

Evolution Artificielle

pour l'apprentissage supervisé

relationnel

Michèle Sebag, LRI, Orsay

Plan

- Définitions
- G-NET
- La barrière de la transition de phase

Apprentissage relationnel

Input

- Exemples

$$\mathcal{E} = \{(y_i \leftarrow x_i)\}$$

Espace des exemples

- Logique des prédicats d'ordre 1
- Clauses définies

Exemple

La mutagenèse

- Domaine de la chimie organique
- Discrimination des molécules cancérigènes

$$\text{carcin}(m) : - \text{atm}(m, m_1, \text{carb}, 22), \text{atm}(m, m_2, \text{hydr}, 3), \text{atm}(m, \\ \text{bind}(m, m_1, m_3), \text{bind}(m, m_2, m_3))$$

carcin, *atm*, *cc*

symboles de prédicat

atm(*m*, *m*₁, *carbon*, 22)

littéral fondé sur le prédicat *atm*

carcin(*m*)

tête de la clause

atm(*m*, *m*₁, *carbon*, 22) . . . ,

corps de la clause

. . .

bind(*m*, *m*₁, *m*₃), *bind*(*m*, *m*₂, *m*₃)

Relation de couverture

La θ -subsumption La clause $\mathcal{C}$ subsume la clause $\mathcal{D}$ ssi il existe une substitution θ , envoyant les variables de $\mathcal{C}$ sur les constantes et variables de $\mathcal{D}$ telle que $\mathcal{C}\theta \subseteq \mathcal{D}$.

Exemple

$$\begin{aligned} \text{carcin}(m) \quad : - \quad & \text{atm}(m, m_1, \text{carb}, 22), \text{atm}(m, m_2, \text{hydr}, 3), \\ & \text{atm}(m, m_3, \text{carb}, 21) \\ & \text{bind}(m, m_1, m_3), \text{bind}(m, m_2, m_3) \end{aligned}$$

$$\begin{aligned} \text{carcin}(X) \quad : - \quad & \text{atm}(X, Y, \text{carb}, Z), \text{atm}(X, Y', \text{carb}, Z') \\ & \text{bind}(X, Y, Y') \end{aligned}$$

$$\theta = \{X/m, Y/m_1/, Y'/m_3, Z/22, Z'/21\}$$

Relation de couverture, 2

par rapport au cas propositionnel

Difficultés additionnelles

- θ -subsumption de complexité exponentielle (pire cas)
- Deux clauses distinctes peuvent être équivalentes
- Implication sur l'exploration de l'espace des hypothèses :
 - Eviter la redondance (explorer plusieurs fois la même clause)
 - Garantir la complétude (explorer possiblement toutes les clauses)
- La barrière de la transition de phase

Programmation Logique Inductive

Algorithmes

- FOIL : extension de ID3
- PROLOG
- G-Net : EA + co-évolution

Applications

- Chimie
- Langage naturel
- Fouille de données multi-relationnelles

G-NET

Anglano, Giordana et al, 1998

Approche : pb de couverture

- Trouver un ensemble de clauses $\mathcal{C}_1, ..\mathcal{C}_n$
- Chaque clause correspond à une modalité du concept
(un sous-ensemble des exemples positifs)
- Chaque exemple positif est couvert par au moins une clause
- Simplicité – Intelligibilité : Minimiser n

Learning from Structured Examples

Data Description Language



Attributes	obj ₁	obj ₂	obj ₃	obj ₄
position	1	2	3	4
shape	engine	tank-car	open-top	freight-car
color	green	red	blue	green
length	1	1	1	2
Load-type	0	2	4	1
n. wheel	4	4	4	8

Espace de recherche

On se donne un template

$$\begin{aligned}\mathcal{L} = & \textit{Color}(X, [rd, bl, gr, yl]) \wedge \textit{Shape}(X, [sq, tr, cr]) \\ & \wedge \textit{Color}(Y, [rd, bl, gr, yl]) \wedge \textit{Shape}(Y, [sq, tr, cr]) \\ & \wedge \textit{Distance}(X, Y, [0, 1, 2, 3, 4])\end{aligned}$$

→ Optimisation paramétrique

	Color(X)	Shape(X)	Color(Y)	Shape(Y)	Distance(X,Y)
Individu	[1111]	[111]	[1111]	[111]	[11111]

Disjonction interne possible : wagon rouge ou bleu.

Espace de recherche, 2

Utilisation de négation possible

$$\neg \text{Color}(X) \quad \neg \text{Shape}(X) \quad \neg \text{Color}(Y) \quad \neg \text{Shape}(Y) \quad \neg \text{Distance}(X,Y)$$
$$\begin{array}{ccccc} [1] & [0001] & [0] & [111] & [1] & [1000] & [0] & [111] & [0] & [01000] \end{array}$$

$\equiv$ Un wagon non jaune suivi d'un wagon non rouge

Critiqué par la communauté ILP

- Est-ce du vrai ILP !@\$#
- ... biais d'apprentissage ...

Fonction objectif

Minimum Description Length (Rissanen 1978)

Etant donné

- Un ensemble de données D
- Un ensemble de programmes $h_1, \dots, h_n$

Le meilleur programme h_i est celui qui minimise

- le nombre de bits nécessaire pour coder h_i
- + le nombre de bits pour coder les erreurs de h_i .

Remarques

- Intérêt : évaluation d'une hyp. ou d'un ensemble d'hyp.
- Possible par le template.

Schéma d'évolution

- Un schéma distribué (island)
- Chaque ile est (temporairement) associée à un exemple
- Les règles migrent
- Une mémoire globale stocke les meilleures règles
- Equilibrer la pression d'apprentissage sur les exemples
- Opérateurs de variation dédiés

Initialisation

Seeding

- Tirer des exemples positifs
- Chaque exemple e_i est alloué à un processeur
- Sous-population i : initialisée d'après e_i
- Degré de généralité initial : paramètre utilisateur
- Clauses initiales un minimum pertinentes : couvrent e_i

Importance de l'initialisation !

Daida GECCO 99

Importance de la diversité

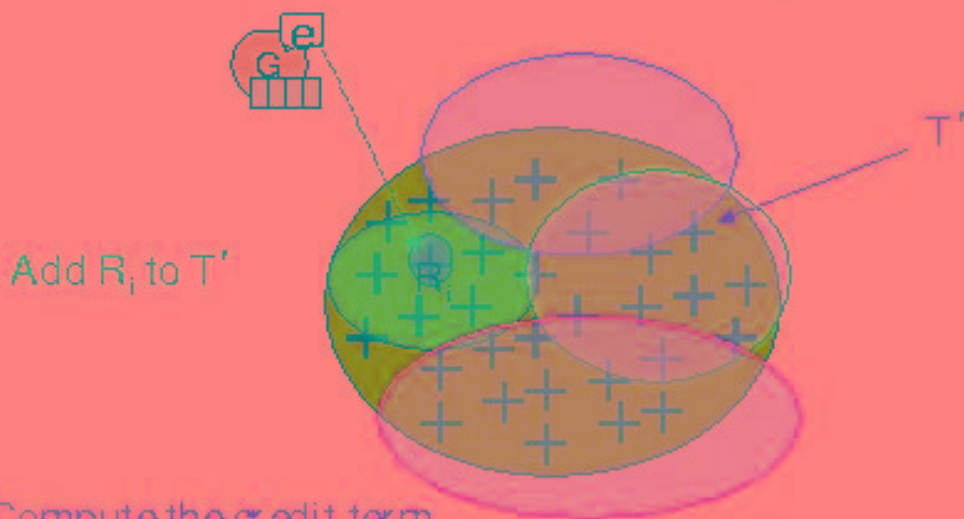
Usage de l'évolution

- Si les parents sont différents, croisement, sinon mutation
- Forcer les enfants différents des parents

Usage de la co-évolution

- Chaque exemple positif $\equiv$ un objectif, une population
- Répartir les efforts :
 - Allocation d'efforts aux objectifs
 - Allocation de crédits aux solutions

How G-nodes Account for Cooperation



$$f_L(R_i) = f(R_i) + f(T' + R_i) - f(T)$$

Opérateurs de variation

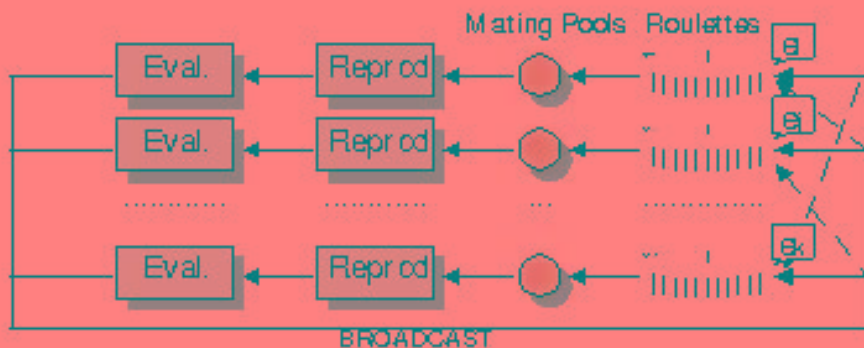
Croisement

- Appariement guidé
- probabilité p : croisement à deux points
- sinon, si parents corrects, croisement généralisant (OU)
- sinon croisement spécialisant (ET)

Mutation

- probabilité p' : seeding
- sinon, si parent correct, mutation généralisante
- sinon mutation spécialisante

1) Guided Mating



Individuals (Horn clauses) enter in a population G_i only when they cover the positive example e_i

Clauses sharing many examples have many chances of crossing
 Mutual exclusive clauses have little chances of crossing

Experimental Evaluation

COMPARISON WITH SMOG AND C4.5

? several standard (simple) benchmarks
from the U.C. Irvine repository
? first run results

Problem	Dataset size	Average Error %			Avg. N. of Rules
		G-NET	Smog	C4.5	G-NET
monk1	432 10-fold	0.00 <u>+</u> 0.00	0.00 <u>+</u> 0.00	0.00 <u>+</u> 0.00	3.00
monk2	432 10-fold	2.80 <u>+</u> 3.80	0.00 <u>+</u> 0.00	32.8 <u>+</u> 10.7	26.0
monk3	432 10-fold	0.00 <u>+</u> 0.00	0.00 <u>+</u> 0.00	0.00 <u>+</u> 0.00	3.00
tictactoe	958 10-fold	0.97 <u>+</u> 0.62	2.82 <u>+</u> 1.97	7.07 <u>+</u> 1.82	10.5
credit	690 10-fold	15.8 <u>+</u> 4.40	19.6 <u>+</u> 5.08	14.0 <u>+</u> 3.28	14.0
breast	699 10-fold	5.29 <u>+</u> 2.89	6.72 <u>+</u> 2.44	5.85 <u>+</u> 3.32	2.60
vote	435 10-fold	5.10 <u>+</u> 3.20	5.29 <u>+</u> 2.64	4.63 <u>+</u> 3.05	2.00

Experimental Evaluation

COMPARISON WITH STILL AND SMART+FONN

?FOL complex tasks
?first run results

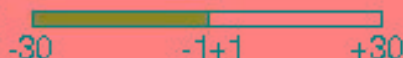
Problem	Dataset size	Average Error %			Avg. N. of Rules
		G-NET	STILL	SMART + FONN	G-NET
mutagenesis	230 10-fold	8.80 <u>+</u> 7.90	6.40 <u>+</u> 4.50	n.a	3
office-documents	210 + 160	0.89 <u>+</u> 0.72	n.a	0.80	11
Train check-out n.3	500 + 6000	11.3 <u>+</u> 0.47	n.a	16.8	2

Experimental Evaluation THE SPLICE-JUNCTIONS DATASET

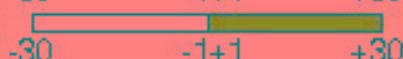
Given a window of 60 DNA base-pairs

Decide if it is a

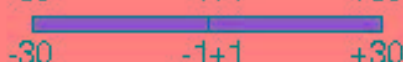
a) Exon-Intron boundary (E/I)



b) Intron-Exon boundary (I/E)



c) Neither (N)



?Complex propositional task

?3 classes (E/I, I/E, N)

?3190 examples (Genbank 64.1)

?Learning set: 2000 examples

?Test set: 1190 examples

		Average Error %		N. of Rules	
		G-NET	REGAL	G-NET	REGAL
E/I		3.40 %	4.40 %	7	19
I/E		2.90 %	4.20 %	10	26
N		3.30 %	5.20 %	11	21

Conclusion sur G-Net

Forces

- Apprentissage efficace, solutions intelligibles
- Coût (parallélisation)

Faiblesses

- Choix du template
- Pression d'apprentissage : récompense le bruit ?

Disponible

- Single process : www.di.unito.it/attilio/gnet
- Multi-process : PVM, MPI

Et maintenant, une histoire d'échec...

Intérêt de parler des échecs

- Evite les tromperies
- Atteste la maturité du champ
- Situe les near-miss : un bon objet d'apprentissage
- ... mais difficile à publier ...

AI winter

ILP et Transition de phase

Quel Rapport ?

Déduction Logique: $\equiv \theta$ -subsumption

C θ -subsume D si $\exists \theta / C\theta \subseteq D$

$C :$ $p_0(x_0, x_1), \quad p_1(x_0, x_3), \quad p_2(x_1, x_3)$

$D :$ $p_0(a_2, a_7), \quad p_1(a_1, a_7), \quad p_2(a_1, a_4),$
 $p_0(a_2, a_7), \quad p_1(a_3, a_7), \quad p_2(a_2, a_7),$

..

$\Leftrightarrow$ **Satisfaction de Contraintes**

Contrainte :

$$\boxed{p_0(x_0, x_1) \wedge p_1(x_0, x_3) \wedge p_2(x_1, x_3)}$$

Relation $p_0 :$

Affectation $\theta :$

Solution :

$$\begin{aligned} & \{p_0(a_1, a_2), p_0(a_2, a_7), ..\} \\ & \{x_0, x_1, .., x_n\} \rightarrow \{a_0, a_1, .., a_L\} \\ & p_0(\theta(x_0), \theta(x_1)) \in \text{Relation } p_0 \\ & p_1(\theta(x_0), \theta(x_3)) \in \text{Relation } p_1 \\ & p_2(\theta(x_1), \theta(x_3)) \in \text{Relation } p_2 \end{aligned}$$

CSP et Transition de Phase

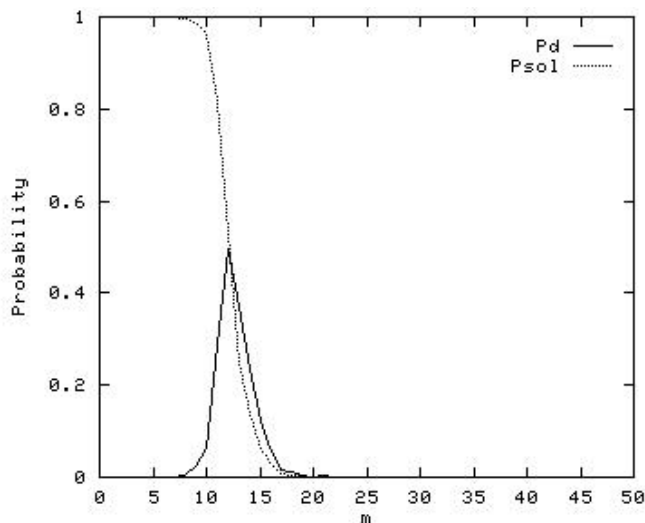
Les problèmes sont tous exponentiels (pire cas)
mais certains sont plus difficiles que d'autres...

m	nb de contraintes	N	taille de relation
n	nb de variables	L	nombre de constantes

Paramètres d'ordre: $p_1 = \frac{2m}{n(n-1)}$
 $p_2 = 1 - \frac{N}{L^2}$

Région satisfiable — Transition de Phase — Région insatisfiable

Transition de Phase : les problèmes les plus difficiles en moyenne.



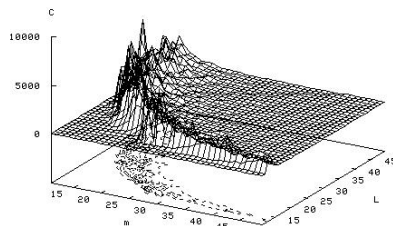
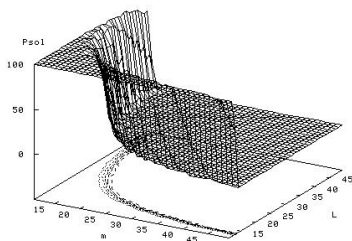
Transition de Phase et Appariement

- Le problème existe-t-il ?

OUI

Botta et al, IJCAI99

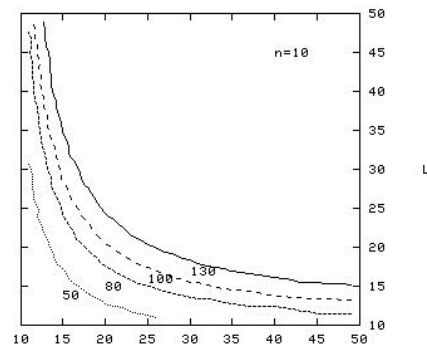
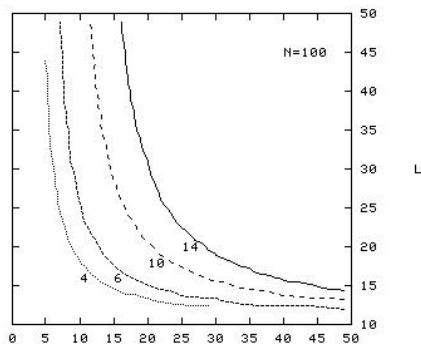
Probabilité($C \prec D$) et Coût de la subsumption



Position de la Transition de Phase

a) $n = 4, 6, 10, 14, N = 100$

b) $n = 10, N = 50, 80, 100, 130$



Plan de travail

1. Problèmes artificiels

- Région oui
- Transition de phase
- Région non

2. Apprentis

- FOIL
- Smart+
- G-Net

3. Etude : impact de la position du problème

- sur la prédiction
- sur la découverte
- sur le coût

$n = 4$	nb variables	$N = 100$	nb littéraux
m	nb prédicats	L	nb constantes

Problèmes artificiels

1. Choisir m et L
2. Construire le concept $\mathcal{C}$

$$p_1(x_{1,1}, x_{1,2}), \dots, p_m(x_{m,1}, x_{m,2})$$

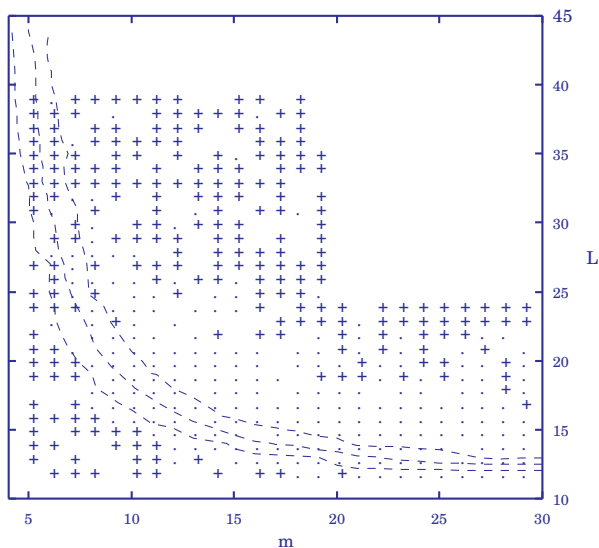
$x_{i,j}$ in $\{x_1, \dots, x_4\} / \mathcal{C}$ connectée

3. Base d'apprentissage et base de test :
200 exemples, 100 positifs, 100 négatifs

$$Ex = \bigwedge_{k=1}^m p_k(a_{1,1}, a_{1,2}), \dots, p_k(a_{N,1}, a_{N,2})$$

Carte des compétences de FOIL

- + Succès ($> 80\%$ sur la base de test)
- Echec



Résultats

$\mathcal{C}$		$\bar{\mathcal{C}}$		Performances			Qualité		
m	L	K	$\hat{m}$	$train.$	$test$	CPU	Exact	%	
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

Problèmes faciles

$\mathcal{C}$		$\hat{\mathcal{C}}$		Performances			Qualité	
m	L	K	$\hat{m}$	<i>train.</i>	<i>test</i>	CPU	Exact	%
8	16	1	8	100	100	106.2	O	O
10	13	1	14	100	99	144.2	O	O
11	13	1	11	100	100	92.2	O	O

$\mathcal{C}$ appartient à la région p.s. satisfiable

8	27	1	7	100	100	58.8	O	O
---	----	---	---	-----	-----	------	---	---

$\mathcal{C}$ appartient à la transition de phase

Problèmes faciles

- Région satisfiable
- Concept cible petit

Problèmes faisables

$\mathcal{C}$		$\hat{\mathcal{C}}$		Performances			Qualité	
m	L	K	$\hat{m}$	$train.$	$test$	CPU	Exact	%
12	13	3	14	98.5	83	516.4	N	O

$\mathcal{C}$ appartient à la région p.s. satisfiable

15	29	1	6	100	100	185.3	N	O
15	35	2	6	97.5	84.5	894.6	N	O
18	35	1	6	100	100	201.0	N	O
25	24	1	6	100	99	135.9	N	O
29	17	1	12	100	99.5	144.9	N	O

$\mathcal{C}$ appartient à la région p.s. insatisfiable

13	26	1	9	100	99	476.8	N	O
----	----	---	---	-----	----	-------	---	---

$\mathcal{C}$ appartient à la transition de phase

Problèmes faisables

- Du bon côté de la TP
- Dans la région insatisfiable,

très loin de la TP.

Problèmes difficiles

$\mathcal{C}$		$\bar{\mathcal{C}}$		Performances			Qualité	
m	L	K	$\hat{m}$	<i>train.</i>	<i>test</i>	CPU	Exact	%
10	16	8	11.75	88	48.5	783.5	N	N
11	15	6	13.5	85	53.5	986.2	N	N

D**D** $\mathcal{C}$ appartient à la région p.s. satisfiable

21	18	8	4.13	81.5	58	1394.9	N	N
----	----	---	------	------	----	--------	---	---

D $\mathcal{C}$ appartient à la région p.s. insatisfiable

6	28	12	8.08	91.5	50.5	815.4	N	N
7	28	11	7.63	91.5	60.5	1034.2	N	N
17	14	8	15	93	46	294.6	N	N
18	16	8	8.87	91	58.5	404.0	N	N
26	12	3	24.33	80	58	361.4	N	N

D**D****D****D****D** $\mathcal{C}$ appartient à la transition de phase

Problèmes durs

- Grands concepts cibles
- Proches de la TP.

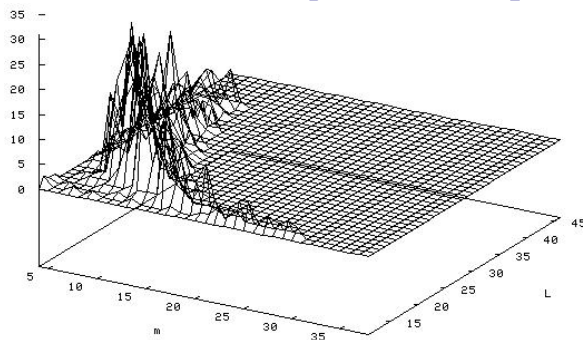
Conclusions expérimentales

- La transition de phase est un attracteur d'apprentissage
- Le paysage relationnel comprend une “tâche aveugle”
- Plus complexe $\nRightarrow$ Plus difficile (?!)

La TP est un attracteur de la PLI

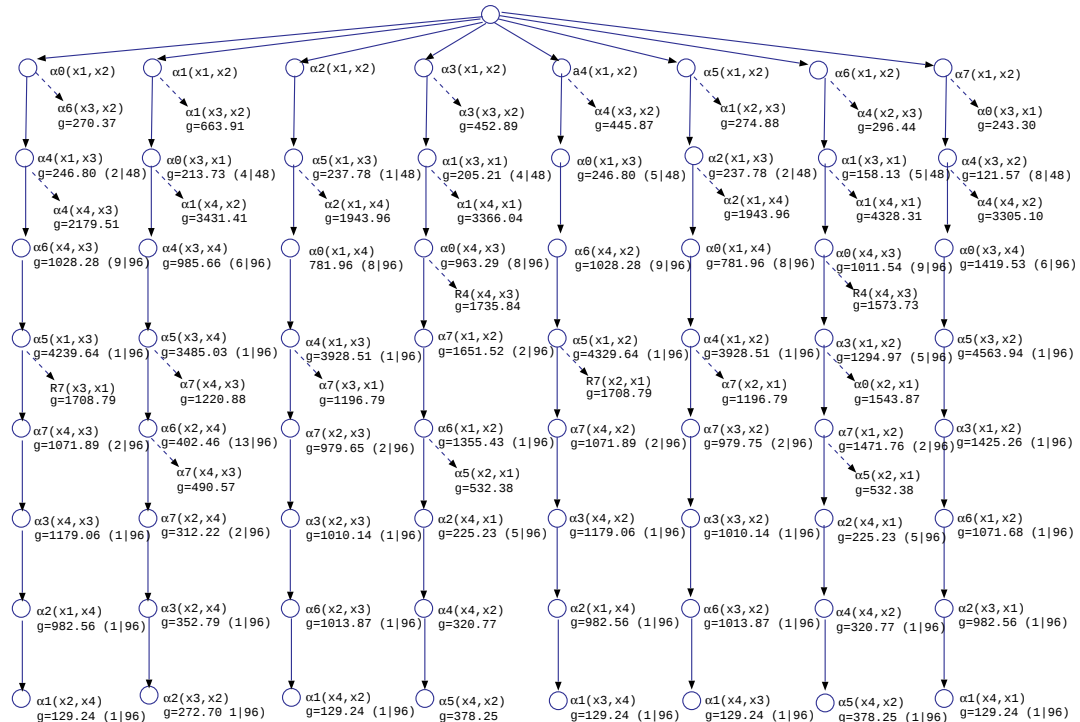
Les hypothèses retenues sont dans la TP

que le concept cible y soit ou non.



La zone aveugle de la PLI

Le gain d'information est un critère trompeur...



Pourquoi plus complexe $\nRightarrow$ plus difficile ?

Bonnes approximations

Soit $\mathcal{C}$ ds la région non satisfiable,

Soit $\mathcal{G} \prec \mathcal{C}$.

$\mathcal{G}$ appartient à la TP (versant insatisfiable)

$\Rightarrow \mathcal{G}$ p.s. correct et complet

Probabilité de les trouver

$$\# \{ \mathcal{G} / \mathcal{G} \prec \mathcal{C}, \mathcal{G} \in TP \}$$

exponentiel en la taille de $\mathcal{C}$.

Perspectives

1. Autres critères (biais de recherche)
2. Autre espace de recherche :
N'explorer que la TP ?
on demande : opérateurs propres, privés, parfaits...
3. Approches ascendantes

Voir Serra et al, IJCAI 01
Ales-Bianchetti et al, ICML 02