A Phase Transition-Based Perspective on Multiple Instance Kernels

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

[1] LRI; Univ. Paris-Sud, CNRS; F-91405 Orsay, France,
{romaric.gaudel,sebag}@lri.fr
[2] École Normale Supérieure de Cachan
[3] AgroParisTech / INRA; UMR 518 Mathématiques et Informatique Appliqués;
F-75005 Paris, France,
antoine.cornuejols@agroparistech.fr

Abstract. This paper is concerned with Relational Support Vector Machines, at the intersection of Support Vector Machines (SVM) and Inductive Logic Programming or Relational Learning. The so-called phase transition framework, primarily developed for constraint satisfaction problems (CSP), has been extended to relational learning, providing relevant insights into the limitations and difficulties thereof. The goal of this paper is to examine relational SVMs and specifically Multiple Instance Kernels along the phase transition framework; a specific CSP formalization for multiple instance problems, inspired by chemometry applications, is proposed. Ample empirical evidence based on a set of order parameters shows the existence of an unsatisfiability region for standard MIP-SVM approaches. A statistical analysis for these findings is proposed, establishing a lower bound of the generalization error depending on the satisfiability probability.

Key words: Phase Transition, Multiple Instance Problems, Relational Learning, Relational Kernels, MIP-Support Vector Machines.

1 Introduction

This paper is concerned with Relational Support Vector Machines, at the intersection of Support Vector Machines (SVM) [16] and Inductive Logic Programming or Relational Learning [13]. After the so-called kernel trick, the extension of SVMs to relational representations relies on the design of specific kernels (see [3, 5]). Relational kernels thus achieve a particular type of propositionalization [10], mapping every relational example in the problem domain onto a propositional space defined after the training examples. However, relational representations intrinsically embed combinatorial issues, such as the Plotkin's θ-subsumption test [1]. The fact that relational learning involves the resolution of CSPs as a core routine has far-fetched consequences besides exponential (worst-case) complexity, referred to as the Phase Transition (PT) paradigm [2, 7].

The question investigated in this paper is whether relational SVMs avoid the limitations of relational learners related to the PT region [6, 1]. This question is

examined w.r.t. a particular relational setting, known as the multiple instance problem (MIP) [4, 12]. This paper presents three contributions. Firstly, the MIP-SVM search is rewritten in terms of CSP, and a lower bound on the generalization error of the MIP-SVM is established in terms of the probability of satisfiability of the CSP. 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 a failure region for MIP-SVMs, conditionned by the value of some order parameters.

The paper is organized as follows. For the sake of self-containedness, the phase transition framework is briefly introduced in Section 2 together with MIP kernels. Section 3 rewrites the MIP-SVM setting as a constrained satisfaction problem, and relates the satisfiability of this CSP with the generalization error of the MIP-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 [2]. This remark led to developing the so-called *phase transition framework* (PT) [7], 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 [6], and was shown to be instrumental in identifying and analyzing some limitations of relational learning [1] or grammatical inference [14] algorithms.

This paper will investigate the PT approach in a specific setting known as Multiple Instance Learning [4], which is viewed as intermediate between relational and propositional settings. In the MIP setting, each example is a bag of instances; the example is positive iff some of its instances satisfy the target concept. Under the so-called *linearity bias* assumption a positive example only needs one of its instances to belong to the target concept.

Actually, an example or bag of instances can be viewed as a set of literals built on a single predicate symbol; equivalently, an example is a set of rows in a matrix where the columns are the arguments of the predicate. Formally, we shall restrict ourselves to the following MIP representation [4]: each example $\mathbf{x}_i$ is a set of N_i instances noted $\mathbf{x}_{i,1}, \ldots, \mathbf{x}_{i,N_i}$; we further assume in the rest of this section that the instance space is $\mathbb{R}^d$ ($\mathbf{x}_{i,j} \in \mathbb{R}^d$).

Besides early approaches [4], specific kernels were designed for MIP problems [5, 3, 12, 11]. The basic idea is to define the kernel K of two bags of instances as the average of the kernels k between their instances:

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

Note that such kernels do not rely on the linearity assumption in any way; $K(\mathbf{x}_i, \mathbf{x}_j)$ only reflects the average similarity between the instances of both examples.

3 Overview

After the above remarks, MIP kernels characterize the similarity of two examples, bags of instances, as the average similarity between their instances. The question examined in this paper is to which extent this average information is sufficient to reconstruct the existential relational information (do the instances of any example satisfy the target concept).

3.1 When MIP 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 [1].

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 MIP 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 domain space, mapping every MIP 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 MIP-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 MIP-SVMs and which correctly classifies the test set (question $\mathrm{Q}(\mathcal{L}, \mathcal{T})$), as opposed to, does the separating hyperplane which would have been learned using MIP-SVM algorithms from the training set, correctly classify the test set (question $\mathrm{Q'}(\mathcal{L}, \mathcal{T})$).

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

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 much deeper insights into the quality of the propositionalization based on the kernel trick. Formally, with inspiration from [8], 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 MIP-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 MIP-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 MIP-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 conditionnally 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 $1 - exp(-2\eta^2 N)$,

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

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

After the standard PT setting, the distribution of the problems is parametrized based on order parameters respectively devoted to the characterization of instances, target concept and examples.

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

$\lvert\Sigma\rvert$	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 example	[30,100]
N_{ic}	Number of instances in tc for a negative example	[0, 100]
nm	Number of target balls not visited by neg. examples	20
P_U	Number of balls of the universe concept	30
nm_U	Number of universe balls not visited by pos. examples	15

Table 1. Order parameters for the MIP constraint satisfaction problem and their range of variations

4.1 Experimental setting

Unless otherwise specified, the order parameter values are fixed or vary in the intervals as described in Table 1.

For each set of order parameter values, 40 MIP-SVM problems are constructed by independently drawing the target concept, the training set $\mathcal{L}$ and the test set $\mathcal{L}$, using a Gaussian kernel as instance kernel. Correspondingly, CSP $\mathrm{Q}(\mathcal{L}, \mathcal{T})$ is constructed (section 3.1), involving $n' = 200$ constraints and $n + 1 = 61$ variables; it is solved using the GLPK package. The average satisfiability of $\mathrm{Q}(\mathcal{L}, \mathcal{T})$ 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 $P_{ic} = N_{ic}$, $\mathrm{Q}(\mathcal{L}, \mathcal{T})$ might be unsatisfiable as the MIP kernel only describes the averaged instance similarity.

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. 1).

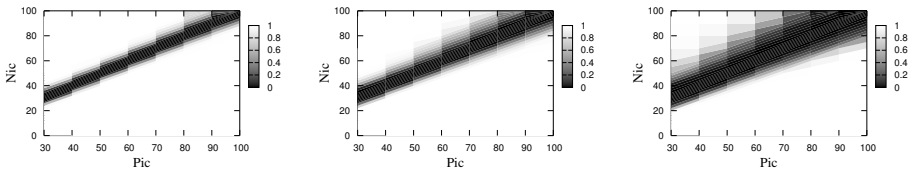


Fig. 1. Fraction of satisfiable $\mathrm{Q}(\mathcal{L}, \mathcal{T})$ in plane P_{ic}, N_{ic} out of 40 runs. Influence of the near-miss parameter: **Left:** $nm = 10$. **Center:** $nm = 20$. **Right:** $nm = 25$.

These results are explained as follows. The MIP 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. 2.

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 expectation 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 $\frac{1}{P}\frac{P_{ic}}{N^+}\frac{N_{ic}}{N^-}(C - c) + c$.

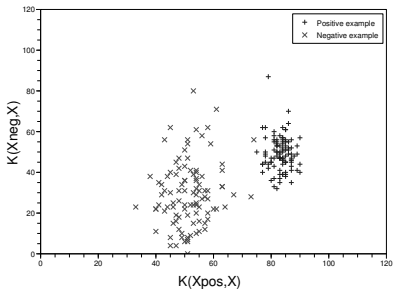


Fig. 2. Distribution of $\Phi_{\mathcal{L}}(\mathbf{x})$ for $\mathbf{x}$ positive (legend +) and $\mathbf{x}$ negative (legend $\times$), where $P = 30,\ nm = 20,\ P_{ic} = 50,\ N_{ic} = 30$. The first (resp. second) axis is derived from a positive (resp. negative) training example.

Therefore when $P_{ic} = N_{ic}$[4], 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.

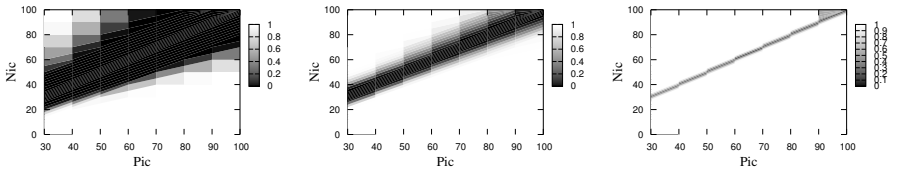
Note that although the near-miss parameter nm has no effect on the center of both distributions, the variance of the propositionalization increases with nm. The larger dispersion of the propositional examples thus adversely affects the satisfiability of the (Q) CSP, as shown on Fig. 1.

4.3 Size of the training and test sets

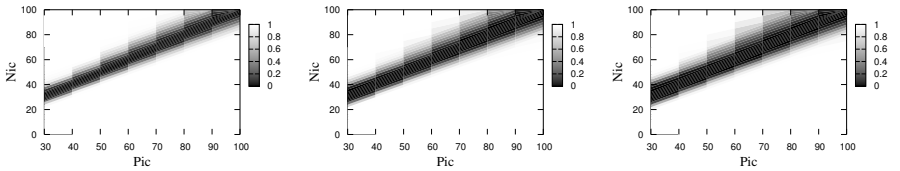
As could have been expected, increasing the number of training examples n makes the failure region to decrease (Fig. 3 (a)); indeed the learning task is made easier as more training examples are available. On one hand − provided that $N_{ic} \neq P_{ic}$ −, the distance between the centers of the propositionalized positive and negative examples 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. 3 (b)); clearly, the more constraints in $\mathrm{Q}(\mathcal{L}, \mathcal{T})$, the lower its probability of satisfiability is.

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



(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:** $n' = 400$.

Fig. 3. Fraction of satisfiable $\mathrm{Q}(\mathcal{L}, \mathcal{T})$ in plane P_{ic}, N_{ic} out of 40 runs.
the

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.

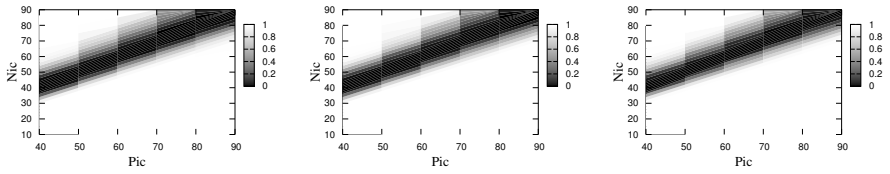
When Δ increases, the size of the failure region decreases (Fig. 4 (a)); indeed, the higher variance among the training examples makes it more likely that one of them will derive a propositional attribute with good discrimination power.

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. 4 (b)); as the higher variance among the test examples makes it more likely to generate inconsistent constraints.

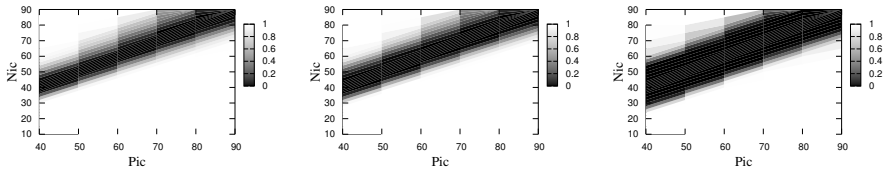
Finally, if the number of instances in all training and test examples varies, the overall effect is to increase the failure region: even though there are propositional attributes with better discriminant power, there are more inconsistent constraints too, and the percentage of satisfiable problems decreases.

4.5 Sensitivity Analysis w.r.t. Example size

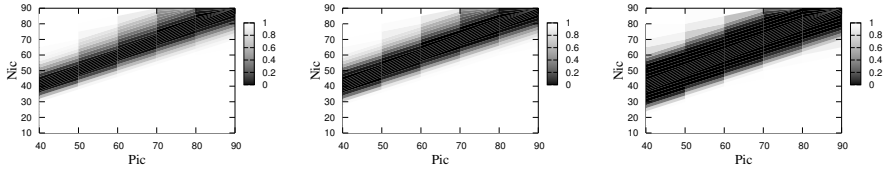
The impact of default instances (not belonging to any elementary target concept) is studied through increasing the example size N^+ and N^-. Experimentally, the



(a) Variation only for training examples.



(b) Variation only for test examples.



(c) Variation for both training and test examples.

Fig. 4. Fraction of satisfiable CSP (Q) in plane P_{ic}, N_{ic} out of 40 runs. Influence of the variability Δ on P_{ic} and N_{ic}. **Left:** $\Delta = 0$. **Center:** $\Delta = 5$. **Right:** $\Delta = 10$.

failure region increases with N^+ and N^- (Fig. 5). The interpretation proposed for this finding goes as follows.

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.

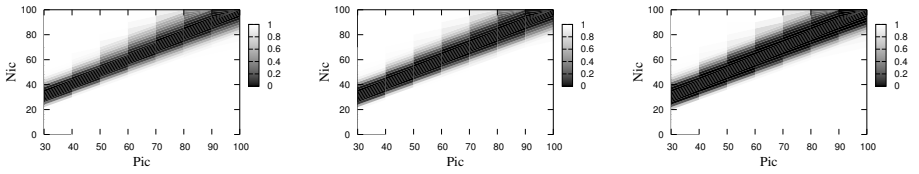


Fig. 5. Fraction of satisfiable CSP (Q) in plane P_{ic}, N_{ic} out of 40 runs. Influence of the size of the examples. **Left:** $N^+ = N^- = 100$.**Center:** $N^+ = N^- = 200$. **Right:** $N^+ = N^- = 400$.

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).

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 (calculations omitted for space limitations). 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. 6); for large values of P_U, the failure region is similar to the non-Universe case. For intermediate values, a larger failure region is observed, compared to the non-Universe case.

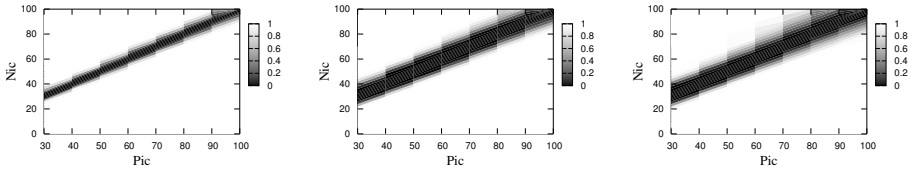


Fig. 6. Fraction of satisfiable CSP (Q) in plane P_{ic}, N_{ic} out of 40 runs. Influence of the size P_U of the Universe when $nm_U = 0$. **Left:** $P_U = 5$. **Center:** $P_U = 30$. **Right:** $P_U = 1000$.

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 examples) have similar effects : the variance of $\Phi_{\mathcal{L}}(\mathbf{x})$ increases with nm and nm_U, and the probability for the CSP (Q) to be satisfied decreases accordingly.

Note however that the impact of nm is maximal for large value of P_{ic} and N_{ic} (Fig. 1), while the opposite holds for nm_U (Fig. 7). 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.

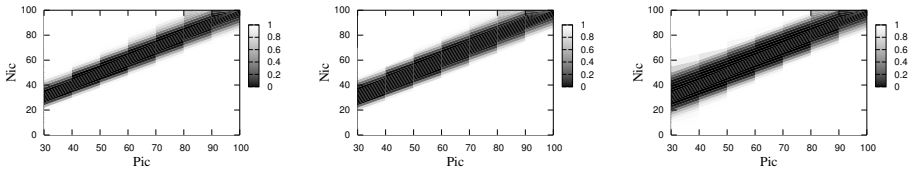


Fig. 7. Fraction of satisfiable CSP (Q) in plane P_{ic}, N_{ic} out of 40 runs. Influence of the size of the near-miss factor of the Universe. **Left:** $nm_U = 0$. **Center:** $nm_U = 15$. **Right:** $nm_U = 25$.

Overall, the Universe is shown to amplify the variations due to the example size, as the instances not related to the target concept now influence the variance of the propositionalized distribution (Fig. 8).

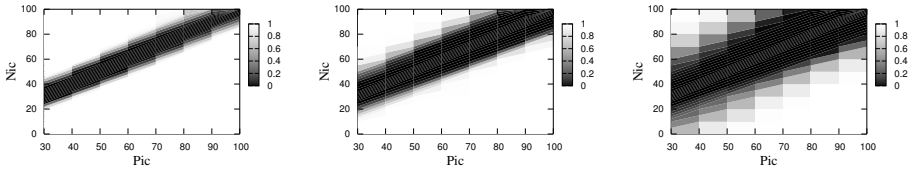


Fig. 8. Fraction of satisfiable CSP (Q) in plane P_{ic}, N_{ic} out of 40 runs. Influence of the size of the example using a Universe. **Left:** $N^+ = N^- = 100$. **Center:** $N^+ = N^- = 200$. **Right:** $N^+ = N^- = 400$.

5 Discussion and Perspectives

The main contribution of this paper is to evidence some Phase Transition-related limitations of MIP 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 MIP kernel.

Clearly, some care must be exercised to interpret the limitations of the well-founded MIP-SVM algorithms suggested by our experiments on artificial problems.

Still, the question of whether MIP-SVM algorithms 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 [12], 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 [1], where the test error could be 2% or lower although the concept learned was a gross overgeneralization of the true target concept.

A research perspective opened by this work is based on the further investigation of the CSP, hybridizing the CSP resolution and the kernel-based propositionalization.

References

1. Botta, M., Giordana, A., Saitta, L., Sebag, M.: Relational learning as search in a critical region. Journal of Machine Learning Research **4** (2003) 431–463.
2. Cheeseman, P., Kanefsky, B., Taylor, W.: Where the really hard problems are. Proc. of Int. Joint Conf. on Artificial Intelligence (1991) 331–337.
3. Cuturi, M., Vert, J.-P.: Semigroup kernels on finite sets. NIPS04 (2004) 329–336.
4. Dietterich, T., Lathrop, R., Lozano-Perez, T.: Solving the multiple-instance problem with axis-parallel rectangles. Artificial Intelligence **89** (1-2) (1997) 31–71.
5. Gärtner, T., Flach, P.A., Kowalczyk, A., Smola, A. J.: Multi-instance kernels. Proc. ICML02 (2006) 179–186.
6. Giordana, A., Saitta, L.: Phase transitions in relational learning. Machine Learning **41** (2000) 217–251.
7. Hogg, T., Huberman, B.A., C., Williams, C.P.: Phase transitions and the search problem. Artificial intelligence **81** (1-2) (1996) 1–15.
8. Kearns, M., Li, M.: Learning in the presence of malicious errors. SIAM J. Comput. **22** (1993) 807–837.
9. Kersting, K., Raedt, L.D.: Adaptative Bayesian logic programs. Proc. of the 11th Int. Conf. on Inductive Logic Programming (2001) 104–117.
10. Kramer, S., Lavrac, N., Flach, P.: Propositionalization approaches to relational data mining. In S. Dzeroski and N. Lavrac (eds.): Relational data mining (2001) 262–291.
11. Kwok, J., Cheung, P.-M.: Marginalized Multi-Instance Kernels. Proc. of the 20th Int. Joint Conf. on Aritificial Intelligence (2007) 901–906.
12. Mahé, P., Ralaivola, L., Stoven, V., Vert, J.-P.: The pharmacophore kernel for virtual screening with support vector machines. Journal of Chemical Information and Modeling **46** (2006) 2003–2014.
13. Muggleton, S., De Raedt, L.: Inductive logic programming: Theory and methods. Journal of Logic Programming **19** (1994) 629–679.
14. Pernot, N., Cornuéjols, A., Sebag, M.: Phase transitions within grammatical inference. Proc. Int. Conf. on Artificial Intelligence (2005) 811–816.
15. Rückert, U., Kramer, S., De Raedt, L.: Stochastic local search in k-term dnf learning. Proc. of the Int. Conf. on Machine Learning (2003) 648–655.
16. Vapnik, V.N.: Statistical learning theory. Wiley-Interscience (1998).