

# Moindres généralisés

Fabien Torre

Université de Lille

Mercredi 30 septembre et 7 octobre 2009

# Motivations et Intuitions

## Difficultés pour les arbres de décision

- tester plusieurs attributs à la fois ;
- capturer les problèmes disjonctifs (*tic-tac-toe* par exemple).

x	x	x	x	o	o	x	o	o	positive
x	x	x	o	o	b	b	b	b	positive
x	x	x	o	x	o	x	o	o	positive
x	x	x	?	?	?	?	?	?	positive

... et sur des attributs continus ?

5	1	1	1	2	1	3	1	benign
5	4	4	5	7	10	3	2	benign
[5;5]	[1;4]	[1;4]	[1;5]	[2;7]	[1;10]	[3;3]	[1;2]	benign
6	8	8	1	3	4	3	7	benign
[5;6]	[1;8]	[1;8]	[1;5]	[2;7]	[1;10]	[3;3]	[1;7]	benign

# Plus formellement

## Définition : moindre généralisé

Étant donné un ensemble d'exemples  $E \subseteq \mathcal{X}$ , une hypothèse  $h \in \mathcal{H}$  est dite *moindre généralisée* de  $E$  si et seulement si :

- $\forall e \in E : h \succeq e$  ;
- il n'existe pas  $h'$  vérifiant  $\forall e \in E : h' \succeq e$  et  $h \succeq h'$ .

Cette définition implique-t-elle l'unicité ? Soit  $\mathcal{E} = \mathbb{R}^2$ , pour chaque  $\mathcal{H}$  possible :

- est-ce que le moindre généralisé est unique ?
- comment le calculer ?

Candidats : rectangles, carrés, cercles, etc.

## Deux profils

Deux vues algorithmiques :

- $\text{MG}(e_1, e_2, \dots, e_n \in \mathcal{X})$  returns  $h \in \mathcal{H}$ ;
- $\text{MG}(h_{n-1}, e_n)$  returns  $h \in \mathcal{H}$ .

on préfère la deuxième version, plus efficace pour l'apprentissage.

La suite : utiliser les classes !

## Intuitions

Généralisation d'exemples d'une même classe sans couvrir aucun exemple d'une autre classe.

×	×	×
×		○
	○	○

×	×	×
	○	
○	○	×

moindre généralisé →

×	×	×
?	?	?
?	○	?

×		○
×	○	×
×		○

×	?	?
×	?	?
?	?	○

subsume →

×		×
×	×	
○	○	○

## Calcul [Torre, 1999]

une graine et sa classe	exemples de la même classe	généralisation maximale correcte
$x_1, y_1$	$x_5, x_8, \cancel{x_2}, x_{14}, \dots$	$g_1 = \text{MG}(\{x_1, x_5, x_8, x_{14}, \dots\})$ $g_1 \rightarrow y_1$
$x_2, y_2$	$\cancel{x_1}, x_3, \cancel{x_1}, x_{12}, \dots$	$g_2 = \text{MG}(\{x_2, x_3, x_{12}, \dots\})$ $g_2 \rightarrow y_2$

- Exemple : âge  $\in [25, 40] \rightarrow$  positif ;
- instable : dépend de la graine et de l'ordre des exemples ;
- pour un exemple donné, un moindre généralisé correct conclut sur une unique classe ( $-1$  ou  $+1$ ) ou s'abstient ( $0$ ).

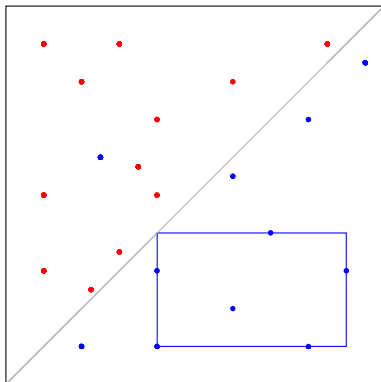
# Algorithme MGC

**Entrées :**  $E = [e_1, \dots, e_n] \subseteq \mathcal{X}$  un ensemble *ordonné* de  $n$  exemples de la même classe,  $N$  un ensemble de contre-exemples.

**Sortie :**  $h \in \mathcal{H}$  une généralisation de  $E$ , **maximalement** correcte par rapport à  $E$  et  $N$ .

- 1:  $h = e_1$
- 2: **for**  $i = 2$  to  $n$  **do**
- 3:      $h' = \text{MG}(h, e_i)$  {Généralisation entre deux hypothèses.}
- 4:     **if**  $(\forall e \in N : h' \not\subseteq e)$  **then**
- 5:          $h = h'$  { $h'$  (correcte) devient la généralisation courante.}
- 6:     **end if**
- 7: **end for**
- 8: **return**  $h$

# Déroutement de MGC





## L'algorithme DLG [Webb and Agar, 1992]

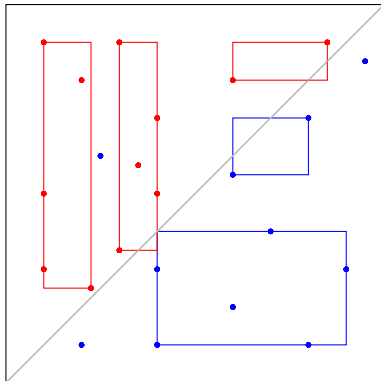
**Entrées** :  $A$  un ensemble de  $n$  exemples  $(x_i, y_i)$ .

**Sortie** :  $H$  un ensemble de règles.

- 1:  $H = \emptyset$ ;  $O = A$ ;  $j = 1$
- 2: **while** ( $O \neq \emptyset$ ) **do**
- 3:      $target$  = classe du premier exemple de  $O$
- 4:      $P = [x_i \in O | y_i = target]$
- 5:      $N = [x_i \in A | y_i \neq target]$
- 6:      $h_j = \text{MGC}(P, N)$
- 7:      $O = O - [x \in O : h_j \succeq x]$
- 8:     ajouter  $h_j$  à  $H$ ;  $j = j + 1$
- 9: **end while**
- 10: **return**  $H$

# Déroulement

Couverture gloutonne des exemples : on répète le calcul de moindre généralisé correct sur les exemples non couverts jusqu'à couverture complète des exemples.



# Bilan DLG

## Avantages

- rapide ;
- permet d'appréhender les attributs pertinents ;
- donne des indications sur la difficulté du problème d'apprentissage : nombre de règles apprises, couverture de chaque règle.

## Inconvénients

- glouton ;
- peu prédictif ;
- peu compréhensible.

# Intuitions

Combattre la dépendance à l'ordre des exemples et chercher la compréhensibilité.

## Principes de GloBo

- 1 calculer plusieurs moindre-généralisés en utilisant des exemples différents comme graine et des exemples de la même classe mélangés ;
- 2 retenir les règles qui permettent une couverture minimale des exemples.

Si chaque exemple sert à un moment de graine, alors au final tout exemple est couvert par au moins une hypothèse.

## Algorithme GloBo [Torre, 1999] (1)

**Entrées :**  $A$  un ensemble de  $n$  exemples  $(x_i, y_i)$ .

**Sortie :**  $H$  un ensemble de règles.

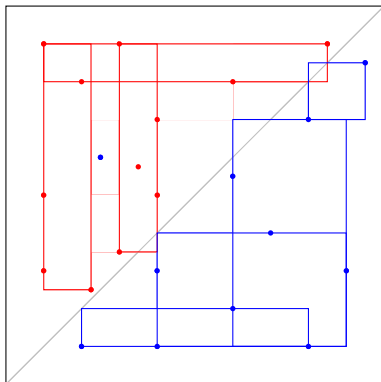
- 1:  $H' = \emptyset$
- 2: **for**  $i = 1$  to  $n$  **do**
- 3:      $P = [x_j | y_j = y_i \wedge i \neq j]$
- 4:      $N = [x_j | y_j \neq y_i]$
- 5:     **mélanger P aléatoirement**
- 6:      $h_i = \text{MGC}(x_i :: P, N)$
- 7:     ajouter  $h_i$  à  $H'$
- 8: **end for**

## Algorithme GloBo (2)

- 1:  $H = \emptyset$
- 2: **while**  $(\exists x_i, \forall h_j \in H, h_j \not\supseteq x_i)$  **do**
- 3:      $h = \text{ArgMax}_{h_i \in H'} |[x_j : h_i \supseteq x_j \wedge \nexists h_k \in H : h_k \supseteq x_j]|$
- 4:     ajouter  $h$  à  $H$
- 5: **end while**
- 6: **return**  $H$

Couverture minimale : problème NP-complet, heuristique quadratique. Justifiée ici ?

# Déroulement



# Bilan GloBo

## Avantages

- compréhensible ;
- meilleures prédictions que DLG.

## Inconvénients

- nombre quadratique de calculs de MG ;
- nombre cubique de tests de subsomption ;
- peut être battu en prédiction par des systèmes moins compréhensibles.



# Protocole

- Algorithmes en présence : C4.5, DLG, GloBo ;
- 20 problèmes du *repository UCI* [[Blake and Merz, 1998](#)];
- validations croisées 10 fois ;
- chaque apprentissage de GloBo est répété 10 fois ;
- nombre d'erreurs moyen ;
- visualisation des performances.

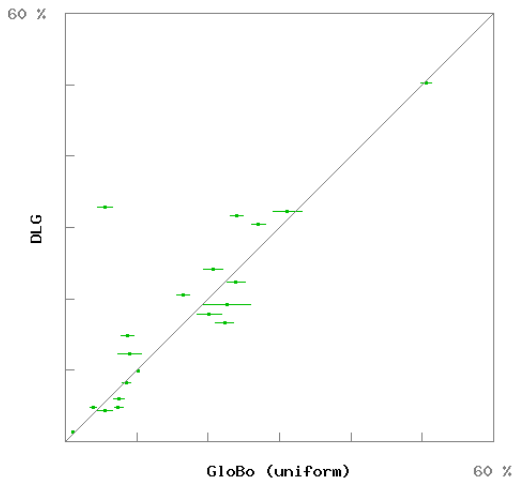
## Tableau de résultats I

Problème	C4.5	DLG	GloBo
audiology	<b>18.20</b>	24.16	20.76
breast-cancer	4.87	4.87	<b>3.89</b>
car	<b>7.69</b>	9.95	10.17
cmc	<b>48.07</b>	50.31	50.60
crx	<b>14.79</b>	20.57	16.50
dermatology	<b>6.23</b>	8.18	8.51
ecoli	<b>15.89</b>	22.35	23.88
glass	28.72	32.91	<b>5.57</b>
hepatitis	20.70	<b>17.88</b>	20.09
horse-colic	<b>13.63</b>	16.63	22.29
house-votes-84	<b>3.22</b>	4.83	7.39
ionosphere	<b>7.96</b>	14.84	8.74

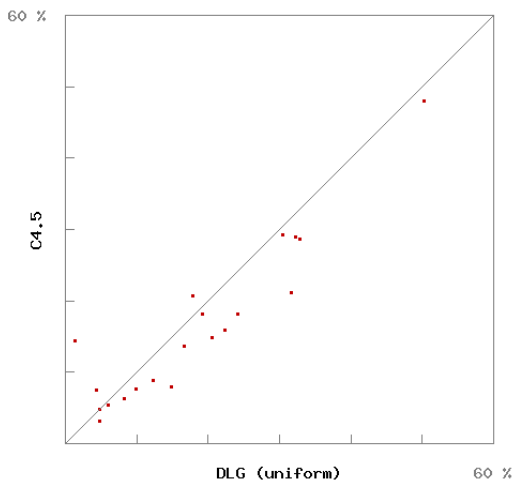
## Tableau de résultats II

iris	<b>5.33</b>	6.00	7.47
pima	29.29	30.47	<b>27.04</b>
promoters	<b>18.17</b>	19.17	22.60
sonar	<b>28.97</b>	32.18	31.09
tic-tac-toe	14.40	1.35	<b>1.11</b>
vowel	<b>21.21</b>	31.72	23.99
wine	<b>8.83</b>	12.33	8.94
zoo	7.51	<b>4.38</b>	5.55
<i>Moyennes</i>	<b>16.18</b>	<i>18.25</i>	<i>16.31</i>
	C4.5	DLG	GloBo

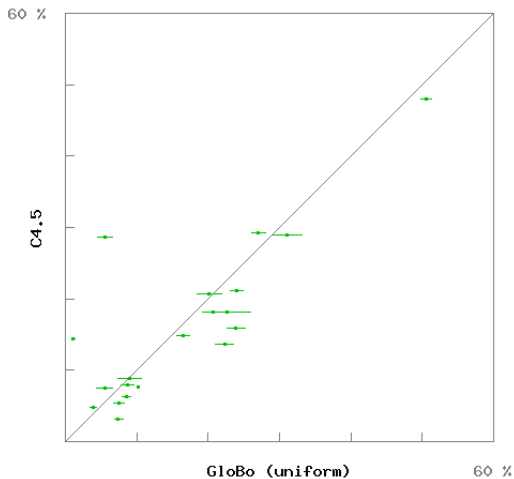
# GloBo vs DLG



# DLG vs C4.5



# GloBo vs C4.5



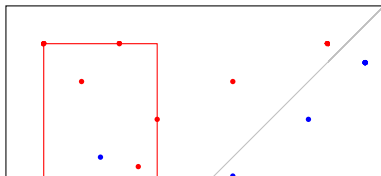
## Bruit et précision de Laplace [Clark and Niblett, 1989]

## Idée

Autoriser la couverture d'exemples d'autres classes : trouver un compromis entre le nombre total d'exemples couverts par une règle ( $t$ ) et le nombre d'exemples bien classés par cette règle ( $b$ ).

## Mesures

$$\begin{aligned} \text{Précision} &= \frac{b}{t} = \frac{1}{1} = 100\% = \frac{2}{2} = 100\% = \frac{7}{8} = 87.5\% = \\ \text{Précision de Laplace} &= \frac{b+1}{t+k} = \frac{1+1}{1+2} = 66.67\% = \frac{2+1}{2+2} = 75\% = \frac{7+1}{8+2} = \end{aligned}$$



## Données bruitées (2)

### Solution pour le bruit

- on réclamait un maintien absolu de la correction ;
- on veut maintenant que la précision de Laplace aille croissante.

$$\text{Précision de Laplace} = \frac{b + 1}{t + k}$$

Le critère de validation d'une généralisation dans MGC devient :

- 1: **if** (PrecisionLaplace( $h'$ )  $\geq$  PrecisionLaplace( $h$ )) **then**
- 2:      $h = h'$
- 3: **end if**



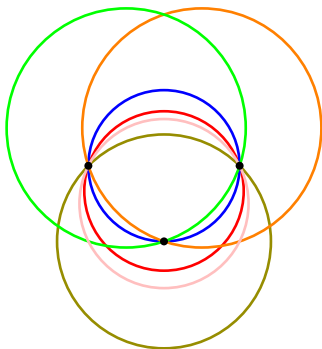
# Des hypothèses-cercles

- $\mathcal{E} = \mathbb{R}^2$ ;
- $d(A, B) = \sqrt{(x_A - x_B)^2 + (y_A - y_B)^2}$ ;
- $\mathcal{H} = \{(c, r) \mid (c \in \mathbb{R}^2) \text{ et } (r \in \mathbb{R})\}$ ;
- $h \succeq e \Leftrightarrow d(c_h, e) \leq r_h$ .

Unicité et calcul du moindré généralisé ?

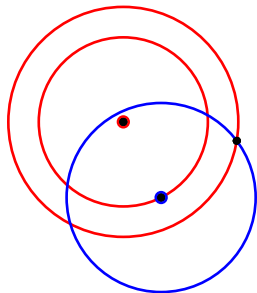
# Cercles multiples

Pour un ensemble de points de  $\mathbb{R}^2$ , il y a une infinité de cercles qui enveloppent ces points...



# Cercles : unicité algorithmique (idée 1)

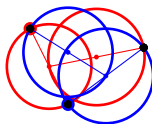
Idée : la graine sert de centre et on augmente le rayon...



Pourquoi est-ce une mauvaise idée ?

## Cercles : unicité algorithmique (idée 2)

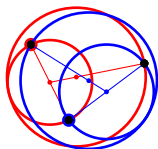
Idée : on prend le nouveau centre entre le précédent centre et le nouveau point.



Bonne ou mauvaise idée ?

## Cercles : unicité algorithmique (idée 3)

Idée : on calcule le rayon minimal pour capturer le cercle courant et le nouveau point, on en déduit le nouveau centre.



À tester !

Exercice : définir précisément cet algorithme.

# Pour poursuivre

- D'autres idées pour un calcul de cercle ?
- et si nous changions de norme ?

$$d(A, B) = |x_A - x_B| + |y_A - y_B|$$

- et si nous prenions des ellipses ?

# Résumé de l'architecture

Trois niveaux :

- le premier niveau fournit l'opération MG permettant de calculer l'hypothèse moindre généralisée d'un ensemble d'exemples quelconque, découle de  $\mathcal{H}$  et  $\succeq$  ;
- le deuxième prend en compte les classes des exemples pour produire des hypothèses correctes, ou quasi-correctes si du bruit de classe est présent (MGC) ;
- le dernier niveau permet l'apprentissage d'un classifieur complet, par combinaison de règles élémentaires apprises par le niveau précédent (DLG, GloBo, etc.).

Seul le premier dépend des langages de représentation  $\mathcal{E}$  et  $\mathcal{H}$ .

# Bilan




- Constat des difficultés des arbres de décision, choix d'un apprentissage ascendant, guidé par les exemples; cheminement défendu par [Fürnkranz, 2002];
- ascendant guidé : on part des exemples et on les généralise pour construire des hypothèses;
- plusieurs algorithmes à disposition, valables pour un nombre quelconque de classes, il suffit de définir le test de subsomption et le calcul de moindre généralisé;
- « fiche signalétique » enrichie :
  - préciser les choix de  $\mathcal{E}$  et de  $\mathcal{H}$ ;
  - évaluer la VCdim de  $\mathcal{H}$ ;
  - expliciter le test de subsomption  $\succeq$ ;
  - déterminer si  $(\mathcal{H}, \succeq)$  implique l'unicité du moindre généralisé;
  - proposer un algorithme  $MG(h \in \mathcal{H}, e \in \mathcal{X})$ .



## Un exercice de réflexion pour finir

- Décrire ce qu'il advient des méthodes d'apprentissage vues aujourd'hui :
  - ① si le MG colle de très près aux exemples ;
  - ② si le MG au contraire décolle très vite ;
- rapprocher votre constat de résultats théoriques vus précédemment.

# Bibliographie I

-  Blake, C. and Merz, C. (1998).  
UCI repository of machine learning databases  
[<http://archive.ics.uci.edu/ml/>].
-  Clark, P. and Niblett, T. (1989).  
The cn2 induction algorithm.  
*Machine Learning*, 3(4) :261–283.
-  Fürnkranz, J. (2002).  
A pathology of bottom-up hill-climbing in inductive rule  
learning.  
In *Proceedings of the 13th European Conference on Algorithmic  
Learning Theory (ALT-02)*, pages 263–277. Springer-Verlag.

## Bibliographie II



Torre, F. (1999).

GloBo : un algorithme stochastique pour l'apprentissage supervisé et non-supervisé.

In Sebag, M., editor, *Actes de la Première Conférence d'Apprentissage*, pages 161–168.



Webb, G. I. and Agar, J. W. M. (1992).

Inducing diagnostic rules for glomerular disease with the DLG machine learning algorithm.

*Artificial Intelligence in Medicine*, 4 :419–430.