ERIC
ENTREPÔTS, REPRÉSENTATION & INGÉNIERIE des CONNAISSANCES

21 Janvier 2013

UNIVERSITÉ LUMIÈRE LYON 2
UNIVERSITÉ DE LYON

Utiliser le front de Pareto pour évaluer les performances d'un algorithme non supervisé de construction d'attributs

Marian-Andrei Rizoiu

Laboratoire ERIC
Université Lumière Lyon 2
France

Les données : Représentation matricielle attribut-valeur ;
Les valeurs sont **booléennes**.

Les défis :

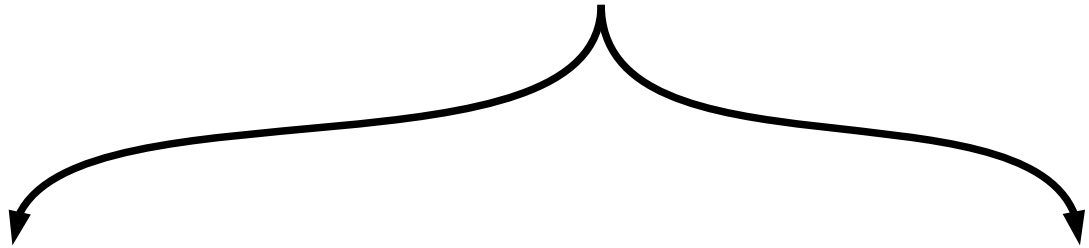
- Corrélation entre les attributs ;
- Mauvaise représentation des données.

	a_1	a_2	a_3	a_4	a_5	
φ_1		0	0			x_1^d
φ_2		1	0			x_2^d
φ_3		0	0			x_3^d
φ_4		0	0			x_4^d
φ_5		1	1			x_5^d
φ_6		1	1			x_6^d

L'objectif : Réduire les corrélations entre les attributs ;
Découvrir des informations manquantes sur les relations entre les attributs.

Étant donné l'ensemble des attributs :

{eau, cascade, manifestation, urbain, groupe, intérieur}



Relations structurelles :

cascade ⟷ *eau*

(*e.g.* type-de, part-du)

Relations sémantiques :

manifestation ⟷ *urbain*

(issues des données)

L'objectif : Réduire les corrélations entre les attributs ;
Découvrir des informations manquantes sur les relations entre les attributs.

{groupes, rue, bâtiment, intérieur}

groupes ∧ ¬*rue* ∧ *intérieur*

groupes ∧ *rue* ∧ bâtiment

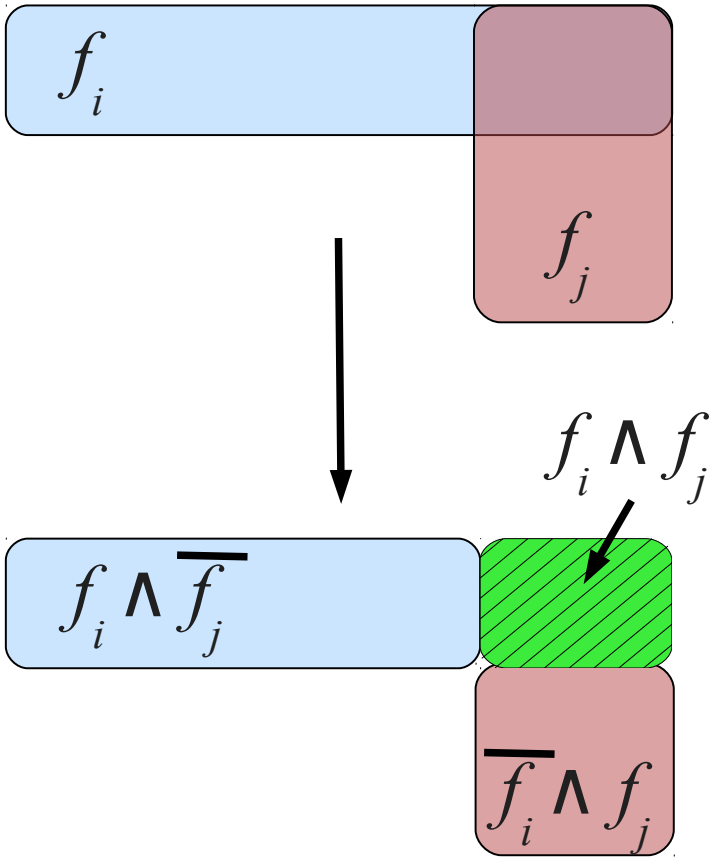
Le plan:

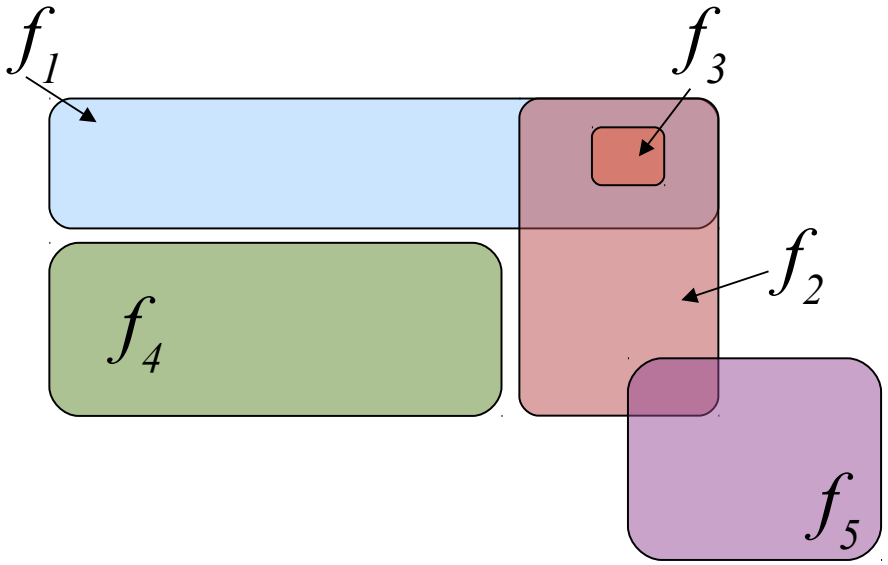
La solution proposée : Remplacer les paires d'attributs très corrélées par des conjonctions d'attributs et leur négations.

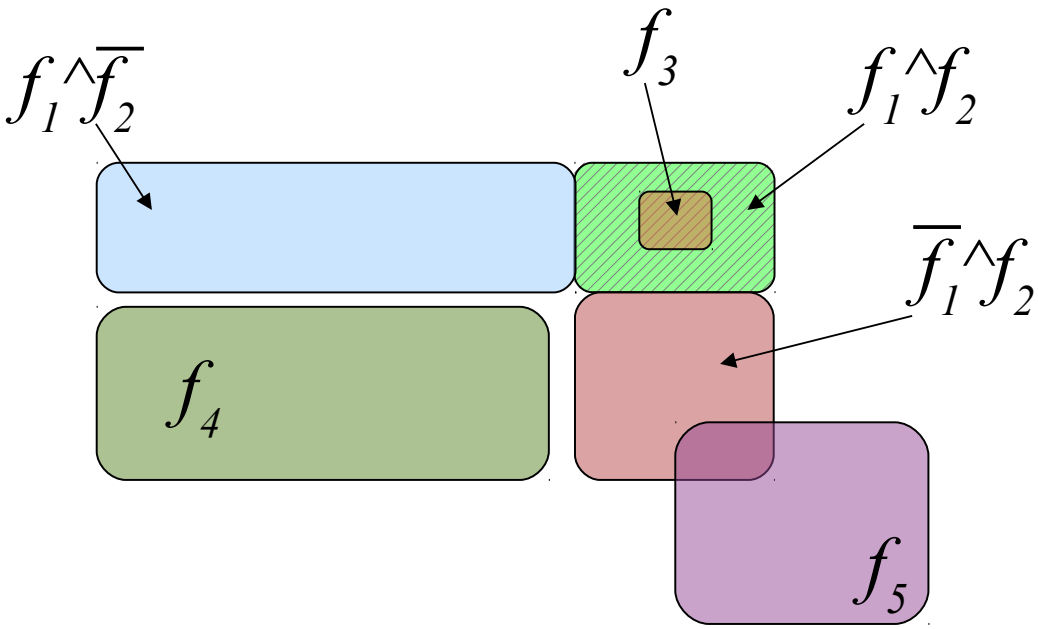
$$\{f_i, f_j\} \longrightarrow \{f_i \wedge f_j, f_i \wedge \overline{f_j}, \overline{f_i} \wedge f_j\}$$

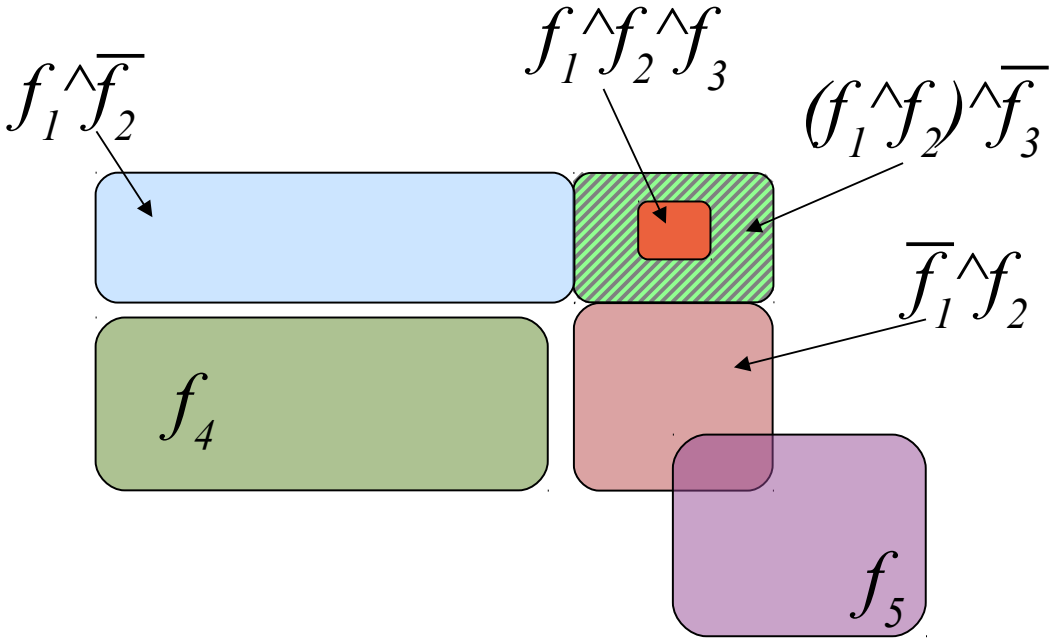
Algorithme *uFC* :

- Chercher des paires d'attributs corrélées (coefficient de Pearson) ;
- Construire de nouveaux attributs.
- Éliminer les anciens attributs et les attributs sans support.




f_1
f_3
f_2
f_4
f_5


$f_1 \wedge \overline{f_2}$
f_3
$f_1 \wedge f_2$
$\overline{f_1} \wedge f_2$
f_4
f_5


$f_1 \wedge \overline{f_2}$
$f_1 \wedge f_2 \wedge f_3$
$(f_1 \wedge f_2) \wedge \overline{f_3}$
$\overline{f_1} \wedge f_2$
f_4
f_5

Évaluation d'un ensemble d'attributs - deux mesures contraires :

Overlapping Index : basée sur la formule de Poincaré, mesure la corrélation totale de la collection des attributs

→ $OI(F) \in [0,1], meilleur\ vers\ 0$

Complexité : nombre des attributs par rapport au nombre maximal qui peut être construit ;

→ A chaque itération, le nombre des attributs augmente ;

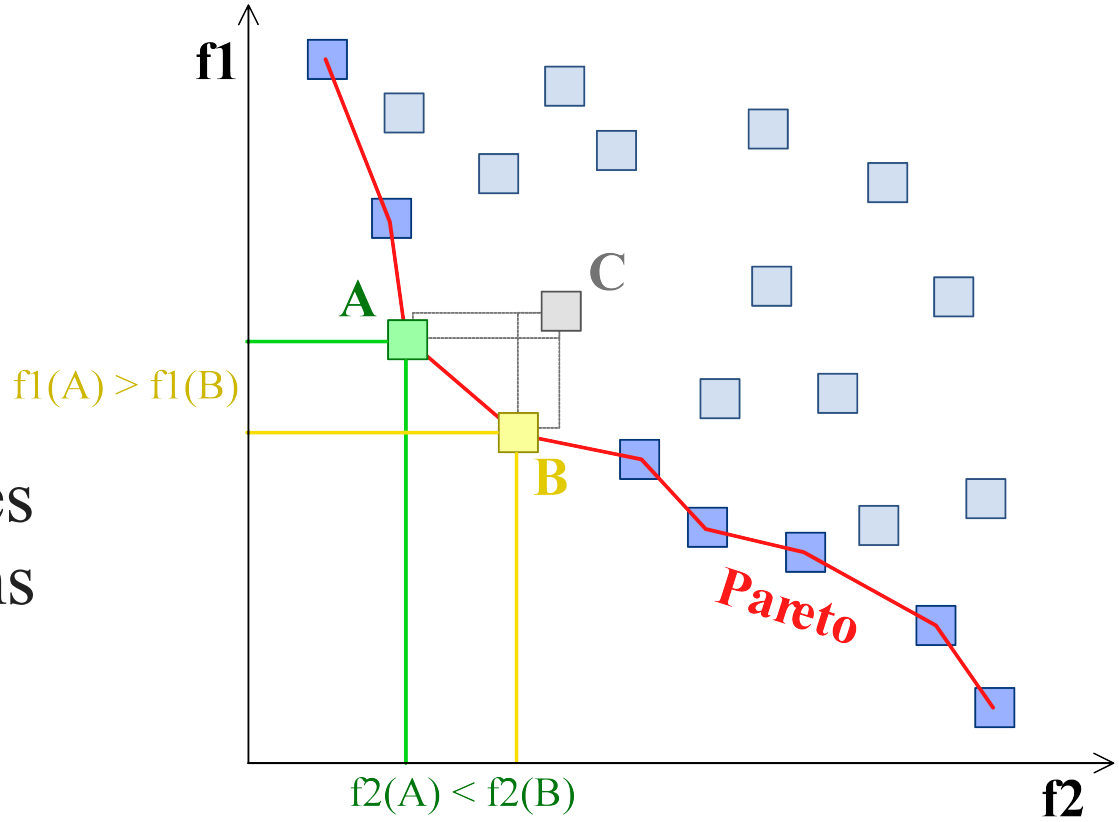
→ La longueur moyenne des attributs augmente ;

→ $C_0(F) \in [0,1], meilleur\ vers\ 0$

Le compromis entre les deux critères opposés.

- Notre problème se réduit à optimiser 2 critères opposés simultanément ;
- On utilise la notion d'optimalité Pareto ;
- Pas de solution unique, mais un ensemble de solutions – **Front de Pareto**.

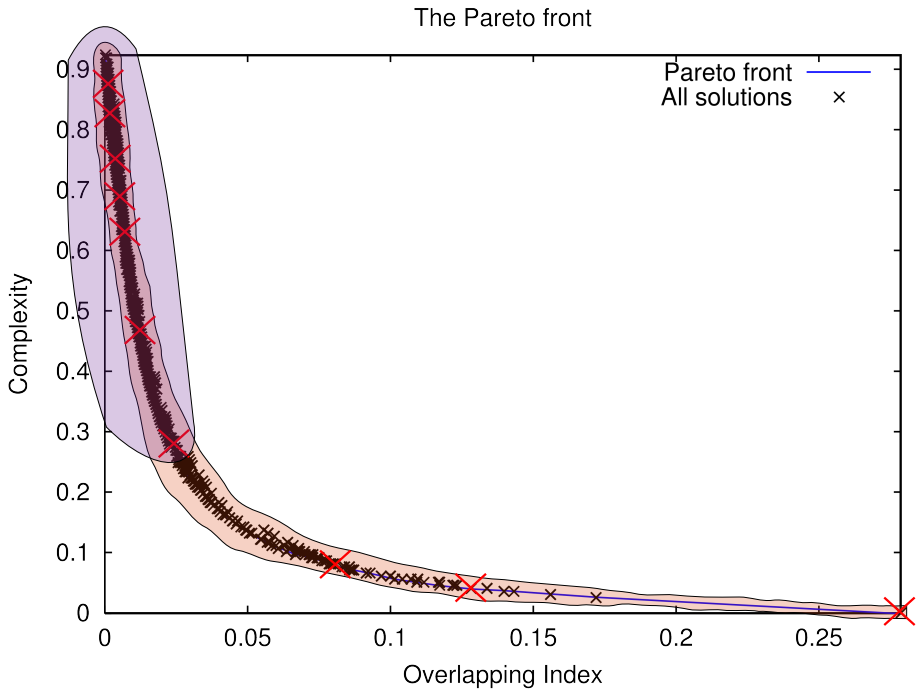
Idée:
Faire varier les 2 paramètres et projeter les solutions dans l'espace (OI, C_0)

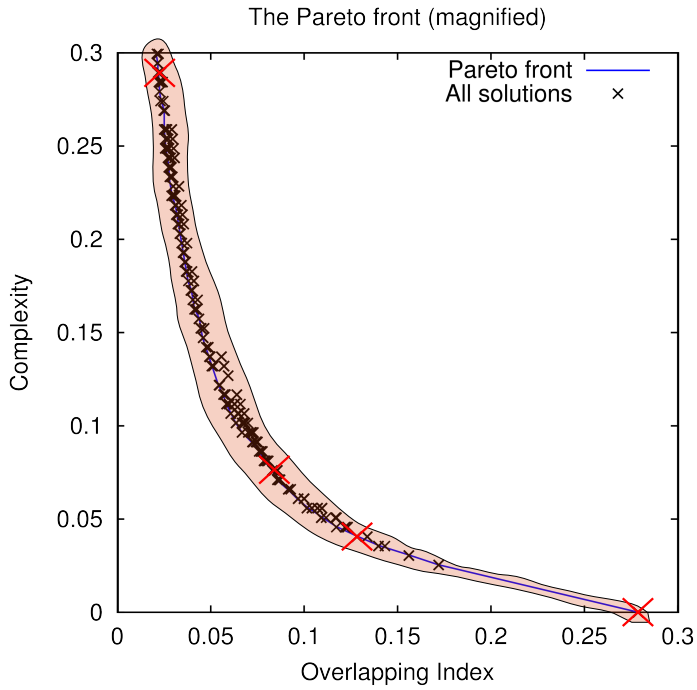


Le plan:

Premières conclusions :

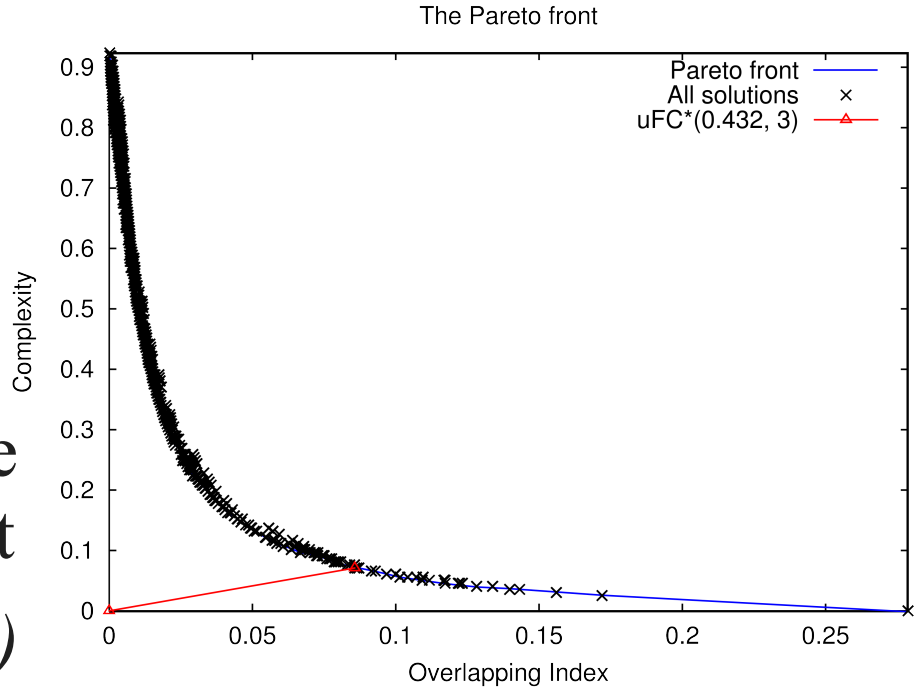
- Visualisation de l'espace des solutions ;
- Stabilité de l'algorithme ;
- Détecter empiriquement le **sur-apprentissage** ;
- Vitesse de convergence.

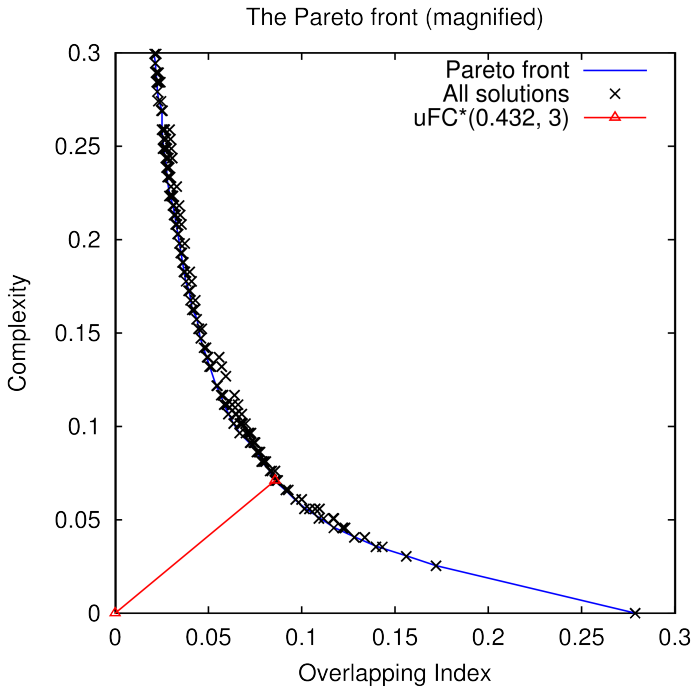




Heuristique pour choisir les paramètres :

l'heuristique "Closest-point" consiste à choisir sur le front Pareto le point où la perte de complexité (C_0) et le gain de score de co-occurrence *(OI)* sont relativement égaux.





groupes ∧ *rue* ∧ *intérieur*
groupes ∧ *rue* ∧ *intérieur*
groupes ∧ *rue* ∧ *intérieur*
eau ∧ *cascade* ∧ *arbre* ∧ *forêt*
eau ∧ *cascade* ∧ *arbre* ∧ *forêt*
eau ∧ *cascade* ∧ *arbre* ∧ *forêt*
ciel ∧ *bâtiment* ∧ *arbre* ∧ *forêt*
ciel ∧ *bâtiment* ∧ *arbre* ∧ *forêt*
ciel ∧ *bâtiment* ∧ *arbre* ∧ *forêt*
ciel ∧ *bâtiment* ∧ *panorama*
ciel ∧ *bâtiment* ∧ *panorama*

ciel ∧ *bâtiment* ∧ *panorama*
groupes ∧ *rue* ∧ *personne*
groupes ∧ *rue* ∧ *personne*
groupes ∧ *rue* ∧ *personne*
ciel ∧ *bâtiment* ∧ *groupes* ∧ *rue*
ciel ∧ *bâtiment* ∧ *groupes* ∧ *rue*
ciel ∧ *bâtiment* ∧ *groupes* ∧ *rue*
eau ∧ *cascade*
arbre ∧ *forêt*
gazon
statue

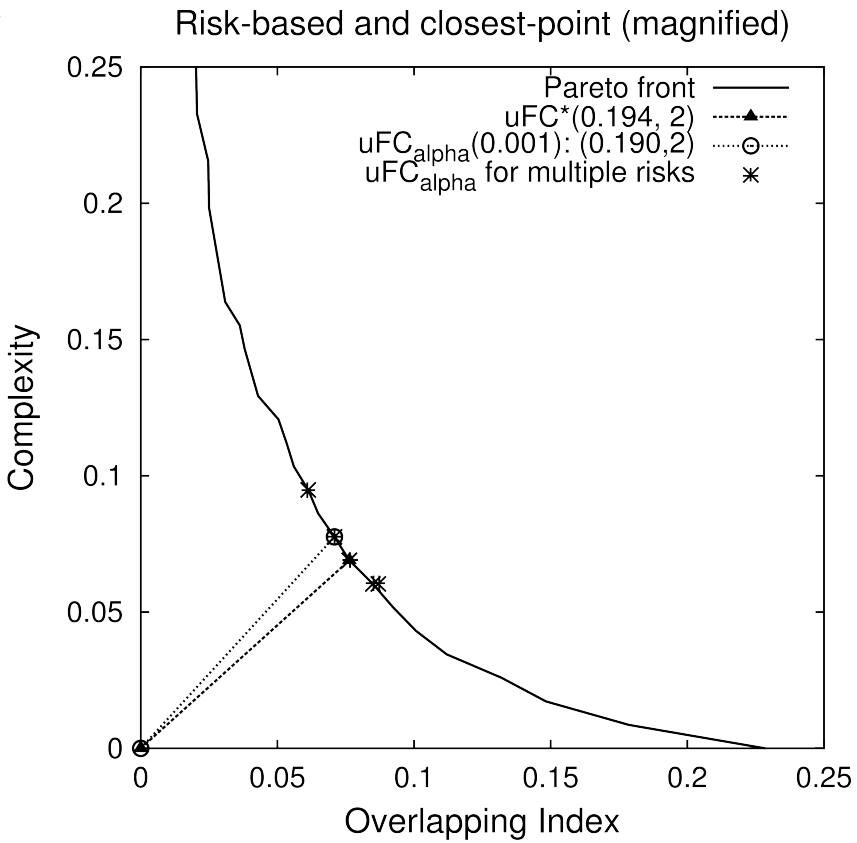
L'heuristique "risk-based" :

Un des paramètres (λ) est choisi en utilisant des tests statistiques

Le deuxième est transformé dans une condition d'arrêt des itérations.

Ça évite les exécutions multiples pour déterminer les valeurs des paramètres.

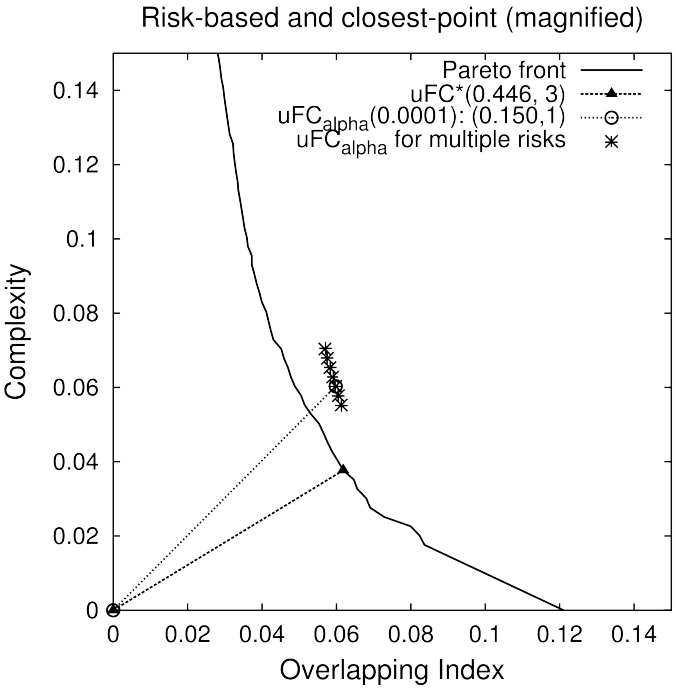
Nous utilisons le front de Pareto pour évaluer les solutions choisies par rapport à la solution "closest-point"

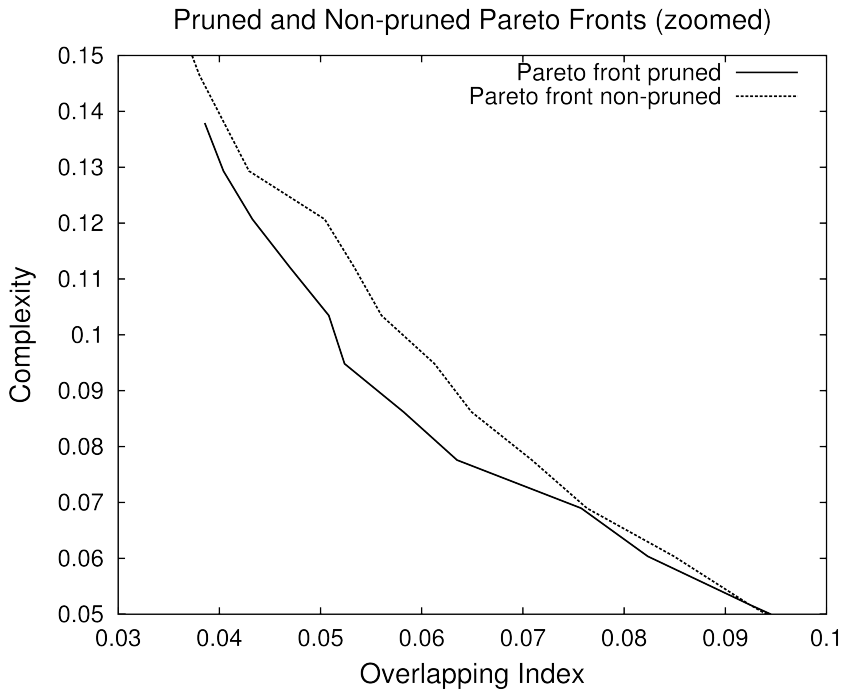


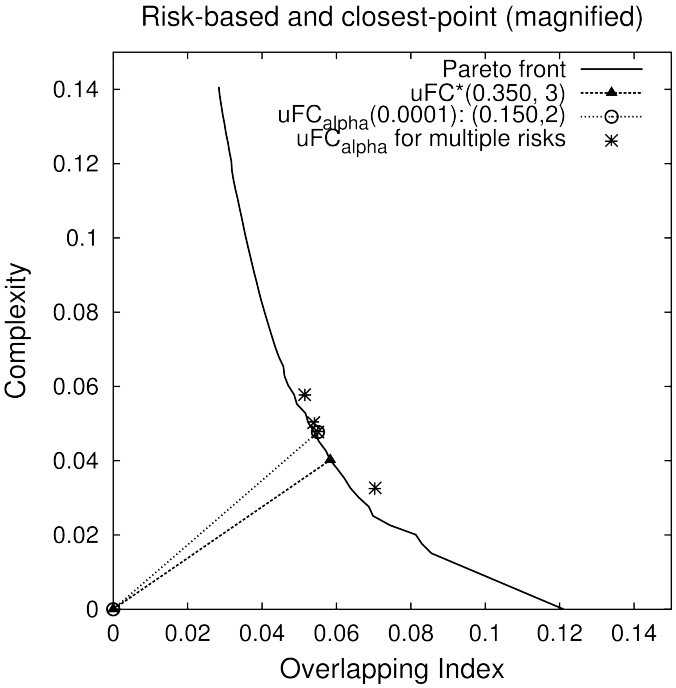
On utilise l'optimalité Pareto pour identifier des problèmes :

L'heuristique "risk-based" ne donne pas toujours de bons résultats

On propose un filtrage théorique







Le plan:

Conclusion :

Le front de Pareto nous a aidé à :

- visualiser l'espace de nos solutions;
- évaluer empiriquement la convergence et les performances;
- choisir les paramètres de notre algorithme;
- comparer les performances des deux algorithmes.

pourtant...

On n'a pas utilisé tout le pouvoir de l'optimalité et du front de Pareto.

Perspectives :

Comment mieux utiliser l'agrégation multi-critères dans notre contexte?

Est-ce que vous avez des idées?

Merci!

Bibliographie :

[a] Rizoiu, M.-A., Velcin, J., & Lallich, S. (2013). Unsupervised feature construction for improving data representation and semantics. *Journal of Intelligent Information Systems*.

[b] Sawaragi Y, Nakayama H, Tanino T (1985) *Theory of multiobjective optimization*, vol 176. Academic Press New York.