A Phase Transition-based Perspective on Multiple Instance Kernels

Romaric Gaudel[1,2], Michèle Sebag[1,3], Antoine Cornuéjols[4]

[1] Équipe Inférence et Apprentissage, Laboratoire de Recherche en Informatique,
Bâtiment 490, Université Paris-Sud, 91405 - Orsay Cedex (France)
romaric,sebag@lri.fr

[2] École Normale Supérieure de Cachan

[3] CNRS

[4] UMR AgroParisTech/INRA 518
16, rue Claude Bernard, F-75231 Paris Cedex 05 (France)
antoine.cornuejols@agroparistech.fr
http://www.lri.fr/∼antoine

Abstract : This paper is concerned with relational Support Vector Machines, at the intersection of Support Vector Machines (SVM) and relational learning or Inductive Logic Programming (ILP). The so-called phase transition framework, primarily developed for constraint satisfaction problems (CSP), has been extended to ILP, providing relevant insights into the limitations and difficulties thereof. The goal of this paper is to examine relational SVMs and specifically Multiple Instance-SVMs in the phase transition perspective. Introducing a relaxed CSP formalization of MI-SVMs, we first derive a *lower bound* on the MI-SVM generalization error in terms of the CSP satisfiability probability. Further, ample empirical evidence based on systematic experimentations demonstrates the existence of a unsatisfiability region, entailing the failure of MI-SVM approaches.

Key-words : Phase Transition, Multiple Instance Learning, Relational Kernels, MIP-Support Vector Machine

1 Introduction

This paper is concerned with Relational Support Vector Machines, at the intersection of Support Vector Machines (SVM) (Vapnik, 1998) and Inductive Logic Programming or Relational Learning (Muggleton & De Raedt, 1994). After the so-called kernel trick, the extension of SVMs to relational representations relies on the design of specific kernels (see (Lodhi et al., 2000; Gärtner et al., 2006)).

Relational kernels thus achieve a particular type of propositionalization (Kramer et al., 2001), mapping every relational example in the problem domain onto a propositional

space defined after the training examples. However, relational representations intrinsically embed constrained satisfaction problems; the covering test commonly used in ILP, referred to as Plotkin's θ-subsumption test, is equivalent to a CSP (Botta et al., 2003).

The fact that relational learning involves the resolution of CSPs as a core routine has far-fetched consequences besides exponential (worst-case) complexity, the study of which is at the core of the recent Phase Transition (PT) paradigm in Machine Learning (Cheeseman et al., 1991; Hogg et al., 1996; Giordana & Saitta, 2000) (more on this in section 2).

The question investigated in this paper is whether relational SVMs avoid the limitations of relational learners which has been uncovered in PT studies (Giordana & Saitta, 2000; Botta et al., 2003). Specifically, it was found that a large class of relational learning problems are intrinsically hard to solve. Especially, there are problems for which the learned concepts appearing to perform well are actually very remotely related to the target concepts. This question is examined here w.r.t. a particular relational setting, known as the multiple instance (MI) problem (Dietterich et al., 1997; Mahé et al., 2006).

This paper presents three contributions. Firstly, a relaxed constraint satisfaction problem formalizing the MI-SVM learning search is presented, and a a lower bound on the MI-SVM generalization error is established with respect to the CSP satisfiability probability.

Secondly, a set of order parameters is proposed to describe the critical factors of difficulty for multiple instance learning. Thirdly, extensive and principled experiments show the existence of an unsatisfiability region conditioned by the value of some order parameters, where MI-SVM approaches are doomed to fail.

The paper is organized as follows. For the sake of self-containedness, the phase transition framework is briefly introduced in Section 2 together with MI kernels.

Section 3 rewrites the MI-SVM setting as a constrained satisfaction problem, and relates the satisfiability of this CSP to the generalization error of the MI-SVM problem.

Section 4 reports on the experimental evidence gathered and the paper ends with some perspective for further research.

2 State of the Art

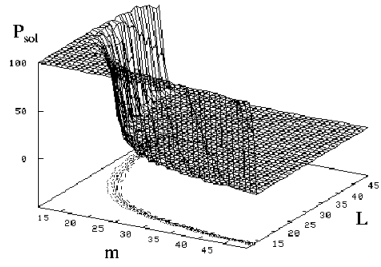
It is widely acknowledged that there is a huge gap between the empirical and the worst case complexity analysis for CSPs (Cheeseman et al., 1991). This remark led to developing the so-called *phase transition framework* (PT) (Hogg et al., 1996), which considers the satisfiability and the resolution complexity of CSP instances as random variables depending on order parameters of the problem instance (e.g. constraint density and tightness).

The phase transition paradigm has been transported to relational machine learning and inductive logic programming (ILP) by (Giordana & Saitta, 2000), based on the fact that the relational covering test, aka θ-subsumption test, is equivalent to a CSP. Fig. 1, left, shows the probability for clause $\mathcal{C}$ to cover example E conditioned by the number m of predicates in $\mathcal{C}$ and the number L of constants in E, for constant values of the number n of variables in $\mathcal{C}$ and the number N of literals per predicate symbols in E

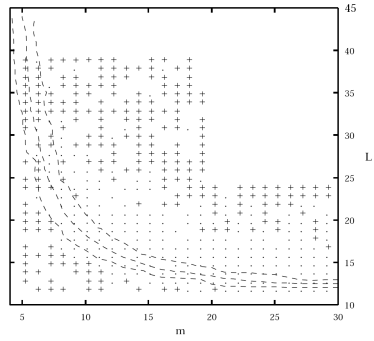
($n = 4,\ N = 100$). Typically, the covering probability is close to 1 when clause $\mathcal{C}$ is general relatively to example E (for small values of m and L), and close to 0 when $\mathcal{C}$ is specific relatively to E. The covering probability drops abruptly in a narrow region, referred to as phase transition.

The phase transition phenomenon has been further investigated in relationship with the success of relational learning, considering the prominent FOIL (relational decision tree) algorithm and other learners (Botta et al., 2003). Artificial learning problems were generated; extensive and principled experimentations show that FOIL and other algorithms fail to learn, i.e. produce hypotheses with test error close to 1/2 when the parameters of the target concept and the training examples are close to the PT region (Fig. 1, right).

Comparable results have been obtained in the field of grammatical inference (Pernot et al., 2005), raising the question of whether the PT-related failure phenomenon can be avoided in relational learning settings.



(a) Probability that a random clause $\mathcal{C}$ covers a random example E, averaged over one thousand pairs $(\mathcal{C}, E)$ for each (m, L) point.



(b) FOIL competence map in plane (m, L): success (legend '+') and failure (legend '.') regions. Dashed curves indicates the phase transition region.

Figure 1: Relational Learning: Phase transition of the covering test, and failure region of the FOIL algorithm in plane (m, L), where m stands for the number of predicates in the clause/target concept, and L for the number of constants in the (training) examples. See text for more details.

This question is investigated in this paper considering the so-called Multiple Instance Learning setting defined by (Dietterich et al., 1997), which is viewed as intermediate between relational and propositional settings. In the MI setting, each example $\mathbf{x}_i$ is a bag of N_i propositional instances $\mathbf{x}_{i,1}, \ldots, \mathbf{x}_{i,N_i}$, where $\mathbf{x}_i$ is positive iff some of its instances satisfy the (propositional) target concept.

Besides early approaches (Dietterich et al., 1997), specific kernels were designed for MI problems (Gärtner et al., 2006; Mahé et al., 2006; Kwok & Cheung, 2007), basically defining the kernel K of two bags of instances as the average of the kernels k

between their instances[1]:

$$K(\mathbf{x}_i, \mathbf{x}_j) = \frac{1}{f_{norm}(\mathbf{x}_i)} \frac{1}{f_{norm}(\mathbf{x}_j)} \sum_{k=1}^{N_i} \sum_{\ell=1}^{N_j} k(\mathbf{x}_{i,k}, \mathbf{x}_{j,\ell}) \quad (1)$$

where f_{norm} denotes some normative function (by example $f_{norm}(\mathbf{x}) = 1$, $f_{norm}(\mathbf{x}) = Card(\mathbf{x})$ or $f_{norm}(\mathbf{x}) = \sqrt{K(\mathbf{x}, \mathbf{x})}$).

After (Gärtner et al., 2006), the approach is efficient under the so-called linearity assumption, that is, the fact that an example is positive iff it contains (at least) one instance pertaining to the target concept.

3 Overview

After the above remarks, MI kernels characterize the similarity of two examples (i.e. two bags of instances) as the average similarity between their instances. The question examined in this paper is to which extent this average similarity is sufficient to reconstruct the existential relational information (do some instances of any example satisfy the target concept) when the linearity assumption does not hold.

Indeed, for quite a few applications tackled as MI problems, such as chemometry for instance (Mahé et al., 2006), it might be doubted whether the linearity assumption holds. As the aim is to discover if a molecule is bio-active or not considering some small sub-parts of this molecule, more than one sub-part of the molecule will most probably be involved to determine its class.

3.1 When MI learning meets CSPs

In order to investigate the above question, one standard procedure is to generate artificial problems, where each problem is made of a training set and a test set, and to compute the test error of the hypothesis learned from the training set. The test error, averaged over a sample of artificial problems generated after a set of parameter values, indeed measures the competence of the algorithm conditionally to these parameter values (Botta et al., 2003).

A different approach is followed in the present paper, for the following reason. Our goal is to examine how kernel tricks can be used to alleviate the specific difficulties of relational learning; in relational terms, the question is about the quality of the propositionalization achieved through relational kernels. In other words, the focus is on the representation (the capacity of the hypothesis search space defined after the MI kernel) instead of a particular algorithm (the quality of the best hypothesis retrieved by this algorithm in this search space).

Accordingly, the methodology we followed is based on the generation of artificial problems composed of a training set $\mathcal{L} = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)\}$ and a test set $\mathcal{T}$ = $\{(\mathbf{x'}_1, y'_1), \ldots, (\mathbf{x'}_{n'}, y'_{n'})\}$. The training set $\mathcal{L}$ induces a propositionalization of the

[1]More sophisticated kernels compare the instance distributions in both bags (Cuturi & Vert, 2004). We shall return to this point in section 5.

domain space, mapping every MI example $\mathbf{x}$ on the n-dimensional real vector $\Phi_{\mathcal{L}}(\mathbf{x}) = (K(\mathbf{x}_1, \mathbf{x}), \ldots, K(\mathbf{x}_n, \mathbf{x}))$. Let $\mathcal{R}_{\mathcal{L}}$ denote this propositional representation based on the training set $\mathcal{L}$.

The novelty of the proposed methodology is to rewrite the MI-SVM learning problem as a constraint satisfaction problem in the $\mathcal{R}_{\mathcal{L}}$ representation.

Specifically, the question examined is: does there exist a separating hyperplane in the propositionalized representation $\mathcal{R}_{\mathcal{L}}$ defined from the training set, which belongs to the search space of MI-SVMs and which correctly classifies the test set (question Q($\mathcal{L}, \mathcal{T}$)), as opposed to, does the separating hyperplane which would have been learned using MI-SVM algorithms from the training set, correctly classify the test set (question Q'($\mathcal{L}, \mathcal{T}$)).

$$\exists\, \vec{\alpha} \in \mathbb{R}^n,\ b \in \mathbb{R} \text{ s.t. } \left\{ \begin{array}{ll} y'_j\ (< \vec{\alpha}, \Phi_{\mathcal{L}}(\mathbf{x}'_j) > +b) \geq 1 & j = 1 \ldots n' \\ \alpha_i \geq 0 & i = 1 \ldots n \end{array} \right. \qquad \text{Q}(\mathcal{L}, \mathcal{T})$$

Clearly, Q($\mathcal{L}, \mathcal{T}$) is much less constrained than Q'($\mathcal{L}, \mathcal{T}$), as Q($\mathcal{L}, \mathcal{T}$) is allowed to use the *test* examples (i.e. cheat...) in order to find the α_i coefficients. The claim is that Q($\mathcal{L}, \mathcal{T}$) gives deep insights into the quality of propositionalization $\mathcal{R}_{\mathcal{L}}$, while Q'($\mathcal{L}, \mathcal{T}$) additionally depends on the quality of a particular algorithm operating on $\mathcal{R}_{\mathcal{L}}$. Formally, with inspiration from (Kearns & Li, 1993), we show that the percentage of times Q($\mathcal{L}, \mathcal{T}$) succeeds induces a lower bound on the generalization error reachable in representation $\mathcal{R}_{\mathcal{L}}$.

Proposition

Within a MI-SVM setting, let $\mathcal{L}$ be a training set of size n, $\mathcal{R}_{\mathcal{L}}$ the associate propositionalization and $p_{\mathcal{L}}$ the generalization error of the optimal linear classifier $h^*_{\mathcal{L}}$ defined on $\mathcal{R}_{\mathcal{L}}$.

Let $\mathbb{E}_n[p_{\mathcal{L}}]$ denote the expectation of $p_{\mathcal{L}}$ conditionally to $|\mathcal{L}| = n$.

Let MI-SVM problems $(\mathcal{L}_i, \mathcal{T}_i)$, $i = 1 \ldots N$ be drawn independently, where the size of $\mathcal{L}_i$ and $\mathcal{T}_i$ respectively is n and n'. Let $\hat{\tau}_{n,n'}$ denote the fraction of CSPs Q($\mathcal{L}_i, \mathcal{T}_i$) that are satisfiable.

Then for any $\eta > 0$, with probability at least $1 - exp(-2\eta^2\, N)$,

$$\mathbb{E}_n[p_{\mathcal{L}}] \geq 1 - (\hat{\tau}_{n,n'} + \eta)^{\frac{1}{n'}}.$$

Proof

Let the MI-SVM problem and $\mathcal{L}$ be fixed; by construction, the probability for a test dataset $\mathcal{T}$ of size n' to include no example misclassified by $h^*_{\mathcal{L}}$ is $(1 - p_{\mathcal{L}})^{n'}$.

It is straightforward to see that if $\mathcal{T}$ does not contain examples that are misclassified by $h^*_{\mathcal{L}}$, Q($\mathcal{L}, \mathcal{T}$) is satisfiable. Therefore the probability for Q($\mathcal{L}, \mathcal{T}$) to be satisfiable conditionally to $\mathcal{L}$ is greater than $(1 - p_{\mathcal{L}})^{n'}$:

$$\mathbb{E}_{|\mathcal{T}|=n'}[\ \text{Q}(\mathcal{L}, \mathcal{T}) \text{ satisfiable}] \geq (1 - p_{\mathcal{L}})^{n'}$$

Taking the expectation of the above w.r.t. $|\mathcal{L}| = n$, it comes:

$$\mathbb{E}_{|\mathcal{T}|=n',\ |\mathcal{L}|=n}[\ \text{Q}(\mathcal{L}, \mathcal{T}) \text{ satisfiable}] \geq \mathbb{E}_{|\mathcal{L}|=n}[(1 - p_{\mathcal{L}})^{n'}] \geq (1 - \mathbb{E}_n[p_{\mathcal{L}}])^{n'} \quad (2)$$

where the right inequality follows from Jensen's inequality as function $x \mapsto (1-x)^{n'}$ is convex on $[0,1]$. Next step is to bound the left term from its empirical estimate $\hat{\tau}_{n,n'}$, using Hoeffding's bound. With probability at least $1 - exp(-2\eta^2 N)$,

$$\mathbb{E}_{|\mathcal{T}|=n',\ |\mathcal{L}|=n}[\ \mathrm{Q}(\mathcal{L},\mathcal{T}) \text{ satisfiable}] < \hat{\tau}_{n,n'} + \eta \tag{3}$$

From (2) and (3) it comes that with probability at least $1 - exp(-2\eta^2 N)$

$$(1 - \mathbb{E}_n[p_{\mathcal{L}}])^{n'} \leq \hat{\tau}_{n,n'} + \eta$$

which concludes the proof. □

3.2 The Order Parameters

Following the standard PT methodology, problems are uniformly generated after order parameters conditioning the description of instances, examples and target concept.

At the *instance* level, each instance $I = (a, \vec{v})$ is formed of a symbol[2] a drawn in an alphabet Σ, and a d-dimensional vector $\vec{v}$, in $[0,1]^d$. By definition, the ε ball of an instance I denoted $\mathcal{B}_\varepsilon(I)$ includes all instances $I' = (a', \vec{v}')$ such that I and I' bear the same symbol $a = a'$ and for each k coordinate, $k = 1 \ldots d$, the absolute difference $|\vec{v}_k - \vec{v}'_k|$ is less than ε.

At the *concept level*, the target concept is characterized as the conjunction of P elementary concepts C_i, where C_i is the ε ball centered on some target instance I_i uniformly drawn in $[0,1]^d$.

At the *example level*, a positive (respectively negative) example $\mathbf{x}_i$ is characterized as a set of N^+ (resp. N^-) instances $\mathbf{x}_{i,l}$; example $\mathbf{x}_i$ is positive iff each C_j in the target concept contains at least one instance of $\mathbf{x}_i$. The N^+ instances of *positive examples* are drawn as follows (Fig. 2): P_{ic} instances are drawn in the elementary concepts C_i, ensuring that at least one instance is drawn in every C_i ($P_{ic} \geq P$). Likewise, the N^- instances of *negative examples* involve N_{ic} instances drawn in the elementary concepts C_i, ensuring that nm (near-miss) C_i are not visited ($nm \geq 1$).

Instances which do not belong to the target concept balls are drawn either (i) uniformly in $[0,1]^d$ (uniform default instances); or (ii) among P_U balls forming the *Universe concept*, introduced to model the fact that example instances are not uniform in real-world problems (universe default instances). In the latter setting, the Universe concept is made of P_U balls with radius ε, and it is similarly required that not all balls of the Universe be visited by an example; the number of Universe balls not visited by positive examples is set to nm_U.

4 Experiments

After describing the experimental setting, this section reports on the results. All first experiments use uniform default instances; the case of universe default instances is discussed in section 4.6.

[2]The features of instances could be in a countable set or in a continuous space. As the finite product of countable sets is countable, adding a feature in an alphabet can simulate the effect of all countable features.

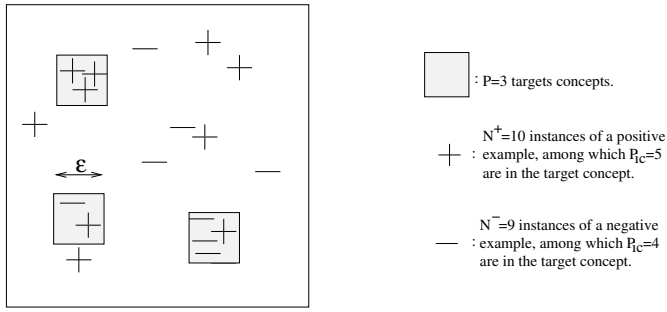


Figure 2: Values of instances of 2 examples in a space of dimension $d = 2$, with an alphabet Σ of size $|\Sigma| = 1$ and $nm = 1$.

4.1 Experimental setting

Unless otherwise specified, the order parameter values are fixed or vary in the intervals as described in Table 1. These values were chosen such that the presented effects could be easier to see.

$\|\Sigma\|$	Size of the alphabet Σ	15
d	Dimension of the instances : $\mathbf{x}_i \in [0,1]^d$	30
P	Number of balls in the target concept	30
ε	Radius of a ball (elementary concept)	.15
n	Number of training examples	60 (30 +, 30 −)
n'	Number of test examples	200 (100 +, 100 −)
N^+, N^-	Number of instances in pos./neg. example	100
P_{ic}	Number of instances in tc for a positive ex.	[30,100]
N_{ic}	Number of instances in tc for a negative ex.	[0, 100]
nm	Number of target balls not visited by neg. ex.	20
P_U	Number of balls of the universe concept	30
nm_U	Number of universe balls not visited by pos. ex.	15

Table 1: Order parameters for CSP Q($\mathcal{L}, \mathcal{T}$) and range of variations

For each set of order parameter values, 40 MI-SVM problems are constructed by independently drawing the target concept, the training set $\mathcal{L}$ and the test set $\mathcal{T}$. The bag kernel is defined as in eq. (1), where the instance kernel is a Gaussian kernel and the normalization factor is set to $f_{norm}(\mathbf{x}) = Card(\mathbf{x})$. Similar results, omitted due to lack of space, are obtained using polynomial kernels (linear, quadratic and of degree 4).

Based on $\mathcal{L}$ and $\mathcal{T}$, the constraint satisfaction problem Q($\mathcal{L}, \mathcal{T}$) is defined (section 3.1), involving $n' = 200$ constraints and $n + 1 = 61$ variables, and solved using the GLPK package. The average satisfiability of Q($\mathcal{L}, \mathcal{T}$) for a set of parameter values is monitored, and displayed in the 2-dimensional plane P_{ic}, N_{ic}; the color code is black (resp. white) if the fraction of satisfiable CSPs is 0 (resp. 100%). It is expected that for

$P_{ic} = N_{ic}$, Q($\mathcal{L}, \mathcal{T}$) might be unsatisfiable; as the MI kernel only describes the average instance similarity, positive and negative examples should have similar distributions in representation $\mathcal{R}_{\mathcal{L}}$.

4.2 Sensitivity analysis w.r.t. Near-miss

Let us first examine the influence of the near-miss parameter nm, ruling the number of elementary concepts which are not visited by instances of negative examples. As expected, a failure region centered on the diagonal $P_{ic} = N_{ic}$ can be observed; furthermore the failure region increases as the near-miss parameter increases (Fig. 3).

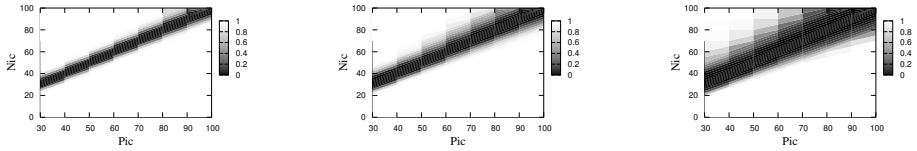


Figure 3: Fraction of satisfiable Q($\mathcal{L}, \mathcal{T}$) in plane P_{ic}, N_{ic} out of 40 runs. Influence of the near-miss parameter: **Left:** nm = 10. **Center:** nm = 20. **Right:** nm = 25.

These results are explained as follows. The MI-SVM propositionalization maps every example $\mathbf{x}$ onto the n-dimensional vector $\Phi_{\mathcal{L}}(\mathbf{x}) = (K(\mathbf{x}_1, \mathbf{x}), \cdots, K(\mathbf{x}_n, \mathbf{x}))$. The distribution of propositionalized examples, in the 2D plan defined from a positive and a negative training example, is displayed on Fig. 4.

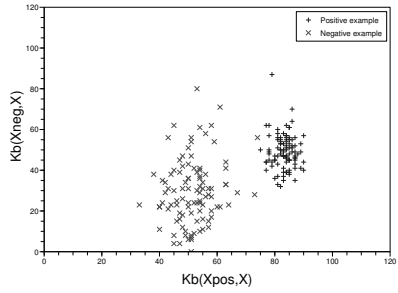


Figure 4: Distribution of $\Phi_{\mathcal{L}}(\mathbf{x})$ for $\mathbf{x}$ positive (legend $+$) and $\mathbf{x}$ negative (legend $\times$), where $P = 30,\ nm = 20,\ P_{ic} = 50,\ N_{ic} = 30$. The first (resp. second) axis is derived from a positive (resp. negative) training example.

Let C (resp. c) denote the mean value of $k(I, I')$ for two instances I and I' belonging to the same elementary concept (resp. drawn uniformly in the instance space). These values depend on both the instance kernel and the instance order parameters d and $|\Sigma|$, set to constant values in the experiments.

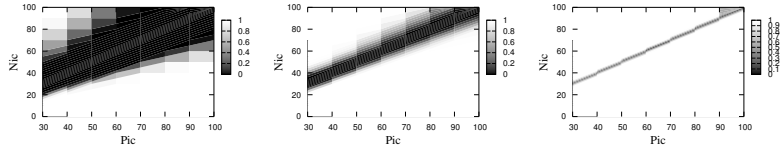
With no difficulty, it is shown that when $\mathbf{x}_i$ and $\mathbf{x}$ are positive, the expectation of $K(\mathbf{x}_i, \mathbf{x})$ is $\frac{1}{P}(\frac{P_{ic}}{N^+})^2(C - c) + c$. Likewise, if both examples are negative, the expec-

tation of $K(\mathbf{x}_i, \mathbf{x})$ is $\frac{1}{P}(\frac{N_{ic}}{N^-})^2(C-c)+c$. Last, if both examples belong to different classes, the expectation of $K(\mathbf{x}_i, \mathbf{x})$ is $\frac{1}{P}\frac{P_{ic}}{N^+}\frac{N_{ic}}{N^-}(C-c)+c$.

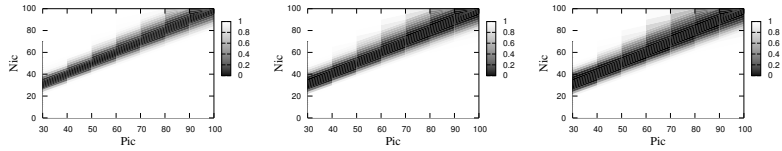
Therefore, when $P_{ic} = N_{ic}$[3], the distribution of $K(\mathbf{x}_i, \mathbf{x})$ does not depend on the class of $\mathbf{x}$, which clearly hinders the linear discrimination task.

In the general case (when $P_{ic} \neq N_{ic}$), both distributions differ by their average value and by their variance. Still, as the distributions of positive and negative test examples in the propositionalized representation $\mathcal{R}_\mathcal{L}$ overlap, their linear separation is only made possible as the number of training examples increases.

Note that although the near-miss parameter nm has no effect on the center of both distributions, the variance of the propositionalization increases with nm. The larger dispersion of the propositional examples thus adversely affects the satisfiability of Q($\mathcal{L}, \mathcal{T}$), as shown on Fig. 3.



(a) Influence of the size of the training set. **Left:** n = 20. **Center:** n = 60. **Right:** $n = 180$.



(b) Influence of the size of the test set. **Left:** n' = 100.**Center:** n' = 200. **Right:** $n' = 400$.

Figure 5: Fraction of satisfiable Q($\mathcal{L}, \mathcal{T}$) in plane P_{ic}, N_{ic} out of 40 runs.

4.3 Size of the training and test sets

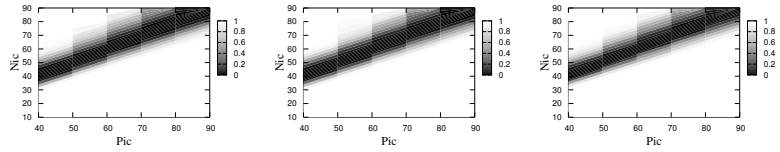
As could have been expected, increasing the number of training examples n makes the failure region to decrease (Fig. 5.a); the learning task is easier as more training examples are available. On one hand − provided that $N_{ic} \neq P_{ic}$ −, the distance between the centers of the propositionalized positive and negative example distributions increases proportionally to $\sqrt{n}$, where n is the number of training examples. On the other hand, the more training examples, the more likely one of them will derive a propositional attribute with good discrimination power.

In contrast, the size of the failure region increases with the size of the test set (Fig. 5.b); clearly, the more constraints in Q($\mathcal{L}, \mathcal{T}$), the lower its probability of satisfiability is.

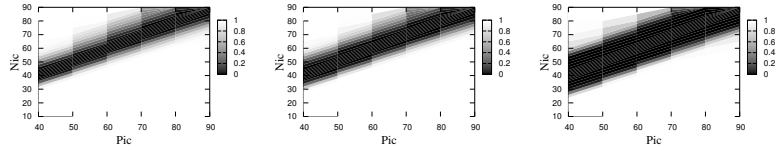
[3]Actually, the failure region corresponds to $\frac{P_{ic}}{N^+} = \frac{N_{ic}}{N^-}$. The distinction is not made as for experiments $N^+ = N^-$.

4.4 Sensitivity analysis w.r.t. P_{ic} and N_{ic}

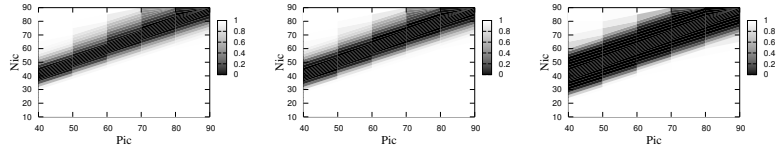
The influence of the dispersion of P_{ic} and N_{ic} is examined as follows. Firstly, the number of instances in positive (respectively, negative) training examples is uniformly drawn in $[P_{ic} - \Delta, P_{ic} + \Delta]$ (resp. $[N_{ic} - \Delta, N_{ic} + \Delta]$), with Δ varying in [0,10] while the number of instances in test examples is kept fixed.



(a) Variation only for training examples.



(b) Variation only for test examples.



(c) Variation for both training and test examples.

Figure 6: Fraction of satisfiable Q($\mathcal{L}, \mathcal{T}$) in plane P_{ic}, N_{ic} out of 40 runs. Influence of the variability Δ on P_{ic} and N_{ic}. **Left:** Δ = 0. **Center:** Δ = 5. **Right:** Δ = 10.

When Δ increases, the size of the failure region decreases (Fig. 6.a); indeed, the higher variance among the training examples makes it more likely that one of them will derive a propositional attribute with good discrimination power.

Secondly, the number of instances for training examples is fixed while the number of instances in positive (respectively, negative) test examples is uniformly drawn in $[P_{ic} - \Delta, P_{ic} + \Delta]$ (resp. $[N_{ic} - \Delta, N_{ic} + \Delta]$), with Δ varying in [0,10]. Here, the failure region increases with Δ (Fig. 6.b); the higher variance of the test examples makes it more likely to generate inconsistent constraints.

Finally, if the number of instances varies for both training and test examples, the overall effect is to increase the failure region: even though there are propositional attributes with better discriminant power, there are more inconsistent constraints too, and the percentage of satisfiable problems decreases.

4.5 Sensitivity Analysis w.r.t. Example size

The impact of default instances (not belonging to any elementary target concept) is studied through increasing the example size N^+ and N^-. Experimentally, the failure region increases with N^+ and N^- (Fig. 7). The interpretation proposed for this finding goes as follows.

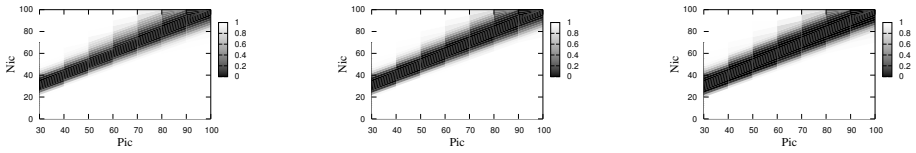


Figure 7: Fraction of satisfiable Q($\mathcal{L}, \mathcal{T}$) in plane P_{ic}, N_{ic} out of 40 runs. Influence of the size of the examples. **Left:** $N^+ = N^-$ = 100.**Center:** $N^+ = N^-$ = 200. **Right:** $N^+ = N^- = 400$.

On one hand, the distance between positive and negative example distributions is increasingly due to the influence of default instances as N^+ and N^- increase. On the other hand, the instances in positive and negative examples are in majority default ones when N^+ and N^- increase; therefore the ratio signal to noise in the propositional representation decreases and the failure region increases.

On the other hand, the effect of default instances is limited as they are far away from each other (in the uniform default instance setting), comparatively to instances belonging to concept balls. Therefore increasing the number of default instances does not much modify $K(\mathbf{x}, \mathbf{x}')$ on average, which explains why the effect of N^+ and N^- appears to be moderate.

4.6 Sensitivity Analysis w.r.t. the Universe Concept

This section examines the sensitivity of the results when default instances are drawn in the Universe concept (section 3.2).

4.6.1 Effect of the size of the Universe (P_U balls).

The impact of the Universe Concept can be expressed analytically, examining the distributions of positive and negative examples in the propositionalized representation. The largest failure region is observed for $P_{ic} = N_{ic} \approx N \frac{P}{P_U + P}$.

Accordingly, the failure region is very thin for small values of P_U (Fig. 8); for large values of P_U, the failure region is similar to the non-Universe case. For intermediate values of P_U, the failure region is larger than for the non-Universe setting.

4.6.2 Effect of the near miss factor of the Universe.

The number of near-miss nm (number of concept balls not visited by the negative instances) and the number nm_U (number of Universe balls not visited by positive exam-

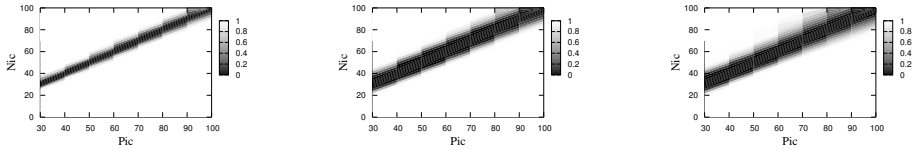


Figure 8: Fraction of satisfiable Q($\mathcal{L}, \mathcal{T}$) in plane P_{ic}, N_{ic} out of 40 runs. Influence of the size P_U of the Universe when nm_U = 0. **Left:** P_U = 5. **Center:** P_U = 30. **Right:** P_U = 1000.

ples) have similar effects : the variance of $\Phi_{\mathcal{L}}(\mathbf{x})$ increases with nm and nm_U, and the satisfiability probability of Q($\mathcal{L}, \mathcal{T}$) decreases accordingly.

Note however that the impact of nm is maximal for large value of P_{ic} and N_{ic} (Fig. 3), while the opposite holds for nm_U (Fig. 9). This is explained as nm influences the distribution of the P_{ic} (resp. N_{ic}) instances in the target concept while nm_U influences the distribution of the $N^+ - P_{ic}$ (resp. $N^- - N_{ic}$) instances drawn in the universe.

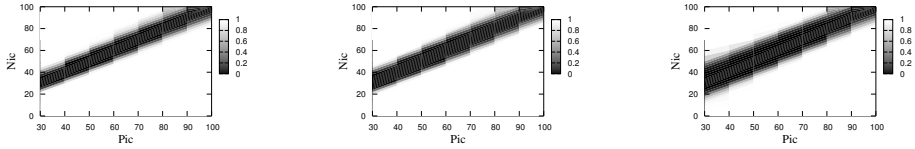


Figure 9: Fraction of satisfiable Q($\mathcal{L}, \mathcal{T}$) in plane P_{ic}, N_{ic} out of 40 runs. Influence of the size of the near-miss factor of the Universe. **Left:** nm_U = 0. **Center:** nm_U = 15. **Right:** nm_U = 25.

Overall, the Universe is shown to amplify the variations due to the example size, as the default instances (not related to the target concept) now influence the variance of the propositionalized distribution (Fig. 10).

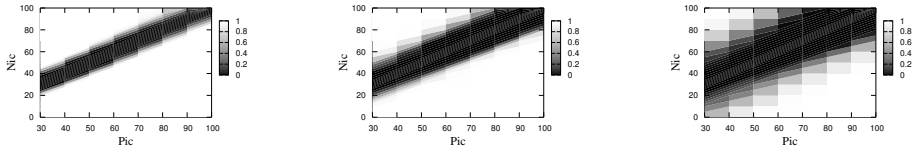


Figure 10: Fraction of satisfiable Q($\mathcal{L}, \mathcal{T}$) in plane P_{ic}, N_{ic} out of 40 runs. Influence of the size of the example using a Universe. **Left:** $N^+ = N^-$ = 100. **Center:** $N^+ = N^-$ = 200. **Right:** $N^+ = N^-$ = 400.

5 Discussion and Perspectives

The main contribution of this paper is to evidence some Phase Transition-related limitations of MI kernels. The presented approach is based on a lower bound of the generalization error, expressed in terms of the satisfaction probability of a CSP on the propositionalized representation induced by a MI kernel.

Clearly, some care must be exercised to interpret the limitations of the well-founded MI-SVM algorithms suggested by our experiments on artificial problems. In particular, more sophisticated kernels proceed by comparing the instance distributions in the examples at hand (Cuturi & Vert, 2004) and they remain to be examined in the proposed CSP framework.

Still, the question of whether Multiple Instance Kernels enable to characterize *existential* properties as opposed to *average* properties makes sense in a relational perspective. Actually, in some domains where the number and/or the diversity of the available examples are limited, as in the domain of chemometry (Mahé et al., 2006), one might learn average properties, these might do well on the test set, and still be poorly related to the target concept; some evidence for the possibility of such a phenomenon was presented in (Botta et al., 2003), where the test error could be 2% or lower although the concept learned was a gross overgeneralization of the true target concept.

A further research perspective opened by this work is based on a tighter coupling between the CSP resolution and the Multiple Instance Kernel-based propositionalisation, in the line of dynamic propositionalization (Blockeel et al., 2005).

References

Blockeel, H., Page, D., Srinivasan, A. (2005). Multiple Instance Decision Tree Learning *Proc. ICML05* (pp. 57–64).

Botta, M., Giordana, A., Saitta, L., & Sebag, M. (2003). Relational learning as search in a critical region. *Journal of Machine Learning Research*, *4*, 431–463.

Cheeseman, P., Kanefsky, B., & Taylor, W. (1991). Where the really hard problems are. *Proc. of Int. Joint Conf. on Artificial Intelligence* (pp. 331–337)

Cuturi, M., & Vert, J.-P. (2004). Semigroup kernels on finite sets. *NIPS04* (pp. 329–336).

Dietterich, T., Lathrop, R., & Lozano-Perez, T. (1997). Solving the multiple-instance problem with axis-parallel rectangles. *Artificial Intelligence*, *89 (1-2)*, 31–71.

Giordana, A., & Saitta, L. (2000). Phase transitions in relational learning. *Machine Learning*, *41*, 217–251.

Gärtner, T., Flach, P. A., Kowalczyk, A., & Smola, A. J. (2006). Multi-instance kernels. *Proc. ICML02* (pp. 179–186).

Hogg, T., Huberman, B., & (Eds), C. W. (1996). *Artificial intelligence: Special issue on frontiers in problem solving: Phase transitions and complexity*, vol. 81(1-2). Elsevier.

Kearns, M., & Li, M. (1993). Learning in the presence of malicious errors. *SIAM J. Comput.*, *22*, 807–837.

Kersting, K., & Raedt, L. D. (2001). Bayesian logic programs. *Proc. of the 11th Int. Conf. on Inductive Logic Programming*.

Kramer, S., Lavrac, N., & Flach, P. (2001). Propositionalization approaches to relational data mining. In S. Dzeroski and N. Lavrac (Eds.), *Relational data mining*, 262–291. Springer Verlag.

Kwok, J., Cheung, P.-M. (2007). Marginalized Multi-Instance Kernels. *Proc. of the 20th Int. Joint Conf. on Aritificial Intelligence* , 2007, 901–906.

Lodhi, H., Shawe-Taylor, J., Cristianini, N., & Watkins, C., J., C., H. (2000). Text Classification using String Kernels. *NIPS00* (pp. 563–569).

Mahé, P., Ralaivola, L., Stoven, V., & Vert, J.-P. (2006). The pharmacophore kernel for virtual screening with support vector machines. *Journal of Chemical Information and Modeling*, *46*, 2003–2014.

Muggleton, S., & De Raedt, L. (1994). Inductive logic programming: Theory and methods. *Journal of Logic Programming*, *19*, 629–679.

Pernot, N., Cornuéjols, A., & Sebag, M. (2005). Phase transitions within grammatical inference. *Proc. Int. Conf. on Artificial Intelligence* (pp. 811–816). IOS Press.

Quinlan, J., R. (1990). Learning logical definitions from relations. *Machine Learning 5* (pp. 239–266).

Rückert, U., Kramer, S., & De Raedt, L. (2003). Stochastic local search in k-term dnf learning. *Proc. of the Int. Conf. on Machine Learning* (pp. 648–655). AAAI Press.

Vapnik, V. N. (1998). *Statistical learning theory*. Wiley.