + \sum_{i \in \mathcal{A}_n^-} \Big[-(\frac{n}{n^-} + \lambda) \min_{h \in \mathcal{H}}(\langle w, \Psi(b_i, h)\rangle) + \frac{n}{n^-} \Big] \Bigg) \end{aligned} \tag{8}$$

u and v are convex on w as a sum of convex functions.

B. Optimization

Once we exhibit the decomposition in difference of convex functions, we solve the resulting difference of convex functions using CCCP [22]. In addition to the CCCP convergence properties [23], this solution offers the possibility to jointly optimize some latent variables with the classifier parameters for each convex sub-problem, resulting in different (and better) local minima. The overall scheme of our CCCP-based optimization is presented in Algorithm 1: we alternate between linearizing the concave part ($-v$) at the current solution (Line 5) and solving the resulting convexified problem (Line 3). We now detail how the problem is solved in the primal and the dual.

1) Primal: The overall algorithm to train SyMIL with CCCP in the primal is given in Algorithm 2. CCCP is an iterative algorithm that alternates between linearizing the concave part ($-v$) at the current solution w_t (Line 5 of Algorithm 2) and solving the resulting convex problem (Line 3 of Algorithm 2). The linearization of the concave part $-v(w)$ consists in upper bounding it by its tangent hyperplane: $-v(w) \leq -\langle w, \nabla_w v(w_t)\rangle$, with:

Algorithm 1 for training SyMIL with CCCP

Require: training set $\{(b_i, y_i)\}_{i=1,\ldots,n}$
1: Set $t = 0$, randomly initialize $\{h_{i,0}^+, h_{i,0}^-\}_{i=1,\ldots,n}$ and linearize the concave part
2: **repeat**
3: Solve convex problem
$w_{t+1} = \text{argmin}_w \mathcal{P}_t^{CCCP}(w)$ or
$\alpha_{t+1} = \text{argmax}_\alpha \mathcal{D}_t^{CCCP}(\alpha)$
4: $t \leftarrow t + 1$
5: Linearize the concave part $-v$ at the current solution w_t / α_t
6: **until** stopping criteria reach
7: **return** w_t / α_t

$$\nabla_w v(w_t) = \Bigg(\sum_{i \in \mathcal{A}_n^+} (\tfrac{n}{n^+} + \lambda) \Psi(b_i, h_{i,t}^+) - \sum_{i \in \mathcal{A}_n^-} (\tfrac{n}{n^-} + \lambda) \Psi(b_i, h_{i,t}^-) \Bigg)$$

where $h_{i,t}^+ = \arg\max_{h \in \mathcal{H}} \langle w_t, \Psi(b_i, h)\rangle$ and $h_{i,t}^- = \arg\min_{h \in \mathcal{H}} \langle w_t, \Psi(b_i, h)\rangle$. After linearization, the resulting optimization problem is:

$$\mathcal{P}_t^{CCCP}(w) = u(w) - \langle w, \nabla_w v(w_t)\rangle \tag{9}$$

$$\begin{aligned} = & \frac{1}{2}\|w\|^2 + \frac{C}{n}\Bigg(\sum_{i \in \mathcal{A}_n^+} \Big[\frac{n}{n^+} \max(0, \max_{h \in \mathcal{H}} \langle w, \Psi(b_i, h)\rangle - 1) \\ & + \lambda \max(1 - \min_{h \in \mathcal{H}} \langle w, \Psi(b_i, h)\rangle, \max_{h \in \mathcal{H}} \langle w, \Psi(b_i, h)\rangle) \Big] \\ & + \sum_{i \in \mathcal{A}_n^-} \Big[\frac{n}{n^-} \max(0, -\min_{h \in \mathcal{H}} \langle w, \Psi(b_i, h)\rangle - 1) \\ & + \lambda \max(1 + \max_{h \in \mathcal{H}} \langle w, \Psi(b_i, h)\rangle, -\min_{h \in \mathcal{H}} \langle w, \Psi(b_i, h)\rangle) \Big] \Bigg) \\ & - \frac{C}{n}\Bigg(\sum_{i \in \mathcal{A}_n^+} (\frac{n}{n^+} + \lambda) \langle w, \Psi(b_i, h_{i,t}^+)\rangle - \sum_{i \in \mathcal{A}_n^-} (\frac{n}{n^-} + \lambda) \langle w, \Psi(b_i, h_{i,t}^-)\rangle \Bigg) \end{aligned} \tag{10}$$

At iteration t, to solve the convexified optimization problem $\min_w \mathcal{P}_t^{CCCP}(w)$ in the primal, we use a Stochastic (sub)Gradient Descent (SGD) strategy [24] that proves to be simple and achieves fast convergence. Although we could use more efficient techniques, such as SAG [25], we find SGD sufficient in our experiments (Section V). The gradient computation is given in Appendix B-A.

Algorithm 2 for training SyMIL with CCCP (Primal)

Require: training set $\{(b_i, y_i)\}_{i=1,\ldots,n}$
1: Set $t = 0$, randomly initialize $h_{i,0}^+$, $h_{i,0}^-$ and
$g_0 = \frac{C}{n}\Big(\sum_{i \in \mathcal{A}_n^+} (\frac{n}{n^+} + \lambda) \Psi(b_i, h_{i,0}^+) - \sum_{i \in \mathcal{A}_n^-} (\frac{n}{n^-} + \lambda) \Psi(b_i, h_{i,0}^-) \Big)$
2: **repeat**
3: Solve $w_{t+1} = \arg\min_w [u(w) - \langle w, g_t\rangle]$
4: $t \leftarrow t + 1$
5: Compute $g_t = \nabla_w v(w_t)$
6: **until** $[u(w_t) - v(w_t)] - [u(w_{t-1}) - v(w_{t-1})] < \varepsilon$
7: **return** w_t

Algorithm 3 Cutting plane algorithm with 1-slack formulation at iteration t

Require: training set $\{(b_i, y_i)\}_{i=1,\dots,n}$, $\{(h^+_{i,t}, h^-_{i,t})\}_{i=1,\dots,n}$

1: Set $T = 0$, $c \leftarrow 0$, $H \leftarrow 0$
2: **repeat**
3: $H \leftarrow (H_{ij})_{1\leq i,j\leq T}$ where $H_{ij} = g^T_{(i)} g_{(j)}$
4: $\alpha \leftarrow \arg\max_{\alpha} \alpha^T c - \alpha^T H \alpha \quad$ s.t. $0 \leq 1^T \alpha \leq C$
5: $\xi \leftarrow \frac{1}{C}(\alpha^T c - \alpha^T H \alpha)$
6: $T \leftarrow T + 1$
7: **for** $i = 1, \dots, n$ **do**
8: $h^+_i = \arg\max_{h\in\mathcal{H}} \langle \sum_{i=1}^{T-1} \alpha_i g_{(i)}, \Psi(b_i, h)\rangle$
9: $h^-_i = \arg\min_{h\in\mathcal{H}} \langle \sum_{i=1}^{T-1} \alpha_i g_{(i)}, \Psi(b_i, h)\rangle$
10: **end for**
11: $g_{(T)} \leftarrow \frac{1}{n}\sum_{i\in\mathcal{A}_n} g_{cave}(b_i, h^+_{i,t}, h^-_{i,t}) - g_{vex}(b_i, h^+_i, h^-_i)$
12: $c_{(T)} \leftarrow \frac{1}{n}\sum_{i\in\mathcal{A}_n} (vex(b_i, h^+_i, h^-_i) - \langle g_{vex}(b_i, h^+_i, h^-_i), \sum_{i=1}^{T-1} \alpha_i g_{(i)}\rangle)$
13: **until** $\langle \sum_{i=1}^{T-1} \alpha_i g_{(i)}, g_{(T)}\rangle \geq c_{(T)} - \xi - \varepsilon$
14: **return** α

2) Dual.: For many applications, nonlinear models are required to achieve good performances. We propose here a kernelized version of our SyMIL scheme. First, we detail the linearization of the concave part (Line 5 in Algorithm 1) in the dual, and then the solving of the convexified problem with cutting-plane.

Linearizing the concave part. To linearize the concave part at iteration $t+1$, we have to fix the latent variables. For a bag b_j, with current solution $\alpha^{(t)}$, the inference of the new latent variable value is:

$$h^+_{j,t+1} = \arg\max_{h\in\mathcal{H}} \Big\langle \sum_k \alpha^{(t)}_k \frac{1}{n} \sum_{i\in\mathcal{A}_n} (g_{cave}(b_i, h^+_{i,t}, h^-_{i,t}) - g_{vex}(b_i, h^+_i, h^-_i)), \Psi(b_j, h)\Big\rangle \quad (11)$$

$$h^-_{j,t+1} = \arg\min_{h\in\mathcal{H}} \Big\langle \sum_k \alpha^{(t)}_k \frac{1}{n} \sum_{i\in\mathcal{A}_n} (g_{cave}(b_i, h^+_{i,t}, h^-_{i,t}) - g_{vex}(b_i, h^+_i, h^-_i)), \Psi(b_j, h)\Big\rangle \quad (12)$$

where the gradient of the convex and concave terms are:

$$g_{cave}(b_i, h^+_{i,t}, h^-_{i,t}) = \begin{cases} (\frac{n}{n^+} + \lambda)\Psi(b_i, h^+_{i,t}) & \text{if } i \in \mathcal{A}^+_n \\ (\frac{n}{n^-} + \lambda)\Psi(b_i, h^-_{i,t}) & \text{if } i \in \mathcal{A}^-_n \end{cases} \quad (13)$$

$$g_{vex}(b_i, h^+_i, h^-_i) = \begin{cases} D + E & \text{if } i \in \mathcal{A}^+_n \\ F + G & \text{if } i \in \mathcal{A}^-_n \end{cases} \quad (14)$$

D, E, F, G are defined in equations (26 - 29) in Appendix B-A. $(h^+_{i,t}, h^-_{i,t})$ are the predicted latent variable for linearizing the concave part at iteration t.

Solving the convexified problem. A direct resolution of the convexified problem in the dual would be intractable, as for many other kernelized (latent) structured output problems. For our SyMIL model, the number of constraints in this dual formulation would be $\prod_{i=1}^n |\mathcal{H}_i|^2$, where $|\mathcal{H}_i|$ is the number of instances for the i^{st} training bag. Therefore, we adopt a cutting-plane strategy to train our SyMIL model, using the 1-slack formulation [26]. The learning algorithm is given in Algorithm 3. Cutting-plane training searches the optimal solution and the set of active constraints simultaneously in an iterative manner. This algorithm is guaranteed to converge to an approximate solution with a reasonable number of outer loops. Starting from an empty working set of constraints, in each iteration it solves the optimization problem (Line 4) with only the constraints of the working set. Then it finds the most violated constraint (Line 7-12) and adds it to the working set. $vex(b_i, h^+_i, h^-_i)$ is the convex term for bag b_i and the equation is given in Eq. (30) of Appendix B-B. The algorithm stops once no constraint can be found that is violated by more than the desired precision ε (Line 13). In our implementation, we use a stopping criterion defined by a fix number of iterations. During each iteration, we use MOSEK (www.mosek.com) to solve the quadratic problem with the given set of active constraints (Line 4).

V. Evaluation

We evaluate the symmetric approach on standard MIL datasets and for weakly-supervised object detection. We also analyze the selected instances and show the complementary of SyMIL with label proportion methods.

A. Toy Experiments

1) Synthetic data: First, we design toy datasets. We model positive and negative bags in a symmetric manner: instances are generated from two different Gaussian distributions, with a parameter α controlling the distance between them, and consequently the proportion of shared instances. (see Figure 4). The smaller the α, the more instances are shared between positives and negatives bags: the overlap region is thus a "background" area that contains "non-discriminative" instances for the classification task.

$\alpha = 0.1$ $\alpha = 0.3$ $\alpha = 0.5$ $\alpha = 0.7$

Fig. 4. Toy datasets generated with different α values. The blue points (resp. red) are the instances of the positive bags (resp. negative bags)

Our experimental setup is as follows. We generate 2d Gaussian distributions and sample bags with 20 instances, with 10 α values in the range $[0.1, 1]$. We train a linear model with 400 positive and 400 negative bags, evaluate the performance (accuracy) on other 100 test positive and negative bags, and average the results over 5 random folds.

Figure 5 shows the performance evolution when varying α for SyMIL and LSVM. For large overlap values ($\alpha \geq 0.6$), the classification task is easy and SyMIL and LSVM have similar

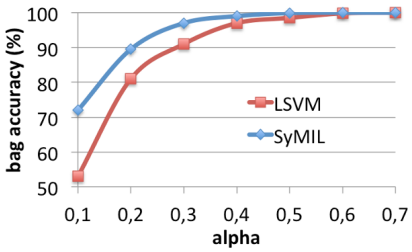
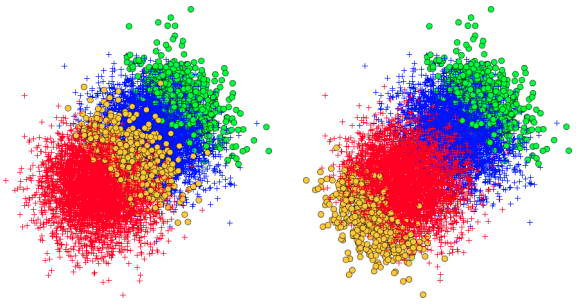


Fig. 5. Toy Examples: test accuracy with respect to α.



a) LSVM b) SyMIL

Fig. 6. Toy Examples: visualization of the selected instances during training ($\alpha = 0.5$): green over positive instances (in blue), orange over negative ones (in red).

performances. When the task becomes more challenging, ($\alpha \leq 0.3$), SyMIL outperforms by more than 5 pt LSVM.

To analyze the performance gain and interpret the latent representation learned by the different models, we visualize the instances selected by SyMIL and LSVM during training in Figure 6a) and 6b), respectively. We can notice that SyMIL selected instances are discriminative for the positive and negative class, *i.e.* they belong to a region in the feature space that is not shared between the two Gaussian distributions. On the contrary, the instances selected by LSVM for the negative class essentially belong to the background area, and are shared by the two classes. This experiment validates the relevance of our model for bag classification. Note that the computation time for LSVM and SyMIL is similar because the most time consuming step is the inference over the instances. In all our experiments, the inference problems are solved with an exhaustive search, so seeking the maximum instance or both maximum and minimum instances both take the same time.

2) Real data: We also validate our hypothesis on Mammal dataset. Mammal dataset [27] is a multi-class dataset, containing 6 categories: bison, deer, elephant, giraffe, llama and rhino. We use the same protocol as in [18]: the latent space $\mathcal{H}$ is composed of constant-size rectangular regions (which is a reasonable assumption for this dataset), and HOG descriptors [28] are used as features $\psi(b, h)$ for each region h. For this experiments, we use only the classes bison and llama. We report in Table I the classification results for bison vs llama (bison (resp. llama) is the positive (resp. negative) class) and llama vs bison (llama (resp. bison) is the positive (resp. negative) class). The performances are evaluated using a 10-fold cross-validation.

We note that our models has better results than LSVM. For bison vs llama experiment, SyMIL seeks discriminative regions for both bison and llama, while LSVM seeks discriminative regions only for bison. We note that reverse the positive and negative class gives different results for LSVM, because it use an asymmetric strategy, whereas our model gives the same results, because it use a symmetric strategy.

	bison vs llama	llama vs bison
LSVM [17]	90.3	87.7
SyMIL	95.7	95.7

TABLE I
CLASSIFICATION PERFORMANCES (ACCURACY) ON MAMMAL DATASET

B. Standard MIL Datasets

We demonstrate the efficiency of SyMIL on standard MIL datasets[4] with 3 different applications: molecule categorization, automatic image annotation, and text categorization. We start by giving details of datasets:

- Musk dataset: consists of descriptions of molecules using multiple low-energy conformations. Each conformation is represented by a 166-dimensional feature vector derived from surface properties.
- Image dataset: an image consists of a set of segments, each characterized by color, texture and shape descriptors. There are three different categories ("elephant", "fox", "tiger"). In each case, the dataset has 100 positive and 100 negative example images. The latter have been randomly drawn from a pool of photos of other animals. The original data are color images from the Corel dataset that have been preprocessed and segmented with the Blobworld system [29].
- Text dataset: starting from the publicly available TREC9 data set, each document is split into passages using overlapping windows of maximal 50 words each. Then, documents are annotated with MeSH terms (Medical Subject Headings), each defining a binary concept. The total number of MeSH terms in TREC9 is 4903. The first seven categories of the pre-test portion with at least 100 positive examples are used to create the dataset.

These datasets do not seem to be adapted for MIL, because the negative class is not everything. For each image dataset (Fox, Tiger, Elephant), positive bags are images that contain the animal, and negative bags are images that contain other animals (also from other categories, not just from the three categories here). Table II provides information about the number of training examples, the average number of instances per bag for each dataset, and the dimension of the features.

Dataset	Image	Musk1	Musk 2	Text
pos/neg bags	100/100	47/45	39/63	200/200
instances/bag	~ 6.5	5.17	64.69	~ 8
feature dimension	230	166	166	$\sim 66\,500^{(*)}$

TABLE II
DATASET STATISTICS. (*) THE FEATURES ARE SPARSE.

[4]The datasets used in this section are available online at http://www.cs.columbia.edu/andrews/mil/datasets.html

The parametrization for our method is the following. Regarding hyper-parameters (Eq. (3)), C is fixed to a large value (10^4). λ is chosen by cross-validation on the training set, on the range $\{0.1, 0.2, 0.5, 1\}$. We evaluate our method with linear and RBF kernels ($k(x, y) = exp(-\gamma\|x - y\|_2^2)$). The scale parameter γ for RBF kernels is determined as the mean pairwise instance distance on the training set. We also try to cross-validate the C and bandwidth parameters, but we do not observe significant differences. For all methods, the initial latent variables are randomly selected. We follow the standard protocol to evaluate performances [2]: the performances are evaluated using a 10-times 10-fold cross-validation.

Method	Image	Musk	Text	
a) re-implemented methods				
mi-SVM	73.4	84.5	81.6	
MI-SVM	75.5	81.7	80.3	
LSVM	74.4	82.7	80	
SyMIL linear	79.1	88.2	**84.8**	
RBF	**80.2**	89.2	-	
b) state-of-the art results				**Avg.**
SyMIL	**80.2**	89.2	**84.8**	**84.7**
MICA	73.9	87.5	82.3	81.2
MIGraph	76.1	**90**	-	
MI-CRF	78.5	86.7	-	
GP-WDA	79	88.4	83.2	83.5
eMIL	77	85.3	82.7	81.7
MILEAGE	77.7	-	-	

TABLE III
CLASSIFICATION ACCURACY (%) ON THE THREE DATASETS. BOLDFACED NUMBERS INDICATE BEST RESULTS.

The overall results for the three kind of datasets (image, text, molecule) are gathered in Table III. Detailed results for each dataset are provided in Table IV and V. A first comparison is given in Table IIIa) with methods the most closely connected to ours: mi-SVM/MI-SVM [2] and LSVM [17]. From a modeling point of view, these approaches basically differ from ours by the way instances are selected in positive and negative bags during training. We re-implement the three methods in order to compare the methods on the same splits. For mi-SVM and MI-SVM, we use linear kernels, that were reported to achieve optimal performances[5] [2]. One can notice SyMIL with linear kernel significantly outperforms mi-SVM, MI-SVM and LSVM: on average in the three types of data, there is a gain of about 4 pt over the best baseline. We perform paired t-test to assess the statistical significance of the difference in each dataset: numbers are given in Table XII and Table XIII in Appendix C. It turns out that SyMIL is statistically better than its competitors with a risk of 1% for all image and molecule datasets, and for all text datasets except TST2 (performance similar with MI-SVM) and TST1 (outperformed by LSVM). Note that even with a risk of 0.01% the improvement remains significant for 8 out of 12 datasets (except TST1, TST2, Musk1 and Tiger). These results clearly highlight the relevance of our model, *i.e.* the importance of seeking discriminative instances in both positive and negative bags.

Using non-linear kernels can further improve performances: ~ 1 pt increase in the image and molecule datasets. However, for the text datasets, the linear model outperforms RBF kernels. Note that this trend is conform to the results reported in GP-WDA [33]. We also evaluate the performance reached when using the LSSVM [19] instantiation corresponding to our prediction function, *i.e.* $\psi(b, y, h) = y \cdot \phi(b, h)$. As explained in Section III, the SyMIL learning scheme is different from this LSSVM instantiation, which translates in ignoring constraints 1&2 in Eq. (2). Results are provided in Table IV: we observe a performance drop between 1 and 3 pt depending on the dataset (Image-Molecule), and on the kernel type (linear *vs* RBF). For example, the superiority of our method is largely significant on Elephant (t-test validation with a risk of 5%).

An absolute performance comparison with recent state-of-the-art works is provided in Table IIIb). On average on the three datasets, our method outperforms all reported results[6]. Competitive approaches in these datasets include recent works such as MILEAGE [16], GP-WDA [33] which solves the MIL problem using Gaussian Processes, eMIL [6] or MI-CRF [9] or MIGraph [15]. Despite the complex models used by these strong competitors, SyMIL outperforms them in the image and text databases. In particular, we can notice the excellent performances for Elephant. Although our method remains very competitive on the Musk datasets, it is slightly outperformed by MIGraph. One explanation may be that MIL assumptions are better satisfied on this historical dataset. Note, however, that MIGraph performs poorly on the image dataset. We use the code available online[7] to perform paired t-test (Table XII). SyMIL is significantly better than MIGraph on the image dataset (risk 1%), and the performance is equivalent on the molecule dataset (risk 5%). To summarize, the excellent results for the three applications exhibit the capacity of our method to successfully handle various types of data. Note that the local information in SyMIL can be combined with a global bag feature, as done in MILEAGE [16] or MI-CRF.

Analysis of parameter λ. We also study the performances with respect to the parameter λ, which is an important parameter for SyMIL model. This parameter adjusts the trade-off between constraints 1 & 2 and constraint 3 during training. A large λ is similar to LSSVM (see section III-B) because the constraints 1 & 2 are negligible with respect to the constraint 3. Figure 7 shows the results on Musk2 and Elephant datasets, for a λ on the range [0.01, 1000]. We observe that the best results are for a lambda around 1 on Musk2, and 0.1 on Elephant. The optimal λ change for each dataset. Use a small λ leads to bad results because the model is not able to predict the relevant instance.

C. Weakly-supervised Object Detection

In weakly-supervised object detection, the goal is to learn a model which jointly classifies the image and localizes the object. Training data only have image-level labels indicating the presence/absence of each object category in an image. The exact object location in the image is unknown and is modeled as a latent variable h. We make experiments on two different datasets: Mammal dataset and PASCAL VOC 2007.

[5]Note that our re-implementation matches the results in reported in [2].

[6]SyMIL results are reported for RBF kernels in the image and molecules datasets, but for linear kernels in the text datasets. This is similar to the setup in [33], since linear models generally lead to better performances.

[7]see MiGraph webpage

method	Elephant	Fox	Tiger	average	Musk1	Musk2	Avg.
re-implemented method							
mi-SVM [2]	81.7±1.7	58.3±1.6	80.2±1.5	73.4	85.5±1.9	83.4±2.1	84.5
MI-SVM [2]	82.2±1.7	60.9±1.9	83.3±1.6	75.5	78.9±3.3	84.4±2.0	81.7
LSVM [17]	81.5±1.9	60.2±1.8	81.6±1.2	74.4	81.6±2.3	83.4±1.5	82.5
SyMIL							
linear	87.2±1.1	64.9±0.9	85.3±0.8	79.1	88.5±1.5	87.8±0.9	88.2
RBF	**88.2±0.7**	66.9±1.2	85.9±0.6	**80.3**	89.5±1.8	88.8±1.9	89.2
Without constraints 1 & 2							
linear	85.0±1.9	64.3±1.7	84.9±1.8	78.1	87.8±2.4	85.9±1.7	86.9
RBF	85.5±1.9	65.5±1.7	85.0±1.2	78.7	88.5±1.2	86.4±1.7	87.5
IAPR [1]	-	-	-	-	92.4[1]	89.2	-
DD [4]	-	-	-	-	88.9	82.5	85.7
EM-DD [30][2]	78.3	56.1	72.1	68.8	84.8	84.9	84.9
MI-kernel (Minimax) [31]	-	-	-	-	**91.6**	86.3	89.0
mi-SVM [2]	82.2	58.2	78.4	72.9	87.4	83.6	85.5
MI-SVM [2]	81.4	57.8	84.0	74.4	77.9	84.3	81.1
ALP-SVM [8]	82.8	65.7	85.2	77.9	86.5	86.1	86.3
AW-SVM [8]	81.9	63.3	82.7	76.0	85.7	83.4	84.6
MICA [32]	80.5	58.7	82.6	73.9	84.4	**90.5**	87.5
MIGraph [15]	85.1	61.2	81.9	76.1	90.0	90.0	**90.0**
miGraph [15]	86.6	61.6	86.0	78.1	88.9	90.3	89.6
MI-CRF [9]	85.0	**67.5**	83.0	78.5	88.0	85.3	86.7
Convex relaxation [10]	86.7	62.5	78.0	75.8	87.7	-	-
GP-WDA [33]	83.8	65.7	87.4	79.0	89.5	87.2	88.4
eMIL [6]	84.0	58.3	**88.8**	77.0	84.5	86.0	85.3
MILEAGE [16]	84.5	64.5	84.0	77.7	-	-	-

1 Musk1 test data set was used to tune IAPR parameters, see [30]. For this reason, we do not report IAPR average performance on Musk, which is not representative of the method potential.

2 The EM-DD results reported in [30] were obtained by selecting the optimal solution using the test data. We report here results published in [2] using the correct algorithm. See [3] pp.935.

TABLE IV
CLASSIFICATION PERFORMANCES (ACCURACY) ON THE IMAGE AND MOLECULE DATASETS. BOLDFACED NUMBERS INDICATE BEST RESULTS.

method	TST1	TST2	TST3	TST4	TST7	TST9	TST10	Avg.
re-implemented								
mi-SVM	94.4±0.5	78.7±1.4	86.7±0.6	82.9±0.6	81.4±0.6	67.4±1.0	79.7±1.3	81.6
MI-SVM	94.0±0.4	84.9±1.2	82.8±0.9	82.3±0.9	77.7±0.7	61.3±0.5	79.4±0.7	80.3
LSVM	**96.0±0.3**	78.9±0.8	85.6±0.8	81.2±0.8	76.9±0.4	61.9±1.3	79.3±0.5	80.0
SyMIL	94.3±0.2	84.3±0.5	**88.8±0.5**	87.1±0.7	**82.3±1.0**	**71.2±0.8**	**85.8±0.5**	**84.8**
EM-DD	85.8	84.0	69.0	80.5	75.4	65.5	78.5	77.0
mi-SVM	93.6	78.2	87.0	82.8	81.3	67.5	79.6	81.4
MI-SVM	93.9	84.5	82.2	82.4	78.0	60.2	79.5	80.1
MICA	94.5	85.0	86.0	**87.7**	78.9	61.4	82.3	82.3
GP-WDA	94.4	**85.3**	86.1	85.3	80.3	70.8	80.4	83.2
eMIL	95.9	79.2	86.8	84.0	80.4	69.0	83.4	82.7

TABLE V
CLASSIFICATION PERFORMANCES (ACCURACY) ON THE TEXT DATASETS. BOLDFACED NUMBERS INDICATE BEST RESULTS.

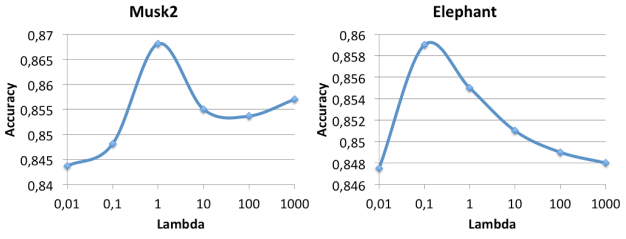


Fig. 7. Accuracy performance with respect to parameter λ (logarithmic scale) on Musk2 and Elephant datasets.

1) Mammal dataset: This dataset is presented in Section V-A2. To do multi class classification, we use 1vs All strategy. The performances are evaluated using a 10-fold cross-validation. We compare the proposed SyMIL RBF model to LSVM [17] and its recently kernelized version [34], using a RBF kernel. In addition, we evaluate M3E [18] by using the code available online[8]. The best performing M3E models use a small value of α, so we fix $\alpha = 5$ for our experiments. We measure prediction performances by using accuracy (ACC), and Mean Average Precision (MAP), to be robust to the $\oplus/\ominus$ unbalance.

	MAP (%)	ACC (%)
LSVM [17]	67.9	89.3
KLSVM [34]	73.3	90.1
M3E 1vsAll [18]	71.9	91.1
SyMIL	**78.7**	**92.1**

TABLE VI
CLASSIFICATION PERFORMANCES ON MAMMAL DATASET

The results are reported in Table VI. As we can see, our SyMIL model outperforms other approaches using both metrics. The trend is the same for both metrics (MAP & ACC).

[8]see M3E webpage.

In addition, all improvements are statistically significant (risk 5%), as validated by the t-tests provided in Table XIV of Appendix D. The prediction results again illustrate the superiority of the symmetric modeling, especially with respect to KLSVM [34] where the comparison directly measures the impact of the min/max selection strategy. Our method also has an edge over M3E, which tackles the weakly supervised learning in a direction complementary to ours (modeling ambiguities between latent variables).

2) VOC 2007: We perform another experiment on the PASCAL VOC 2007 dataset [35], which is the most famous object recognition benchmark used in computer vision. The dataset is composed of 10 000 images and 20 categories. For this experiment, we use our model as top layers of the `vgg-m-2048` deep ConvNet architecture pre-trained on ImageNet [36] and we only optimize the classification layer (no fine-tuning of the pre-trained layers). Our model behaves like a global pooling function which selects the most discriminative region for the class. In this architecture, each image is composed 25 regions and each region is represented by a 2048-dimensional vector (output of the 7-th layer after the ReLU). We follow the standard protocol [35] to evaluate the performances (Mean Average Precision).

The classification results are shown in Table VII. We compare SyMIL and LSVM. As observed in previous experiments, SyMIL outperforms LSVM. It confirms that seeking discriminative instances for both positive and negative class is relevant, even on challenging dataset like VOC 2007.

	LSVM [17]	SyMIL
Classification MAP (%)	76.21	78.37

TABLE VII
CLASSIFICATION PERFORMANCES ON PASCAL VOC 2007.

D. Further Analysis

To give additional insights on the symmetric MIL modeling introduced in this paper, we further analyze the selected instances on real image and text data.

1) Weakly-supervised Object Detection: We analyze the predicted instances (regions) for weakly-supervised object detection. We report localization performances to quantitatively evaluate the quality of the predicted latent values. We use the standard detection metric [35], measuring the overlap between the predicted and ground truth bounding boxes. We consider that a prediction is correct if the overlap is larger than 0.5.

	Train Ov.	Test Ov.	Train MAP	Test MAP
LSVM [17]	59.8	61.3	40.2	40.7
KLSVM [34]	60.9	60.8	39.9	40.1
M3E 1vsAll [18]	62.5	60.9	44.3	42.7
SyMIL	**64.7**	**63.2**	**47.6**	**46.5**

TABLE VIII
DETECTION PERFORMANCES ON MAMMAL DATASET

Table VIII summarizes the average performances for both detection metrics on Mammal dataset. SyMIL outperforms asymmetric approaches for both metrics. In addition, all improvements are statistically significant (risk 5%), as validated by the t-tests provided in Table XV of Appendix D. Detection results are connected to prediction performances. They quantitatively validate the motivation of the method illustrated in Figure 2, *i.e.* the fact that SyMIL is better able than asymmetric MIL models to track the structure of the negative class. In this dataset, we show that SyMIL successfully localizes regions containing object of the five categories composing the negative class. Visualizations of weakly supervised objects detection are given in Figure 8. LSVM tends to predict background regions, because of the asymmetry of the model, whereas SyMIL predicts foreground regions.

	Train (%)	Test (%)
LSVM [17]	36.38	41.99
SyMIL	42.71	43.42

TABLE IX
DETECTION PERFORMANCES ON PASCAL VOC 2007.

For VOC 2007, we normalize the overlap by the area of the predicted bounding box. Used the intersection over union score is not adapted, because we have only one size of box. If the ground truth bounding box is smaller than the size of the bounding boxes, it is not possible to have a good score even if the ground truth is in the predicted region. We observe similar results as on Mammal Dataset. SyMIL achieves better results for classification (+2,1%) and detection (+6% on training set) than LSVM (Table IX).

2) Text classification: We perform experiments on a text dataset from Reuters21578[9] to analyze selected instances for real data. We choose the category *money* as positive examples and *ship, crude* as negative. 100 documents from the 3 categories are randomly selected. Each document is a bag, and each paragraph is an instance. To represent each paragraph, we use tf-idf feature with vocabulary of size 18933. Performances are evaluated using a 10-times 5-fold cross-validation.

a) Predictive accuracy

	LSVM	**SyMIL**
	96.3%	97.6%

b) Similarity between instances and category

bag $\oplus$	74%	73%
bag $\ominus$	67%	78%

c) Examples

bag $\oplus$	bank, currency, money, exchange, treasury,	bank, exchange, rate, currency, monetary
bag $\ominus$	west, finance, bank british, money	oil, opec, shipping port, union

TABLE X
INSTANCE SELECTION FOR TEXT CLASSIFICATION: A) PREDICTIVE ACCURACY B) WORDS IN SELECTED INSTANCES WHICH ARE A SEMANTICALLY CORRELATED TO THE CATEGORY AND C) EXAMPLE OF TOP 5 SELECTED WORDS.

Results given in Table X show that SyMIL outperforms LSVM in terms of predictive accuracy (97.6% *vs* 96.3%). To analyze the instances selected by the two models, we compute the semantic similarity between the words in the selected instances and the related category, using Wu and Palmer (WP) similarity measure [37] on WordNet[10]. More

[9] http://kdd.ics.uci.edu/databases/reuters21578/reuters21578.html
[10] http://wordnet.princeton.edu

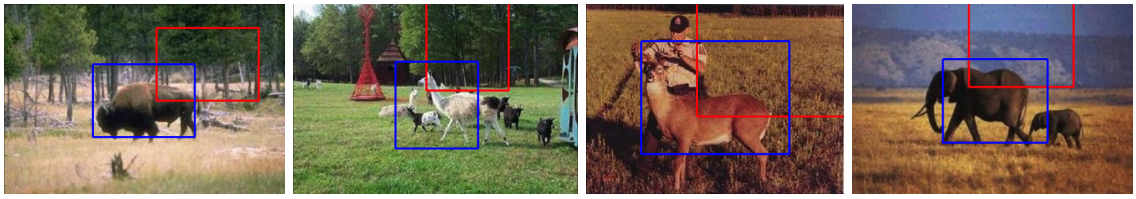


Fig. 8. Visualization of predicted latent variable for negative examples: on Mammal Dataset LSVM (red) and SyMIL (blue)

precisely, the similarity is determined by computing the ratio of words that have a WP similarity with respect to the category larger than a threshold Δ (set here to 0.2). For negative bags, we use the maximum similarity between ship or crude. In Table Xb), we notice that LSVM and SyMIL perform similarly (74% *vs* 73 %) for positive bags, whereas SyMIL is much better than LSVM for negative bags (78% *vs* 67 %). This highlights the superiority of the symmetric modeling to select instances which are representative of the negative class. Finally, Table Xc) shows an example of the 5 words that mostly contribute to the decision function. The top 5 selected words are generated as follows: for each selected instance (*i.e.* paragraph) we compute the top 5 words (i.e. dimensions in the instance space) that mostly contribute to the classification score (largest components of $|w|$), and average over all positive/negative bags. More precisely, for word k, we compute a histogram of contribution $w[k] \times \Psi(x, h^{predict})[k]$. We can point out that SyMIL extracts words that are semantically in touch with the negative class, *e.g.* (oil, OPEC) for crude and (port, shipping) for ship. On the contrary, LSVM selected words are not always semantically meaningful for the negative class, and are even more related to the positive class (money). Seeking discriminative instances for both positive and negative class is more robust than seeking discriminative instances for only the positive negative class. This analysis confirms the toy experiments conclusions of Section V-A.

VI. Conclusion

We introduced SyMIL, a new model for learning from weakly labeled data. Following LLP ideas, SyMIL departs from standard MIL assumptions by modeling positive and negative bags in a symmetric manner. The resulting latent variable model is trained by defining a regularized large margin objective function, which is minimized using CCCP. In addition, we derive a generalization error bound based on the Rademacher complexity. Experiments on various datasets validate the relevance of the proposed model, and an analysis of the SyMIL instance selection strategy reveals the capacity of the symmetric modeling to track the structure of the negative class. To have more robust prediction function, it would be interested to use several instances.

Appendix A
Proof of Lemma 1

In this section, we show that the loss function $E_l(w)$ in Eq. (3) is a surrogate of the $0/1$ loss our the prediction function $g(b) = \text{sign}\left[f_w(b)\right]$ - $f_w(b)$ defined in Eq. (1). We recall that $E_l(w)$ penalizes the violation of constraints 1-3 in Eq. (2), *i.e.*:

$$E_l(w) = \frac{C}{n}\left(\frac{n}{n^+}\sum_{i\in\mathcal{A}_n^+}\left[1-\max_{h\in\mathcal{H}}(\langle w,\Psi(b_i,h)\rangle)\right]_+ \right. \tag{15}$$
$$+\frac{n}{n^-}\sum_{i\in\mathcal{A}_n^-}\left[1+\min_{h\in\mathcal{H}}(\langle w,\Psi(b_i,h)\rangle)\right]_+$$
$$\left.+\lambda\sum_{i\in\mathcal{A}_n}\left[1-y_i\left(\max_{h\in\mathcal{H}}(\langle w,\Psi(b_i,h)\rangle)+\min_{h\in\mathcal{H}}(\langle w,\Psi(b_i,h)\rangle)\right)\right]_+\right)$$

$$=\frac{C\lambda}{n}\left(\frac{n}{\lambda n^+}\sum_{i\in\mathcal{A}_n^+}\left[1-\max_{h\in\mathcal{H}}(\langle w,\Psi(b_i,h)\rangle)\right]_+ \right. \tag{16}$$
$$+\frac{n}{\lambda n^-}\sum_{i\in\mathcal{A}_n^-}\left[1+\min_{h\in\mathcal{H}}(\langle w,\Psi(b_i,h)\rangle)\right]_+$$
$$\left.+\sum_{i\in\mathcal{A}_n}\left[1-y_i\left(\max_{h\in\mathcal{H}}(\langle w,\Psi(b_i,h)\rangle)+\min_{h\in\mathcal{H}}(\langle w,\Psi(b_i,h)\rangle)\right)\right]_+\right)$$

Let us denote $L_{SyMIL}(b_i, y_i)$ the loss for a single training sample with our objective function, and $\hat{y}_i = sign(f_w(x_i))$ the label predicted by our SyMIL model.

- if $y_i = +1$ and $\langle w, \Psi(x_i, h+)\rangle \geq -\langle w, \Psi(x_i, h-)\rangle$, then $\hat{y}_i = +1$ and $\Delta(y_i, \hat{y}_i) = 0$

$$L_{SyMIL}(b_i, y_i) = \frac{n}{\lambda n^+}\underbrace{\left[1-\langle w,\Psi(b_i,h^+)\rangle\right]_+}_{\geq 0} \tag{17}$$
$$+\underbrace{\left[1-\langle w,\Psi(b_i,h^+)+\Psi(b_i,h^-)\rangle\right]_+}_{\geq 0}$$
$$\geq \Delta(y_i,\hat{y}_i) \tag{18}$$

- if $y_i = +1$ and $\langle w, \Psi(x_i, h+)\rangle \leq -\langle w, \Psi(x_i, h-)\rangle$, then $\hat{y}_i = -1$ and $\Delta(y_i, \hat{y}_i) = 1$

$$L_{SyMIL}(b_i, y_i) = \frac{n}{\lambda n^+}\underbrace{[1-\langle w,\Psi(b_i,h^+)\rangle]_+}_{\geq 0} \tag{19}$$
$$\underbrace{+\,[1-\underbrace{\langle w,\Psi(b_i,h^+)+\Psi(b_i,h^-)\rangle}_{\leq 0}]_+}_{\geq 1}$$
$$\geq \Delta(y_i,\hat{y}_i) \tag{20}$$

- if $y_i = -1$ and $\langle w, \Psi(x_i, h+)\rangle \geq -\langle w, \Psi(x_i, h-)\rangle$, then $\hat{y}_i = 1$ and $\Delta(y_i, \hat{y}_i) = 1$

$$L_{SyMIL}(b_i, y_i) = \frac{n}{\lambda n^-} \underbrace{[1 + \langle w, \Psi(b_i, h^-)\rangle]_+}_{\geq 0} + \underbrace{[1 + \underbrace{\langle w, \Psi(b_i, h^+) + \Psi(b_i, h^-)\rangle}_{\geq 0}]_+}_{\geq 1} \tag{21}$$

$$\geq \Delta(y_i, \hat{y}_i) \tag{22}$$

- if $y_i = -1$ and $\langle w, \Psi(x_i, h+)\rangle < -\langle w, \Psi(x_i, h-)\rangle$, then $\hat{y}_i = -1$ and $\Delta(y_i, \hat{y}_i) = 0$

$$L_{SyMIL}(b_i, y_i) = \frac{n}{\lambda n^-} \underbrace{\left[1 + \langle w, \Psi(b_i, h^-)\rangle\right]_+}_{\geq 0} + \underbrace{\left[1 + \langle w, \Psi(b_i, h^+) + \Psi(b_i, h^-)\rangle\right]_+}_{\geq 0} \tag{23}$$

$$\geq \Delta(y_i, \hat{y}_i) \tag{24}$$

We thus have shown that in all cases, our loss $L_{SyMIL}(b_i, y_i)$ is a surrogate of the $0/1$ loss. Another way to prove it is to see that the third constraint in Eq. (2) is an instantiation of LSSVM [19] for binary classification. Therefore, the loss defined on this constraint is a surrogate of the 0/1 loss (by design in LSSVM). Since we add a positive loss on the first two constraints, E_l is necessarily a surrogate of the 0/1 loss.

APPENDIX B
OPTIMIZATION

A. Primal

In this section, we give the gradient for training SyMIL with SGD. For any randomly sampled training data (b_i, y_i), w is updated using the partial sub-gradient of Eq. (10) with respect to (b_i, y_i):

$$\nabla_w \mathcal{P}_t^{CCCP}(w) = \begin{cases} w + \frac{C}{n}(D + E - (\frac{n}{n^+} + \lambda)\Psi(b_i, h_{i,t}^+)) \\ \qquad\qquad \text{if } y_i = +1 \\ w + \frac{C}{n}(F + G + (\frac{n}{n^-} + \lambda)\Psi(b_i, h_{i,t}^-)) \\ \qquad\qquad \text{otherwise} \end{cases} \tag{25}$$

$$D = \begin{cases} \frac{n}{n^+}\Psi(b_i, h_i^+) & \text{if } \langle w, \Psi(b_i, h_i^+)\rangle - 1 > 0 \\ 0 & \text{otherwise} \end{cases} \tag{26}$$

$$E = \begin{cases} -\lambda\Psi(b_i, h_i^-) & \text{if } 1 - \langle w, \Psi(b_i, h_i^-)\rangle > \langle w, \Psi(b_i, h_i^+)\rangle \\ \lambda\Psi(b_i, h_i^+) & \text{otherwise} \end{cases} \tag{27}$$

$$F = \begin{cases} -\frac{n}{n^-}\Psi(b_i, h_i^-) & \text{if } -\langle w, \Psi(b_i, h_i^-)\rangle - 1 > 0 \\ 0 & \text{otherwise} \end{cases} \tag{28}$$

$$G = \begin{cases} \lambda\Psi(b_i, h_i^+) & \text{if } 1 + \langle w, \Psi(b_i, h_i^+)\rangle > -\langle w, \Psi(b_i, h_i^-)\rangle \\ -\lambda\Psi(b_i, h_i^-) & \text{otherwise} \end{cases} \tag{29}$$

B. Dual

In this section, we give the equation of the convex term for a bag b_i: $vex(b_i, h_i^+, h_i^-) =$

$$\begin{cases} \frac{n}{n^+}\max(0, \langle w, \Psi(b_i, h_i^+)\rangle - 1) \\ \quad +\lambda\max(1 - \langle w, \Psi(b_i, h_i^-)\rangle, \langle w, \Psi(b_i, h_i^+)\rangle) \\ \qquad\qquad \text{if } i \in \mathcal{A}_n^+ \\ \frac{n}{n^-}\max(0, -\langle w, \Psi(b_i, h_i^-)\rangle - 1) \\ \quad +\lambda\max(1 + \langle w, \Psi(b_i, h_i^+)\rangle, -\langle w, \Psi(b_i, h_i^-)\rangle) \\ \qquad\qquad \text{if } i \in \mathcal{A}_n^- \end{cases} \tag{30}$$

APPENDIX C
ADDITIONAL RESULTS ON STANDARD MIL DATASETS

Finally, we provide paired t-test to assess the statistical significance of the performance difference of our method compared to its competitors (mi/MI-SVM, LSVM), on each datasets. Table XI reports the critical values for different risks. The results of paired t-test on image and molecule (resp. text) datasets are report in Table XII (resp. Table XIII).

risk	5%	1%	0.1%	0.01%
t_{crit}	2.26	3.25	4.78	6.59

TABLE XI
CRITICAL VALUES FOR DIFFERENT RISKS (N=10)

method	**Eleph.**	**Fox**	**Tiger**	**Musk1**	**Musk2**
SyMIL / mi-SVM	9.24	21.52	11.16	5.92	10.22
SyMIL / MI-SVM	7.62	8.76	4.00	10.28	7.19
SyMIL / LSVM	10.97	7.99	7.69	14.44	10.22
SyMIL / miGraph	6.39	6.04	5.81	-0.18	-1.31

TABLE XII
PAIRED T-TEST RESULTS ON THE MOLECULE AND IMAGE DATASETS

method	**TST1**	**TST2**	**TST3**	**TST4**	**TST7**	**TST9**	**TST10**
mi-SVM	-0.52	11.88	7.56	18.90	2.74	9.11	11.74
MI-SVM	2.90	-1.44	22.58	11.27	11.62	47.70	27.41
LSVM	-16.30	25.55	11.68	14.91	14.61	24.41	28.71

TABLE XIII
PAIRED T-TEST RESULTS ON THE TEXT DATASETS BETWEEN SYMIL AND OTHERS METHODS

APPENDIX D
ADDITIONAL RESULTS ON MAMMAL DATASET

The significant tests for classification (resp. detection) are in Table XIV (resp. XV), and the critical values are given in Table XVI. SyMIL is significantly better than standard MIL approaches in both cases.

method	LSVM	KLSVM	M3E
ACC - SyMIL	8,77	6.78	2.67
MAP - SyMIL	7.32	3.09	4.42

TABLE XIV
MAMMAL DATASET: SIGNIFICANT TESTS FOR CLASSIFICATION

method	LSVM	KLSVM	M3E
SyMIL - Train	29.3	24.2	8.6
SyMIL - Test	13.2	14.5	9.3

TABLE XV
MAMMAL DATASET: SIGNIFICANT TESTS FOR DETECTION

risk	5%	1%	0.1%
t_{crit}	2.26	3.25	4.78

TABLE XVI
CRITICAL VALUES FOR DIFFERENT RISKS

REFERENCES

[1] T. G. Dietterich, R. H. Lathrop, and T. Lozano-Pérez, "Solving the multiple instance problem with axis-parallel rectangles," *Artif. Intell.*, vol. 89, no. 1-2, pp. 31–71, Jan. 1997.

[2] S. Andrews, I. Tsochantaridis, and T. Hofmann, "Support vector machines for multiple-instance learning," in *Advances in Neural Information Processing Systems 15*. MIT Press, 2003, pp. 561–568.

[3] Y. Chen and J. Z. Wang, "Image categorization by learning and reasoning with regions," *J. Mach. Learn. Res.*, vol. 5, pp. 913–939, Dec. 2004.

[4] O. Maron and A. L. Ratan, "Multiple-instance learning for natural scene classification," in *International Conference on Machine Learning (ICML)*, 1998.

[5] T. Gärtner, P. A. Flach, A. Kowalczyk, and A. J. Smola, "Multi-instance kernels," in *International Conf. on Machine Learning*. Morgan Kaufmann, 2002.

[6] G. Krummenacher, C. S. Ong, and J. Buhmann, "Ellipsoidal multiple instance learning," in *International Conference on Machine Learning (ICML)*, 2013.

[7] R. C. Bunescu and R. J. Mooney, "Multiple instance learning for sparse positive bags," in *International Conference on Machine Learning (ICML)*, 2007.

[8] P. Gehler and O. Chapelle, "Deterministic annealing for multiple-instance learning," in *Artificial Intelligence and Statistics*, 2007.

[9] T. Deselaers and V. Ferrari, "A conditional random field for multiple-instance learning," in *International Conference on Machine Learning (ICML)*, 2010.

[10] A. Joulin and F. Bach, "A convex relaxation for weakly supervised classifiers," in *International Conference on Machine Learning (ICML)*, 2012.

[11] N. Quadrianto, A. Smola, T. Caetano, and Q. Le, "Estimating labels from label proportions," *Journal of Machine Learning Research*, vol. 10, pp. 2349–2374, 2009.

[12] S. Rueping, "Svm classifier estimation from group probabilities." in *International Conference on Machine Learning (ICML)*, 2010.

[13] F. X. Yu, D. Liu, S. Kumar, T. Jebara, and S.-F. Chang, "$\propto$svm for learning with label proportions." in *International Conference on Machine Learning (ICML)*, 2013.

[14] K.-T. Lai, F. X. Yu, M.-S. Chen, and S.-F. Chang, "Video event detection by inferring temporal instance labels," in *IEEE Computer Vision and Pattern Recognition (CVPR)*, Columbus, OH, June 2014.

[15] Z.-H. Zhou, Y.-Y. Sun, and Y.-F. Li, "Multi-instance learning by treating instances as non-i.i.d. samples," in *International Conference on Machine Learning (ICML)*, 2009.

[16] D. Zhang, J. He, L. Si, and R. D. Lawrence, "Mileage: Multiple instance learning with global embedding," in *International Conference on Machine Learning (ICML)*, 2013.

[17] P. F. Felzenszwalb, R. B. Girshick, D. A. McAllester, and D. Ramanan, "Object detection with discriminatively trained part-based models," *IEEE Trans. Pattern Anal. Mach. Intell.*, 2010.

[18] K. Miller, M. P. Kumar, B. Packer, D. Goodman, and D. Koller, "Max-margin min-entropy models," in *Proceedings of the Fifteenth International Conference on Artificial Intelligence and Statistics, AISTATS*, 2012.

[19] C.-N. J. Yu and T. Joachims, "Learning structural svms with latent variables," in *International Conference on Machine Learning (ICML)*, 2009.

[20] S. Sabato and N. Tishby, "Multi-instance learning with any hypothesis class," *Journal of Machine Learning Research*, 2012.

[21] P. L. Bartlett and S. Mendelson, "Rademacher and gaussian complexities: Risk bounds and structural results," *J. Mach. Learn. Res.*, vol. 3, Mar. 2003.

[22] A. L. Yuille and A. Rangarajan, "The concave-convex procedure," *Neural Computation*, vol. 15, no. 4, pp. 915–936, 2003.

[23] B. K. Sriperumbudur and G. R. G. Lanckriet, "On the convergence of the concave-convex procedure," in *NIPS*, 2009.

[24] L. Bottou, "Large-scale machine learning with stochastic gradient descent," in *International Conference on Computational Statistics (COMPSTAT)*, 2010.

[25] N. Le Roux, M. Schmidt, and F. Bach, "A Stochastic Gradient Method with an Exponential Convergence Rate for Strongly-Convex Optimization with Finite Training Sets," in *NIPS*, 2012.

[26] T. Joachims, T. Finley, and C.-N. Yu, "Cutting-plane training of structural svms," *Machine Learning*, vol. 77, no. 1, pp. 27–59, 2009.

[27] G. Heitz, G. Elidan, B. Packer, and D. Koller, "Shape-based object localization for descriptive classification," *Int. J. Comput. Vision*, vol. 84, no. 1, Aug. 2009.

[28] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *CVPR*, 2005.

[29] C. Carson, S. Belongie, H. Greenspan, and J. Malik, "Blobworld: Image segmentation using expectation-maximization and its application to image querying," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, pp. 1026–1038, 1999.

[30] Q. Zhang and S. A. Goldman, "Em-dd: An improved multiple-instance learning technique," in *In Advances in Neural Information Processing Systems*. MIT Press, 2001, pp. 1073–1080.

[31] T. Gärtner, P. A. Flach, A. Kowalczyk, and A. J. Smola, "Multi-instance kernels," in *International Conference on Machine Learning (ICML)*, 2002.

[32] O. Mangasarian and E. Wild, "Multiple instance classification via successive linear programming," *Journal of Optimization Theory and Applications*, vol. 137, no. 3, 2008.

[33] M. Kim and F. De la Torre, "Multiple instance learning via gaussian processes," *Data Mining and Knowledge Discovery (DMKD)*, 2013.

[34] W. Yang, Y. Wang, A. Vahdat, and G. Mori, "Kernel latent svm for visual recognition," in *Advances in Neural Information Processing Systems (NIPS)*, 2012.

[35] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The PASCAL Visual Object Classes Challenge 2007 (VOC2007) Results," 2007.

[36] A. Vedaldi and K. Lenc, "Matconvnet – convolutional neural networks for matlab," *CoRR*, vol. abs/1412.4564, 2014.

[37] M. S. Palmer and Z. Wu, "Verb semantics for english-chinese translation," *Machine Translation*, vol. 10, no. 1-2, pp. 59–92, 1995.

Thibaut Durand is a postdoctoral researcher at Simon Fraser University. He received the Ph.D. in Computer Vision and Machine Learning from the University of Pierre et Marie Curie, France, in 2017. He received an M.Sc. in Electrical Engineering by ENSEA, France, and an M.Sc. degrees in computer science from the University of Cergy-Pontoise, France, in 2013.

Nicolas Thome is a full professor at Conservatoire Nationnal des Arts et Métiers (Cnam Paris). He received the Ph.D. degree in computer science from the University of Lyon, France in 2007, and has been associate professor at UPMC-Paris 6 from 2008 to 2016. His research interests include machine learning for computer vision, including applications for semantic understanding of multimedia data. He is involved in several French (ANR), European and international (Canada, Singapore, Brazil) research projects. He is being coordinator of an ANR project on weakly supervised learning for image retrieval in 2013-2018.

Matthieu Cord is a full professor at Sorbonne University. He received the PhD degree computer science from the UCP, France, before working as postdoc at KU Leuven, Belgium. His research interests include computer vision, deep learning and artificial intelligence. He developed several interactive learning-based approaches for CBIR and many models for pattern recognition using deep architectures. Recently, he focused on multimodal (vison and language) understanding. M. Cord is (co-)author of more than 100 international, peer-reviewed publications among including two edited books. He is involved in several French, European and international research projects. In 2009, he was nominated to the prestigious IUF (French Research Institute).