La Programmation Logique Inductive à la lumière de la Transition de Phase

Attilio Giordana[*] — Lorenza Saitta [*]
Michele Sebag [**] — Marco Botta [***]

[*] *DISTA, Università del Piemonte Orientale, Alessandria, Italy*
[**] *LMS, École Polytechnique, Palaiseau, France*
[***] *Dipartimento di Informatica, Università di Torino, Italy*

Attilio.Giordana@unipmn.it, Lorenza.Saitta@di.unito.it,
Michele.Sebag@polytechnique.fr, Marco.Botta@di.unito.it

RÉSUMÉ. *En logique du premier ordre, le test de couverture correspond à un problème de satisfaction de contraintes. Or, dans le domaine de la satisfaction de contraintes, les problèmes les plus difficiles à résoudre en moyenne se concentrent dans une étroite région, appelée transition de phase, qui marque la transition entre les problèmes presque surement satisfiables, et les problèmes presque surement insatisfiables.*
Pour étudier la complexité et la faisabilité de la Programmation Logique Inductive (PLI), nous avons généré quelques centaines de problèmes d'apprentissage, situés dans et hors de la transition de phase ; ces problèmes ont été soumis à FOIL, SMART+ et G-Net.
Ces expérimentations systématiques établissent deux résultats. Tout d'abord, la transition de phase constitue un attracteur pour l'apprentissage, dans le sens où toutes les hypothèses produites appartiennent à cette région. En second lieu, une "zone aveugle de l'apprentissage" apparaît : pour tout problème situé dans cette zone, les systèmes de PLI considérés échouent tous à apprendre quelque hypothèse pertinente que ce soit.
Ces résultats sont interprétés, et la discussion conduit à remettre en cause les biais usuels de la PLI.

ABSTRACT. *A key step of relational learning is testing whether a candidate hypothesis covers a given example. The covering test is equivalent to a Constraint Satisfaction Problem (CSP), which shows a phase transition for critical values of some order parameters. This paper investigates the effects of the phase transition in the covering test on the feasibility and scalability of learning in first order logic languages. Several hundreds of artificial learning problems have been generated. FOIL and other learners have been applied to these problems. The experiments show the presence of a failure region, where all considered learners systematically fail to identify the target concept. Furthermore, the phase transition region behaves as an attractor for the learning search. Interpretations of these findings are proposed and discussed.*

MOTS-CLÉS : *PLI, CSP, complexité, transition de phase, passage à l'échelle, problèmes artificiels.*

KEYWORDS: *ILP, CSP, complexity, phase transition, ILP scalability, artificial problems.*

1. Introduction

La programmation logique inductive (PLI) s'intéresse à l'apprentissage supervisé à partir d'exemples exprimés en logique des prédicats du premier ordre (LPO) [QUI 90, MUG 94]. En LPO, le test de couverture – savoir si une hypothèse couvre un exemple – est équivalent à un problème de satisfaction de contraintes [PRO 96], soit un problème NP difficile. Ce fait peut jeter un certain doute sur les capacités pratiques et le passage à l'échelle de la PLI.

Cependant, il est connu que les instances d'une catégorie de problème NP difficile ne sont pas toutes également difficiles à résoudre [CHE 91]. Dans le domaine de la recherche combinatoire, on montre que les instances les plus difficiles (en moyenne) se concentrent dans une région étroite appelée *transition de phase* [HOG 96]. Cette région marque la transition entre la région des instances presque surement (p.s.) satisfiables, et la région des instances p.s. insatisfiables. Dans ces deux dernières régions, le coût moyen de résolution des problèmes est bien en-dessous de leur complexité au pire cas, soit qu'une instance admette de nombreuses solutions – ce qui rend aisé d'en trouver une – soit qu'une instance soit très contrainte – ce qui rend aisé de prouver qu'elle n'admet pas de solution.

En recherche combinatoire, l'étude de nouveaux algorithmes se concentre ainsi sur leur comportement dans la transition de phase, plus riche d'enseignements pratiques que leur complexité au pire cas. Notre objectif dans cet article est d'examiner la programmation logique inductive à la lumière de la transition de phase (TP).

Dans un article antérieur [GIO 00], l'existence de la TP a été montrée empiriquement sur des problèmes artificiels ou issus du domaine de la mutagenèse [KIN 95], et la TP a été localisée. Dans un second temps, la complexité de l'apprentissage dans la TP a été étudiée, comparant les performances respectives de stratégies déterministes et stochastiques [BOT 99]. Le présent article s'intéresse plus généralement à l'impact de la transition de phase, sur les performances et les capacités de passage à l'échelle (scalability) de la PLI.

La démarche adoptée est expérimentale. Plusieurs centaines de problèmes artificiels échantillonnant les trois régions (p.s. satisfiable, TP et p.s. insatisfiable), ont été construits. Sur ces problèmes, FOIL [QUI 90] a été appliqué ; les résultats obtenus ont été confirmés par des expériences complémentaires fondées sur SMART+ [BOT 93] et G-Net [ANG 98]. Ces expériences systématiques établissent deux résultats. Tout d'abord, la transition de phase constitue un attracteur pour l'apprentissage, dans le sens où les systèmes de PLI finissent par explorer cette région, et y terminent leur recherche quelque soit le concept à apprendre et la stratégie d'apprentissage utilisée. En second lieu, une "zone aveugle de l'apprentissage" apparaît : aucun des trois algorithmes considérés n'apprend d'hypothèse pertinente, dont l'efficacité soit meilleure qu'un diagnostic aléatoire, pour les problèmes de cette zone.

Ces résultats sont discutés et interprétés par rapport aux biais d'apprentissage standard en PLI.

2. Transition de Phase & PLI

Cette section situe brièvement la transition de phase dans le contexte de la PLI.

Nous nous restreindrons au cas le plus simple de l'apprentissage en LPO. Dans la suite, les notations α_i, x_j, v_k désignent respectivement des symboles de prédicats, des variables et des constantes du domaine d'application. Le concept cible $\mathcal{C}$ est une conjonction de littéraux $\alpha_i(x_{i_1}, \ldots x_{i_K})$ quantifiés existentiellement ; une hypothèse est également une conjonction de littéraux quantifiés existentiellement.

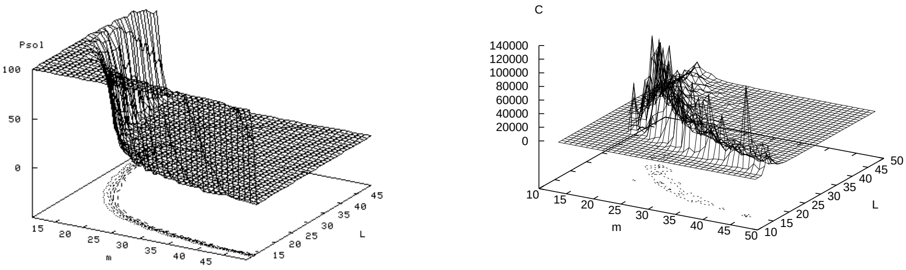
Un exemple E est une conjonction de littéraux totalement instanciés $\alpha_i(v_{i_1},..v_{i_K})$; E est un exemple positif si et seulement si il contient un modèle de $\mathcal{C}$.

Par analogie avec les problèmes de satisfaction de contraintes, nous appelons *relation* l'ensemble des littéraux d'un exemple fondés sur un même symbole de prédicat α_i. Les paramètres décrivant un exemple E sont le nombre L de constantes distinctes apparaissant dans E, et la taille moyenne N des relations de E. Symétriquement, les paramètres décrivant un concept $\mathcal{C}$ sont le nombre m de littéraux et le nombre n de variables distinctes de $\mathcal{C}$.

Prosser a montré que le test de couverture $(\mathcal{C} \prec ?\, E)$, déterminant si $\mathcal{C}$ couvre E au sens de la θ-subsomption [PLO 70], constitue un problème de satisfaction de contraintes caractérisé par le quadruplet (n,N,m,L) [PRO 96]. Or, dans le contexte de la satisfaction de contraintes apparaît un phénomène statistique, la transition de phase, défini par référence à un ensemble d'exemples et de concepts. Les exemples et concepts considérés dans ce qui suit sont construits à partir de distributions uniformes [BOT 99], en se limitant au cas de prédicats binaires :
• Chaque littéral $\alpha_i(x_{i_1},x_{i_2})$ du concept $\mathcal{C}$ est obtenu en tirant x_{i_1} et x_{i_2} uniformément parmi les n variables, sous la condition que $\mathcal{C}$ soit connecté;
• La i-ème relation de E (conjonction des $\alpha_i(v_{i_1,k},v_{i_2,k})$, pour $k = 1..N$) est construite en tirant uniformément et sans remise N paires de valeurs $(v_{i_1,k},v_{i_2,k})$ dans l'ensemble $\{v_1,..v_L\} \times \{v_1,..v_L\}$.

Les résultats expérimentaux obtenus sont résumés dans les figures ci-après. La probabilité P_{cov} que E soit couvert par $\mathcal{C}$, vue comme une fonction de (n,m,N,L), est estimée en considérant plusieurs milliers de tirages de E et $\mathcal{C}$. La Figure 1.a montre P_{cov} en fonction de m et L, pour $n = 4$ et $N = 100$. Il est visible que P_{cov} est voisin de 1 pour les faibles valeurs de m et L (le test de couverture est presque surement vérifié). Symétriquement, P_{cov} est voisin de 0 pour les grandes valeurs de m ou L (le test de couverture n'est presque jamais vérifié). Entre ces deux régions, se manifeste la transition de phase : P_{cov} décroît abruptement de 1 à 0. La complexité de calcul (estimée d'après le nombre de modèles considérés pour savoir si $\mathcal{C}$ couvre E) atteint son maximum dans la transition de phase (Figure 1.b).

Supposons dans ce qui suit que $\mathcal{C}$ et E obéissent à une distribution uniforme, et supposons de plus que n et N sont constants.



(a) $P_{cov} = Pr(\mathcal{C} \prec E)$ (b) Complexité de $(\mathcal{C} \prec E)$

Figure 1. *Transition de phase du test de couverture $(\mathcal{C} \prec E)$ dans le plan $< m,L >$*
m = nb de littéraux de $\mathcal{C}$ L = nb de constantes de E
n = nb de variables de $\mathcal{C}$ (fixé à 4) N = taille des relations de E (fixé à 100)

Le fait que le test de couverture $(\mathcal{C} \prec ?\, E)$ appartienne à la transition de phase dépend à la fois du nombre de littéraux de $\mathcal{C}$ et du nombre de constantes de E. Soit E un exemple donné ; lorsque $\mathcal{C}$ devient plus spécifique (m augmente), le test de couverture se déplace horizontalement vers la droite dans le plan $< m,L >$; on passe d'un test de couverture presque surement vérifié, à un test de couverture presque jamais vérifié. Il est clair que si le nombre m de littéraux est petit, il est aisé de trouver un modèle de $\mathcal{C}$ dans E ; ceci est de moins en moins vrai quand m croît [1].

Les tests de couverture visités durant l'apprentissage correspondent ainsi à un chemin dans le plan $< m,L >$. Dans ce plan, la transition de phase joue un rôle particulier. Soient E et E' deux exemples comprenant L constantes chacun, et soit m_L le nombre critique de littéraux correspondant (i.e., tel que le point (m_L,L) appartienne à la transition de phase). Soit un concept $\mathcal{C}$ comprenant m littéraux ; si m est inférieur à m_L, $\mathcal{C}$ couvre presque surement les deux exemples. Symétriquement, si m est supérieur à m_L, $\mathcal{C}$ rejette presque surement les deux exemples. En résumé, si $\mathcal{C}$ sépare E et E', son nombre de littéraux est presque surement égal à m_L.

3. Protocole expérimental

Le but est d'étudier l'impact de la transition de phase sur les performances de l'apprentissage.

1. Symétriquement, si $\mathcal{C}$ est fixé et que le nombre L de constantes dans E croît, le test de couverture se déplace verticalement vers le haut dans le plan $< m,L >$; on passe d'un test de couverture presque surement vérifié, à un test de couverture presque jamais vérifié.

3.1. *Les problèmes*

Un problème d'apprentissage artificiel π est défini par un concept cible $\mathcal{C}$, un ensemble d'apprentissage et un ensemble de test. Tous les exemples d'un problème π comprennent le même nombre de constantes et le même nombre de littéraux.

451 problèmes artificiels ont été construits. Pour garder un coût de calcul raisonnable, le nombre n de variables de $\mathcal{C}$ est fixé à 4, et la taille N des relations est fixée à 100. Le nombre m de littéraux de $\mathcal{C}$ varie de 5 à 30. Le nombre de littéraux des exemples est de $N \times m = 100m$. Le nombre L de constantes des exemples varie de 11 à 40. L'ensemble des 451 problèmes échantillonne ainsi les régions p.s. satisfiable et p.s. insatisfiable, et la TP (Figure 1.a).

Pour (m,L) donné, le problème d'apprentissage est construit de la façon suivante :
• Le concept cible $\mathcal{C}$ est défini par la conjonction des littéraux $p_i(x_{i_1},x_{i_2})$, avec $i = 1..m$; les variables x_{i_1}, x_{i_2} sont tirées avec probabilité uniforme dans $\{x_1,x_2,x_3,x_4\}$ telles que $\mathcal{C}$ soit connecté.
• Les ensembles de test et d'apprentissage comprennent 200 exemples chacun. Chaque exemple E est défini par m relations correspondant aux prédicats α_i, $i = 1..m$. Une relation est construite par tirage uniforme et sans remise de $N = 100$ paires de valeurs (v_{j_1},v_{j_2}) parmi les L^2 paires de valeurs possibles.
• Cependant, le procédé ci-dessus implique que les exemples sont presque surement tous positifs (couverts par $\mathcal{C}$) lorsque le point (m,L) est en deça de la transition de phase. Symétriquement, les exemples sont presque surement tous négatifs lorsque (m,L) est au delà de la transition de phase. Nous "réparons" donc arbitrairement les exemples, de sorte que tout ensemble de test et d'apprentissage comprenne exactement 100 exemples positifs et 100 exemples négatifs. La réparation d'un exemple négatif E se fait en otant un littéral fondé sur chaque prédicat ; puis, un modèle de $\mathcal{C}$ est généré et ajouté à E, qui devient ainsi positif. La réparation d'un exemple positif E se fait en considérant tous les modèles de $\mathcal{C}$ dans E, et en modifiant l'une des relations de E de façon à prévenir l'existence d'un modèle (voir [BOT 99] pour plus de détails).

3.2. *Les systèmes de PLI : FOIL, SMART+, G-NET*

FOIL effectue une recherche descendante en profondeur d'abord ; il spécialise itérativement l'hypothèse courante $\mathcal{C}_t$ par conjonction avec le meilleur littéral $\alpha_i(x_j,x_k)$ au sens du gain d'information [QUI 90] ou du critère MDL [RIS 78]).
SMART+ effectue une recherche descendante en largeur d'abord [BOT 93]. A la différence de FOIL, il maintient plusieurs hypothèses de front (beam search), la largeur du front étant paramétrée par l'utilisateur.
G-Net utilise un algorithme génétique comme stratégie de recherche [ANG 98]. La population initiale comprend des hypothèses de tailles diverses, qui sont spécialisées ou généralisées de manière aléatoire et selon leur performance.

La plupart des expériences ont été faites avec FOIL ou SMART, G-Net souffrant du surcoût de calcul typique des méthodes d'évolution artificielle[2].

3.3. *Critères d'évaluation*

Notons $\widehat{\mathcal{C}}$ l'ensemble des hypothèses conjonctives découvertes par un système de PLI. Une première mesure de qualité concerne l'efficacité prédictive de $\widehat{\mathcal{C}}$, donnée par le pourcentage d'exemples test correctement classés[3]. Cependant, une bonne qualité de prédiction ne signifie pas que le concept cible $\mathcal{C}$ a été correctement identifié. Un second critère de qualité de l'apprentissage concerne ainsi la structure de $\widehat{\mathcal{C}}$; celle-ci est jugée satisfaisante si elle est proche de celle de $\mathcal{C}$, i.e. si $\widehat{\mathcal{C}}$ comprend une unique hypothèse conjonctive de taille comparable à celle de $\mathcal{C}$.

4. Résultats

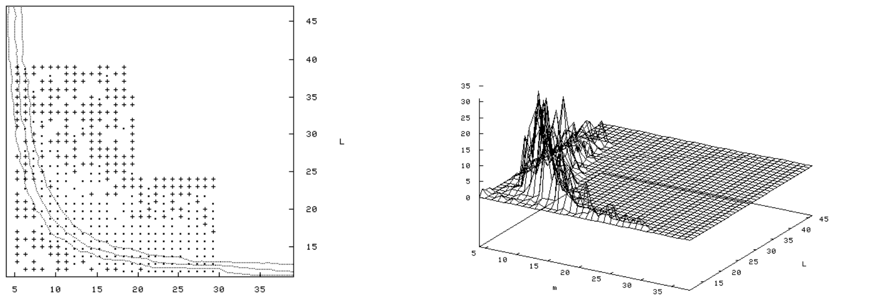
Les résultats obtenus sont évalués selon les critères ci-dessus, qualité prédictive et identification du concept cible. La localisation des hypothèses apprises est également examinée.

4.1. *Qualité prédictive*

La Figure 2.a dresse la "carte des compétences" de FOIL dans le plan $< m,L >$, où les courbes pointillées rappellent la région de la transition de phase ($P_{cov} \in [.1,.9]$). Pour chaque problème, l'apprentissage est couronné de succès (légende $+$) si l'efficacité prédictive est supérieure ou égale à 80% ; au cas contraire, il y a échec (légende $\cdot$). Les résultats détaillés (Table 1) montrent que la qualité de prédiction est soit très élevée ($\geq 95\%$) soit comparable à celle d'un diagnostic aléatoire ($\approx 50\%$). La carte de

2. Quelques essais ont été faits avec PROGOL, qui effectue une recherche descendante utilisant intensivement la connaissance du domaine [MUG 95]. Cependant, PROGOL s'avère peu adapté au traitement d'exemples artificiels de grande taille en l'absence de toute connaissance du domaine.
Quelques essais ont également été faits avec STILL [SEB 97] qui effectue une recherche ascendante fondée sur l'échantillonnage stochastique de l'espace des appariements. Cependant, il s'avère que les heuristiques stochastiques de STILL sont inopérantes face à une distribution uniforme des exemples. Une voie de recherche future concerne l'identification de la distribution des appariements, et l'adaptation en retour des heuristiques de STILL.

3. Nous nous écartons de la procédure usuelle de validation croisée [DIE 98] pour la raison suivante. Les exemples d'apprentissage et de test sont construits selon une même distribution uniforme ; il est donc équivalent d'effectuer une 2CV ou de doubler le nombre de problèmes considérés, ce que nous n'avons pas fait en raison du coût de calcul global déjà fort elevé.
Par ailleurs, même si les résultats indiqués pour un couple (m,L) se fondent sur un seul tirage, ils coïncident très souvent avec les résultats obtenus pour les points (m,L) voisins ; leur stabilité est un gage du fait qu'ils soient significatifs.



(a) Carte des compétences de FOIL
$+$ succès · échec

(b) Histogramme des hypothèses apprises

Figure 2. *Comportement de FOIL dans le plan <m,L>*

compétences est donc largement indépendante du seuil de compétence fixé de 80%.
La portée des résultats obtenus a été confirmée en appliquant SMART+ et G-NET sur une petite fraction des problèmes considérés par FOIL : les trois systèmes échouent ou réussissent de conserve (excepté lorsque SMART+ est lancé avec un nombre d'hypothèses à considérer en parallèle voisin de la taille du concept cible, auquel cas la recherche effectuée est quasi exhaustive).
En résumé, il semble que la région où l'apprentissage échoue, appelée *zone aveugle de l'apprentissage*, ne dépend ni de la stratégie de recherche, ni du seuil d'échec considéré.

Ces expériences tendent à montrer que l'apprentissage en logique du premier ordre est efficace si et seulement si le concept cible est suffisamment court ($m \leq 6$), ou si le problème d'apprentissage est suffisamment loin de la transition de phase. Ce dernier point est contre-intuitif, dans la mesure où il indique qu'apprendre un concept très long peut être plus facile qu'un concept plus court, pour un nombre de constantes L donné. Nous reviendrons sur ce point en Section 5.

Dans un souci de simplicité et par abus de langage, nous dirons que $\mathcal{C}_{.}$ appartient à la transition de phase (resp. est p.s. satisfiable ou p.s. insatisfiable), si le test de couverture ($\mathcal{C}_{.} \prec ?\, E$), où E varie dans la base d'apprentissage ou de test, appartient à la transition de phase (resp. à la région p.s. satisfiable ou p.s. insatisfiable).

4.2. *Identification du concept cible*

Examinons plus en détail les hypothèses apprises par FOIL, en considérant un sous-ensemble représentatif des problèmes considérés. La Table 1 donne tout d'abord les caractéristiques m et L du problème d'apprentissage, puis le nombre K d'hypothèses conjonctives apprises par FOIL, et le nombre moyen $\bar{m}$ de littéraux de ces hypothèses. Ensuite viennent les pourcentages d'exemples bien classés, en apprentissage et en test, puis le coût de l'apprentissage (en secondes, sur une Sparc Enterprise 450). Les deux colonnes suivantes indiquent si la qualité d'identification et de prédiction est ou non satisfaisante. Enfin la dernière colonne donne la catégorie du problème d'apprentissage, discutée ci-dessous.

Tableau 1. *Concept cible $\mathcal{C}$ et Concept appris $\widehat{\mathcal{C}}$*

$\mathcal{C}$		$\widehat{\mathcal{C}}$		Performances			Qualité		
m	L	K	$\hat{m}$	*train.*	*test*	CPU	Identif.	Prédiction	
8	16	1	8	100	100	106.2	O	O	**F**
10	13	1	14	100	99	144.2	O	O	**F**
10	16	8	11.75	88	48.5	783.5	N	N	**D**
11	13	1	11	100	100	92.2	O	O	**F**
11	15	6	13.5	85	53.5	986.2	N	N	**D**
12	13	3	14	98.5	83	516.4	N	O	**f**
$\mathcal{C}$ appartient à la région p.s. satisfiable									
15	29	1	6	100	100	185.3	N	O	**f**
15	35	2	6	97.5	84.5	894.6	N	O	**f**
18	35	1	6	100	100	201.0	N	O	**f**
21	18	8	4.13	81.5	58	1394.9	N	N	**D**
25	24	1	6	100	99	135.9	N	O	**f**
29	17	1	12	100	99.5	144.9	N	O	**f**
$\mathcal{C}$ appartient à la région p.s. insatisfiable									
6	28	12	8.08	91.5	50.5	815.4	N	N	**D**
7	28	11	7.63	91.5	60.5	1034.2	N	N	**D**
8	27	1	7	100	100	58.8	O	O	**F**
13	26	1	9	100	99	476.8	N	O	**f**
17	14	8	15	93	46	294.6	N	N	**D**
18	16	8	8.87	91	58.5	404.0	N	N	**D**
26	12	3	24.33	80	58	361.4	N	N	**D**
$\mathcal{C}$ appartient à la transition de phase									

Les résultats expérimentaux conduisent à distinguer trois catégories de problèmes d'apprentissage :

Problèmes Faciles. FOIL identifie correctement le concept cible $\mathcal{C}$, ou un concept légèrement plus général (un littéral de moins). L'immense majorité des exemples de test et d'apprentissage sont bien classés. Les problèmes faciles sont situés principalement dans la région p.s. satisfiable ; à défaut, ils appartiennent à la transition de phase, pour des valeurs faibles de m.

Problèmes faisables. FOIL détermine une hypothèse conjonctive $\widehat{\mathcal{C}}$, qui classe correctement la très grande majorité des exemples de test et d'apprentissage, *mais surgénéralise* considérablement le concept cible $\mathcal{C}$ (e.g., comprenant 6 au lieu de 18 littéraux).

Les problèmes faisables appartiennent essentiellement à la région p.s. insatisfiable, et

se situent loin de la transition de phase.
Problèmes Difficiles. Ici, FOIL apprend la disjonction $\widehat{\mathcal{C}}$ de nombreuses hypothèses $\mathcal{C}_t$ conjonctives (entre 6 et 15) de taille variable, chacune des $\mathcal{C}_t$ couvrant très peu d'exemples d'apprentissage. La qualité de prédiction de $\widehat{\mathcal{C}}$ sur l'ensemble test est voisine de celle d'un diagnostic aléatoire. Le coût d'apprentissage atteint son maximum pour les problèmes difficiles, d'une part en raison du nombre d'hypothèses apprises, et d'autre part parce que ces hypothèses appartiennent à la transition de phase (voir ci-dessous).
Les problèmes difficiles sont situés dans la transition de phase ou dans son voisinage, pour des valeurs élevées de m.

Ces résultats confirment le fait qu'une bonne qualité de prédiction *ne prouve pas que le bon concept a été découvert*. De toute évidence, il est impossible dans le cas d'une application réelle de faire la différence entre problème facile ou faisable, i.e. de déterminer si le concept cible a été correctement identifié ou largement sur-généralisé.

4.3. *Localisation des hypothèses*

La distribution des hypothèses apprises par FOIL est illustrée par la Figure 2.b. Dans le cas des problèmes faciles, FOIL découvre le concept cible, situé dans la région p.s. satisfiable ; mais ceci n'est pas très visible en raison du faible nombre de problèmes faciles. Dans le cas faisable, FOIL aboutit à une sur-généralisation du concept cible, située dans la transition de phase. Enfin, dans les cas difficiles, FOIL retient un nombre élevé d'hypothèses conjonctives, qui appartiennent en majorité à la transition de phase.
En résumé, la transition de phase constitue un *attracteur* de l'apprentissage : la plupart des hypothèses apprises appartiennent à cette région indépendamment de la position du concept cible.

5. Interprétation

Nous chercherons tout d'abord à expliquer le fait que la transition de phase jouet-elle le rôle d'un attracteur de l'apprentissage. Nous nous intéresserons ensuite à la zone aveugle de l'apprentissage. Enfin nous examinerons le cas où l'apprentissage ne réussit pas à découvrir le concept cible, mais obtient cependant de bonnes performances prédictives.

5.1. *La transition de phase, attracteur de l'apprentissage*

Examinons tout d'abord la stratégie de FOIL. Celui-ci construit une série d'hypothèses $\mathcal{C}_t$ de spécificité croissante, où $\mathcal{C}_1$ comprend un littéral unique. Dans le plan $< m,L >$, si L_π dénote le nombre de constantes des exemples du problème π, les tests de couverture effectués se situent aux points $(1,L_\pi),(2,L_\pi),..(t,L_\pi)$. Les hypothèses considérées tout d'abord sont p.s. satisfiables ; elles appartiennent ensuite à la transition de phase ; la recherche aborde enfin, éventuellement, les hypothèses p.s. insatisfiables. Les contraintes sur les hypothèses retenues sont de deux natures ; $\mathcal{C}_t$ doit en premier lieu être suffisamment représentative, i.e. couvrir un nombre suffisant d'exemples positifs ; l'hypothèse $\mathcal{C}_T$ finalement retenue doit de surcroît être suffisamment correcte, i.e. couvrir peu ou pas d'exemples négatifs. Considérons les implications de cette stratégies en fonction du concept cible $\mathcal{C}$.

Cas 1: $\mathcal{C}$ appartient à la transition de phase.
Par construction, il n'est pas nécessaire de réparer les exemples (Section 3) pour obtenir une équidistribution des exemples positifs et négatifs : si $\mathcal{C}$ appartient à la TP, la probabilité $Pr(\mathcal{C} \prec E)$ est voisine de .5 par construction. En conséquence :

• Aucune hypothèse p.s. satisfiable ne peut être correcte (elle couvre les exemples négatifs). L'exploration se poursuit au moins jusqu'à la transition de phase.

• Symétriquement, aucune hypothèse p.s. insatisfiable ne peut être représentative (elle ne couvre pas les exemples positifs). L'exploration doit donc s'arrêter au tout début de la région p.s. insatisfiable et de préférence avant, i.e. dans la transition de phase.

Les hypothèses retenues appartiennent donc obligatoirement à la transition de phase.

Cas 2: $\mathcal{C}$ appartient à la région p.s. insatisfiable.
Dans ce cas, les exemples négatifs n'ont pas eu besoin d'être réparés ; par le même argument que ci-dessus, aucune hypothèse p.s. satisfiable ne peut donc être correcte, et l'exploration doit au moins continuer jusqu'à la transition de phase. D'un autre côté, une hypothèse p.s. insatisfiable est p.s. correcte ; il n'y a donc aucune raison de continuer la recherche.
Les hypothèses retenues appartiennent donc obligatoirement à la transition de phase, ou au bord de la région p.s. insatisfiable.

Cas 3: $\mathcal{C}$ appartient à la région p.s. satisfiable.
La situation est différente ici, puisqu'il existe des hypothèses correctes p.s. satisfiables, à savoir $\mathcal{C}$ soi-même, et éventuellement certaines de ses généralisations. Si ces hypothèses correctes sont découvertes (la probabilité d'une telle découverte est discutée ci-après), l'exploration s'arrête. En tout état de cause, l'exploration doit s'arrêter avant la région p.s. insatisfiable : puisque les exemples positifs n'ont pas eu besoin d'être réparés, une hypothèse p.s. insatisfiable ne peut les couvrir et n'est donc pas représentative. En résumé une recherche descendante produit des hypothèses p.s. satisfiables ou situées dans la transition de phase.

D'après ce raisonnement détaillé, dans tous les cas l'apprentissage descendant a de fortes chances d'aboutir dans la transition de phase, qui constitue ainsi un attracteur. Excluons maintenant le cas où le concept cible appartient à la région p.s. satisfiable. Il est de fait que *toutes les approches existantes en PLI font intervenir un biais de simplicité* : ce biais conduit à privilégier les hypothèses correctes les plus courtes.

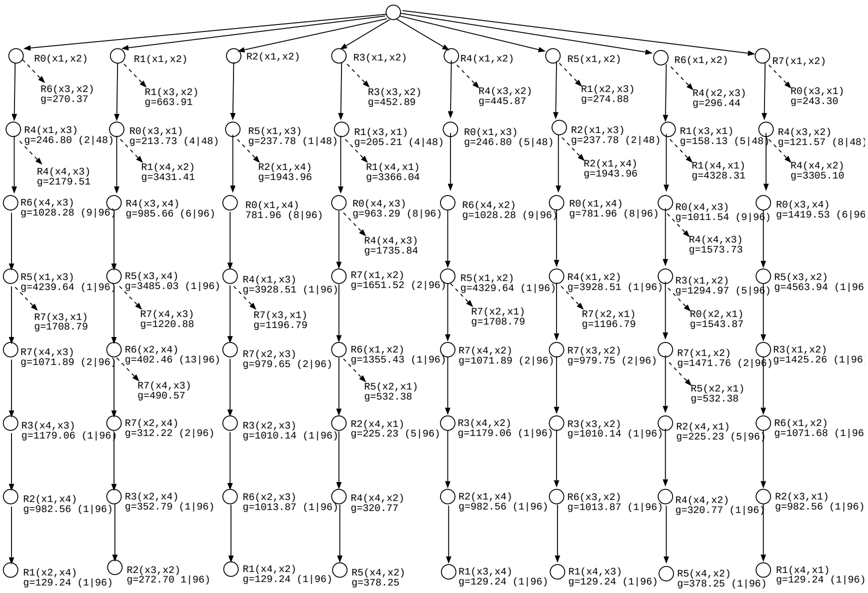


Figure 3. *Exploration descendante : les choix erronés fondés sur le gain d'information ($\searrow$)*

Celles-ci, dans la mesure où nous avons exclu le cas (facile) où le concept cible est p.s. satisfiable, appartiennent à la transition de phase. Ainsi, compte tenu du biais de simplicité et quelle que soit la stratégie de recherche utilisée par ailleurs, la transition de phase constitue un attracteur de la PLI. $\square$

5.2. *Quand l'apprentissage échoue*

Supposons tout d'abord que le concept cible $\mathcal{C}$ appartient à la transition de phase. Nous avons rencontré deux cas de figure (Table 1) : soit $\mathcal{C}$ est court ($m \leq 6$) et il est correctement identifié ; ou bien, de nombreuses hypothèses sont apprises, chacune d'entre elles couvre peu d'exemples positifs, et le diagnostic obtenu sur la base de test est voisin de 50%.

L'échec de l'apprentissage dans le cas d'un concept long est illustré sur l'exemple du concept cible $\mathcal{C}$ ci-dessous :

$$\alpha_0(x_1,x_2) \wedge \alpha_1(x_2,x_3) \wedge \alpha_2(x_2,x_3) \wedge \alpha_3(x_3,x_4) \wedge$$
$$\alpha_4(x_1,x_4) \wedge \alpha_5(x_1,x_4) \wedge \alpha_6(x_3,x_4) \wedge \alpha_7(x_3,x_4)$$

Les hypothèses successivement retenues par FOIL sont montrées sur la Figure 3. Le choix du premier littéral $\mathcal{C}_1$ est aléatoire. En effet le gain d'information ne permet pas de départager les candidats : tout exemple, positif ou négatif, comprend 100 littéraux fondés sur chaque symbole de prédicat. Ceci ne pénalise toutefois pas la recherche puisque tous les prédicats sont pertinents (apparaissent dans le concept cible) par construction [4].

En fonction du premier littéral choisi, les littéraux suivants sont triés d'après leur gain d'information. Le hic vient du fait que le meilleur littéral au sens de ce critère *est faux*, i.e. l'hypothèse $\mathcal{C}_2$ correspondante ne conduit pas au concept cible $\mathcal{C}$ ($\mathcal{C}_2 \not\prec \mathcal{C}$). A partir de ce moment, la recherche ne peut que s'égarer – ou backtracker... En d'autres termes, le critère "maximiser le gain d'information" fourvoie la recherche ; et celle-ci se fourvoie dans 7 des 8 cas considérés, correspondant aux 8 choix possibles pour $\mathcal{C}_1$. Chacune des flèches obliques de la Figure 3 indique que le littéral maximisant le gain d'information (conditionné par l'hypothèse courante, i.e. les noeuds précédents de l'arbre), est *faux*. Afin d'analyser en profondeur le phénomène et les défauts du critère, nous avons chaque fois "réparé" autoritairement le choix, en forçant la sélection du meilleur littéral correct (i.e. le meilleur au sens du gain d'information tel que $\mathcal{C}_t$ soit une généralisation de $\mathcal{C}$).
Globalement, le gain d'information apparait peu fiable : il faut attendre le niveau quatre de l'arbre (i.e. il faut avoir forcé le choix de quatre littéraux corrects), pour que le critère de gain d'information conduise au bon concept cible. En tout état de cause, l'apprentissage descendant est presque certain d'échouer dans un tel cas de figure : les embranchements erronés sont nombreux sur toutes les branches de l'arbre, et de plus ils sont situés aux noeuds supérieurs de l'arbre. La quantité de backtrack nécessaire pour éviter ces embranchements erronés est donc énorme.

Ce résultat s'explique par le fait que le gain d'information dépend du nombre de modèles d'une hypothèse candidate. Or, toute hypothèse p.s. satisfiable admet de nombreux modèles dans tout exemple. Le gain d'information d'un littéral donné est donc peu significatif – sauf lorsque l'hypothèse courante est proche du concept cible. Ceci se produit lorsque le concept cible est lui-même court, ou lorsque l'hypothèse courante est à la fois correcte et suffisamment spécifique. Dans une application réelle, il faudrait donc que l'expert puisse aider considérablement le système (e.g. fournissant quatre littéraux corrects pour un concept cible de huit littéraux), pour se trouver dans ce cas favorable. La variance du nombre de modèles est un facteur supplémentaire brouillant le choix des littéraux ; la variance atteint son maximum lorsque l'hypothèse courante atteint la transition de phase.

L'étude systématique décrite Figure 3 a été effectuée sur les problèmes voisins de la transition de phase. Pour chaque problème d'apprentissage de coordonnées (m,L), on note $t_{m,L}$ la taille minimale de l'hypothèse correcte fournie à FOIL, garantissant que l'on aboutisse au concept cible en suivant le critère de gain d'information. La Figure 4 indique pour chaque point (m,L) le nombre $t_{m,L}$ correspondant. Cette figure peut ainsi être vue comme la carte de fiabilité du gain d'information (plus $t_{m,L}$ est élevé et moins le gain d'information est fiable). De manière équivalente, cette figure

4. Notons que dans une application réelle, le choix du premier littéral se fait sur la base d'une information purement attribut-valeurs.

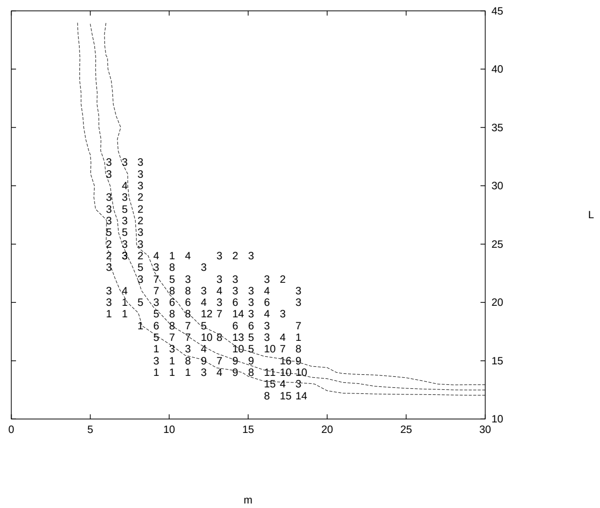


Figure 4. *Taille minimum de $\mathcal{C}_t$ garantissant la fiabilité du gain d'information.*

indique la somme critique de connaissances a priori dont doit disposer FOIL pour que l'apprentissage puisse aboutir.

Ainsi, $t_{m,L}$ prend des valeurs élevées dans la transition de phase et au voisinage de la région p.s. insatisfiable, ce qui explique l'échec systématique de FOIL dans cette région (la probabilité d'éviter tout choix erroné décroît exponentiellement en $t_{m,L}$). Il nous reste à expliquer pourquoi $t_{m,L}$ décroît lorsque l'on s'éloigne de la transition de phase.

5.3. *Quand l'apprentissage trouve une bonne approximation de $\mathcal{C}$*

D'après ce qui précède, les chances de succès de l'apprentissage sont minces lorsque la taille m du concept cible, ou le nombre L de constantes du domaine d'application, sont élevées. Pourtant, lorsque m et L sont tous deux élevés, FOIL détermine des hypothèses de qualité prédictive très satisfaisantes (Table 1).

L'explication proposée est la suivante. Supposons que le concept cible $\mathcal{C}$ soit p.s. insatisfiable (m et L élevés).

Montrons alors que toute généralisation $\mathcal{G}$ de $\mathcal{C}$ située dans la région p.s. insatisfiable classe p.s. correctement tout exemple d'apprentissage ou de test. En effet, si $\mathcal{C}$ est p.s. insatisfiable, les exemples négatifs suivent une distribution uniforme ; tout concept p.s. insatisfiable, en particulier $\mathcal{G}$, les discrimine. Par ailleurs, si $\mathcal{G}$ généralise $\mathcal{C}$, il couvre tout exemple positif. $\mathcal{G}$ est donc correct et complet. $\square$

Il s'ensuit que, si l'exploration tombe sur une généralisation $\mathcal{G}$ de $\mathcal{C}$ proche de la région p.s. insatisfiable, la recherche s'arrête ($\mathcal{G}$ est de bonne qualité sur la base d'apprentissage) et la validation est très satisfaisante ($\mathcal{G}$ est de bonne qualité sur la base de test). Le succès de l'apprentissage relationnel, en terme d'efficacité prédictive, dépend donc de la probabilité de trouver une généralisation $\mathcal{G}$ de $\mathcal{C}$ qui soit p.s. insatisfiable ou située au voisinage de la transition de phase.

Soit m le nombre de littéraux de $\mathcal{C}$. Le nombre $H(v,m)$ de généralisations of $\mathcal{C}$ comprenant v littéraux, a été déterminé de manière analytique, et il atteint son maximum pour $v = \frac{m}{2}$. Ceci peut expliquer pourquoi l'apprentissage relationnel obtient souvent d'excellents résultats lorsque le concept cible est suffisamment long, et plus précisément, lorsque la taille m de $\mathcal{C}$ est supérieure à deux fois la taille critique (pour L donné, la taille critique m_L est telle que le point (m_L,L) appartienne à la transition de phase). Le nombre de généralisations de $\mathcal{C}$ dans la transition de phase étant exponentiel en fonction de m, la chance d'en trouver une devient raisonnable.

6. Conclusion

De récentes avancées en recherche combinatoire montrent qu'il est plus instructif d'étudier le comportement pratique d'un algorithme sur les problèmes "en moyenne les plus difficiles", i.e. situés dans la transition de phase, que de se focaliser sur sa complexité au pire cas [HOG 96].

Nous inspirant de cette démarche, nous avons étudié de manière systématique le comportement de trois algorithmes de PLI sur un grand nombre de problèmes artificiels, échantillonnant la transition de phase ainsi que les régions p.s. satisfiable et insatisfiable. En dépit des simplications faites (distribution uniforme des exemples, prédicats tous pertinents, concepts cibles uniquement conjonctifs), cette étude éclaire certaines des limitations actuelles de la PLI.

Le premier résultat obtenu est que la transition de phase constitue un attracteur de l'apprentissage : la majorité des hypothèses produites appartient à cette région, indépendamment de la stratégie de recherche et de la position du concept cible. Ce résultat est (a posteriori) naturel, dans la mesure où la transition de phase concentre les hypothèses capables de séparer deux exemples quelconques. Un corollaire en est que la PLI ne peut contourner la complexité exponentielle du test de couverture – ce qui jette quelque doute sur les capacités de passage à l'échelle de la PLI.
Le second résultat empirique concerne la fiabilité du critère de gain d'information, largement utilisé dans la littérature. Le rapport "signal à bruit" de ce critère apparaît très faible, spécialement dans les premières étapes d'une recherche descendante. Ceci vient du fait que toute hypothèse courte admet un très grand nombre de modèles dans tout exemple.
En troisième lieu, une large "zone aveugle de la PLI" est mise en évidence : sur tous les problèmes situés dans cette zone, les trois systèmes de PLI considérés échouent.

Ces résultats nous conduisent à reconsidérer les points d'assise de la PLI. Ainsi, de nouveaux biais d'apprentissage semblent nécessaires pour faire face à de larges applications, et apprendre des concepts plus complexes (plus de quatre variables non déterminées). Ceci passe par l'élaboration de nouveaux critères de recherche descendante, moins dépendants du nombre de modèles des hypothèses ; une alternative consisterait

à reconsidérer les stratégies de recherche ascendante.
Plus fondamentalement, il apparait raisonnable de remettre en cause l'espace et les opérateurs de recherche de la PLI. Typiquement, si l'espace pertinent est celui de la transition de phase, la question centrale est de restreindre la recherche à cet espace, et de déterminer les opérateurs permettant d'y naviguer.

7. Bibliographie

[ANG 98] ANGLANO C., GIORDANA A., LOBELLO G., SAITTA L., « An experimental Evaluation of Coevolutive Concept Learning », *Proceedings of the 15th International Conference on Machine Learning*, 1998, p. 19–23.

[BOT 93] BOTTA M., GIORDANA A., « SMART+: A Multi-Strategy Learning Tool », *Proceedings of the Thirteenth International Joint Conference on Artificial Intelligence*, 1993, p. 937-943.

[BOT 99] BOTTA M., GIORDANA A., SAITTA L., « Relational Learning: Hard Problems and Phase Transitions », *Proceedings of the 16th International Joint Conference on Artificial Intelligence*, 1999, p. 1198-1203.

[CHE 91] CHEESEMAN P., KANEFSKY B., TAYLOR W., « Where the Really Hard Problems Are », *Proceedings of the 12th International Joint Conference on Artificial Intelligence*, 1991, p. 331–337.

[DIE 98] DIETTERICH T., « Approximate statistical tests for comparing supervised classification learning algorithms », *Neural Computation*, 1998.

[GIO 00] GIORDANA A., SAITTA L., « Phase Transitions in Relational Learning », *Machine Learning*, 2000, page in press.

[HOG 96] HOGG T., HUBERMAN B., WILLIAMS C., Eds., *Artificial Intelligence: Special Issue on Frontiers in Problem Solving: Phase Transitions and Complexity*, vol. 81(1-2), Elsevier, 1996.

[KIN 95] KING R., SRINIVASAN A., STENBERG M., « Relating Chemical Activity to Structure: an Examination of ILP Successes », *New Generation Computing*, vol. 13, 1995.

[MUG 94] MUGGLETON S., DE RAEDT L., « Inductive Logic Programming: Theory and Methods », *Journal of Logic Programming*, vol. 19, 1994, p. 629-679.

[MUG 95] MUGGLETON S., « Inverse entailment and PROGOL. », *New Gen. Comput.*, vol. 13, 1995, p. 245–286.

[PLO 70] PLOTKIN G., « A note on inductive generalization », *Machine Intelligence*, vol. 5, Edinburgh University Press, 1970.

[PRO 96] PROSSER P., « An empirical study of phase transitions in binary constraint satisfaction problems », *Artificial Intelligence*, vol. 81, 1996, p. 81-110.

[QUI 90] QUINLAN R., « Learning Logical Definitions from Relations », *Machine Learning*, vol. 5, 1990, p. 239-266.

[RIS 78] RISSANEN J., « Modeling by Shortest Data Description », *Automatica*, vol. 14, 1978, p. 465-471.

[SEB 97] SEBAG M., ROUVEIROL C., « Tractable induction and classification in first order logic via Stochastic Matching », *Proceedings of the 15th International Joint Conference on Artificial Intelligence*, 1997, p. 888–893.