Boosting collaboratif de classifieurs non supervisés

une plateforme unificatrice pour le clustering multi-vues, le consensus de clusterings multiple et la recherche de clusterings alternatifs

Jacques-Henri Sublemontier

Laboratoire d'Informatique Fondamentale d'Orléans
Université d'Orléans - ENSI de Bourges
Fouille et Visualisation de Données Massives

26 juin 2013

Plan de la présentation

Cadre : multiplicité des données

Données multi-vues

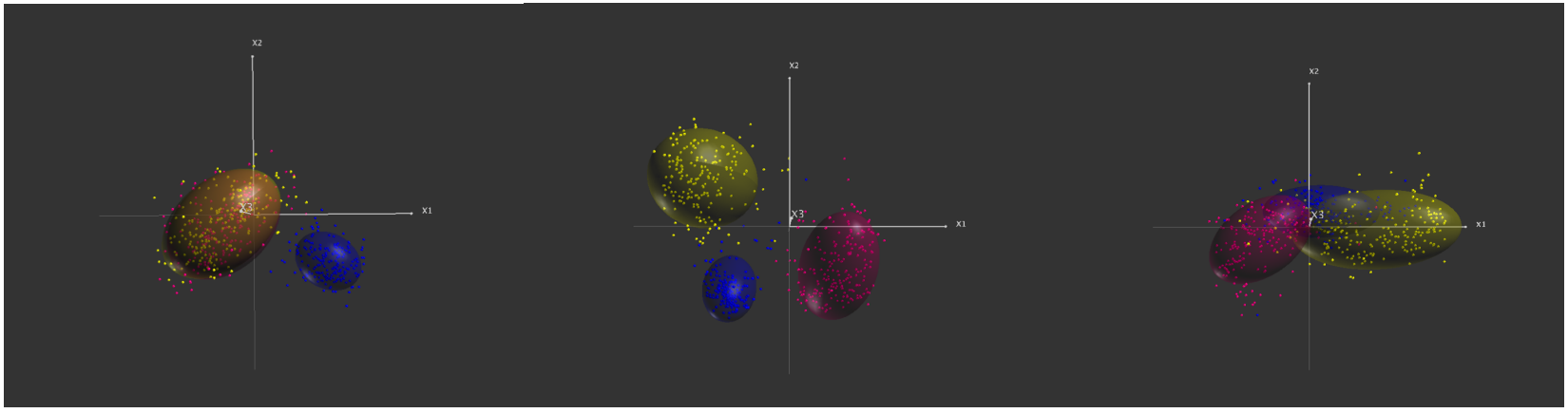
- plusieurs représentations pour chaque individu
- plusieurs mesures de proximité pour les données
- en général décentralisées

Application : caractères manuscrits

Pour une image :

- mesures de formes
- intensité des pixels
- plusieurs transformées

Cadre : multiplicité des données

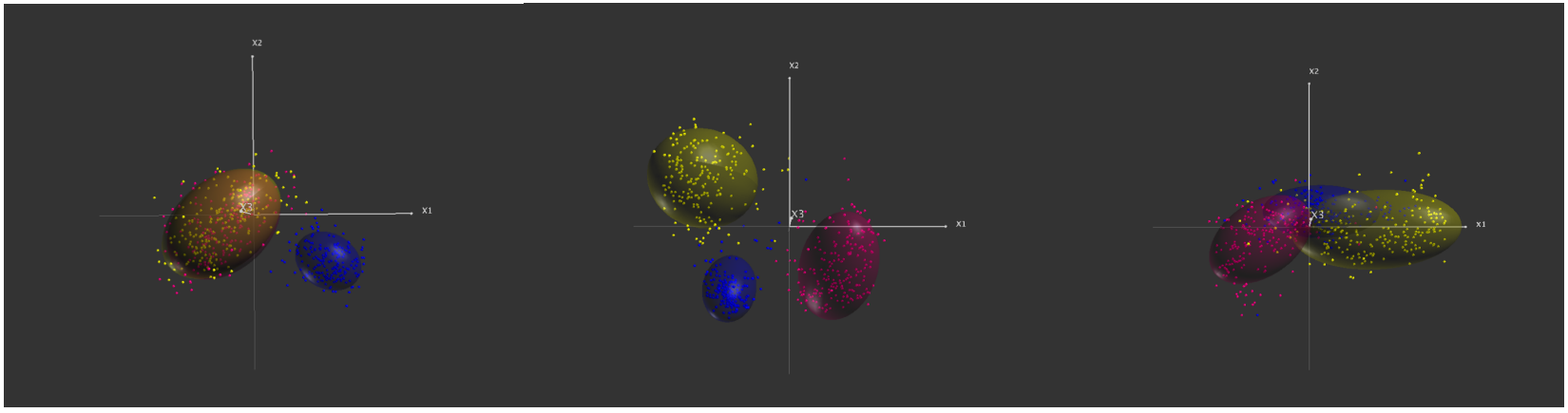


Application : caractères manuscrits

Pour une image :

- mesures de formes
- intensité des pixels
- plusieurs transformées

Cadre : multiplicité des données



Objectif : exploiter avantageusement toutes les informations

Problème : combiner les différentes vues pour produire une *bonne* partition

Hypothèse : une bonne partition réalise un consensus entre les différentes organisations locales naturelles

Cadre : multiplicité des analyses

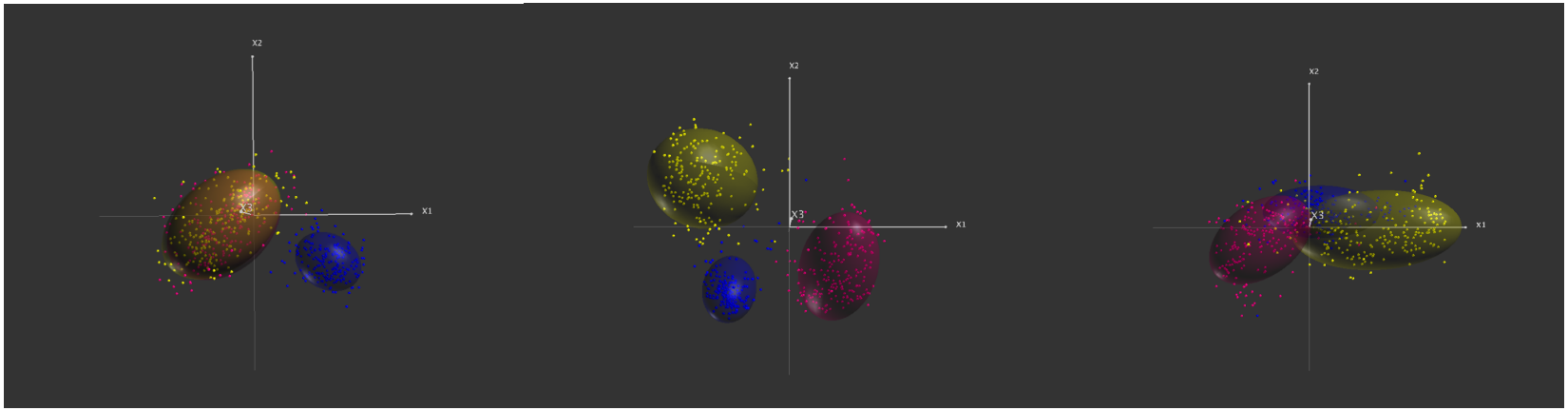
Aide à la décision pour l'utilisateur

- de multiples approches ont été développées
- pour un jeu de données, plusieurs groupements possibles
- *quid* d'une solution la meilleure ? choisir ? combiner ?
- plusieurs solutions intéressantes simultanément ?
- contexte de la grande dimensionnalité

Cadre : multiplicité des analyses

Aide à la décision pour l'utilisateur

- de multiples approches ont été développées
- pour un jeu de données, plusieurs groupements possibles
- *quid* d'une solution la meilleure ? choisir ? combiner ?
- plusieurs solutions intéressantes simultanément ?
- contexte de la grande dimensionnalité



Cadre : multiplicité des analyses

Aide à la décision pour l'utilisateur

- de multiples approches ont été développées
- pour un jeu de données, plusieurs groupements possibles
- *quid* d'une solution la meilleure ? choisir ? combiner ?
- plusieurs solutions intéressantes simultanément ?
- contexte de la grande dimensionnalité

Objectif : fournir de la diversité

Problème : chaque regroupement doit être de bonne qualité

Hypothèse : un regroupement de bonne qualité doit être proche de l'optimal d'un critère

Intérêt

Recherche de consensus

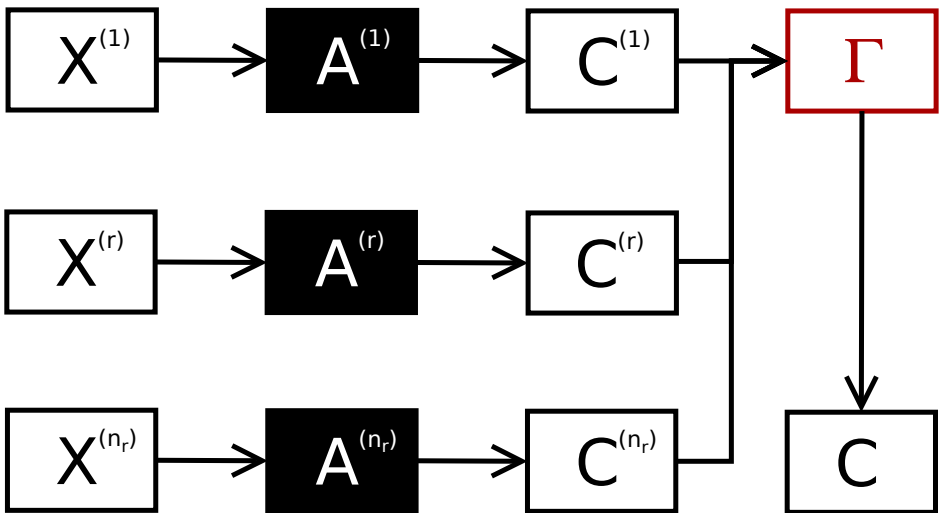
- réutilisation des connaissances
- combinaison de modèles, robustesse
- auto-paramétrage

Recherche d'alternatives

- explorations variées, diverses
- trouver des relations *orthogonales* entre les données
- aide le praticien, utilisateur de la méthode

Problématique

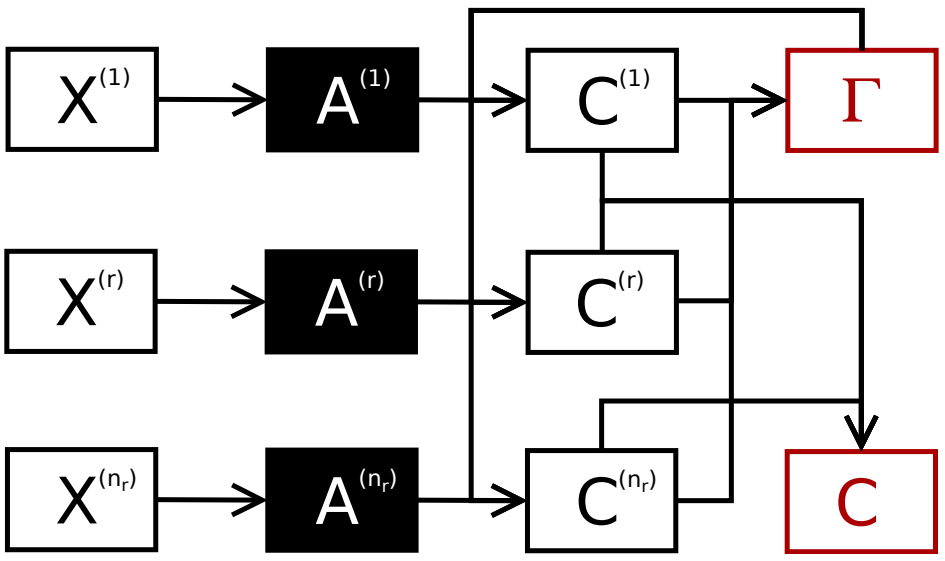
Clustering d'ensemble - consensus de partitions



- *Clustering ensemble* (CE) : MCLA, CSPA, HGPA
- *Fusion-Transfert* (FT)

Problématique

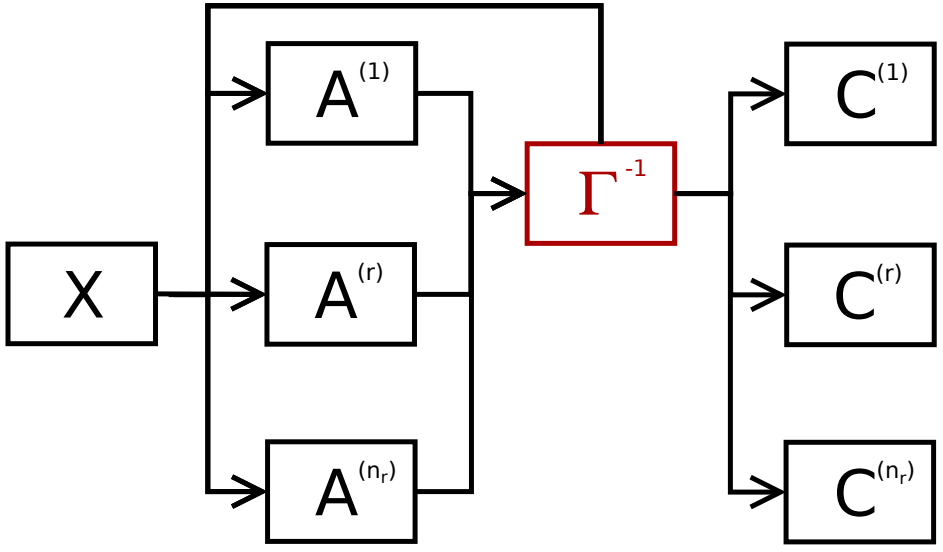
Clustering collaboratif



- SAMARAH
- MOCLE

Problématique

Clusterings alternatifs



- ADFT
- COALA, CAMI

Quelques approches

CE	[Strehl and Ghosh, 2003]
FT	[Guénoche, 2011]
SAMARAH	[Wemmert et al., 2000]
MOCLE	[Faceli et al., 2009]
COALA	[Bae and Bailey, 2006]
ADFT	[Davidson and Qi, 2008]
CAMI	[Dang and Bailey, 2010]

Bilan

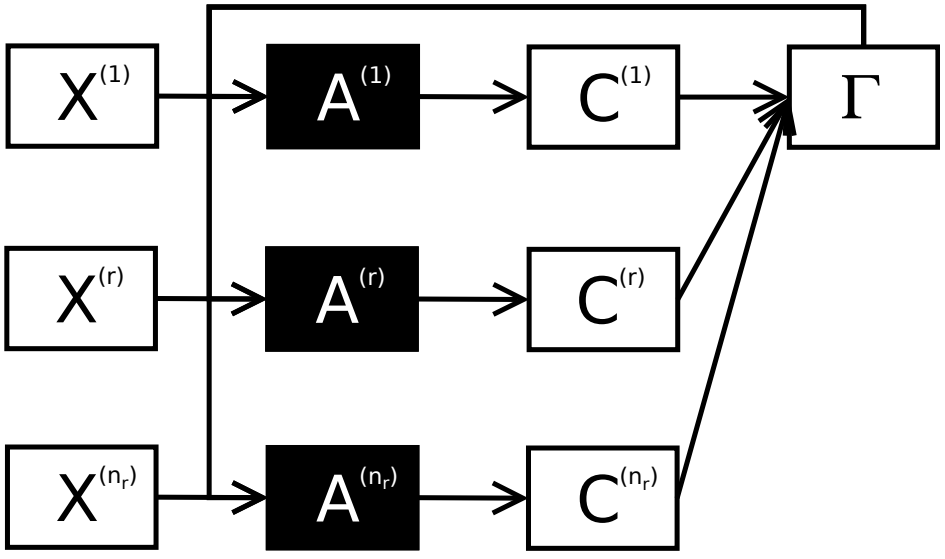
	Objectif		Algorithmes fixés	
	Consensus	Alternative	Non	Oui
CE	•		•	
FT	•		•	
SAMARAH	•		•	
MOCLE	•	•	•	
COALA		•		•
ADFT		•	•	
CAMI		•		•

- pas d'unification des deux problématiques
- algorithmes de *clustering* fixés ou pas de collaboration

Bilan

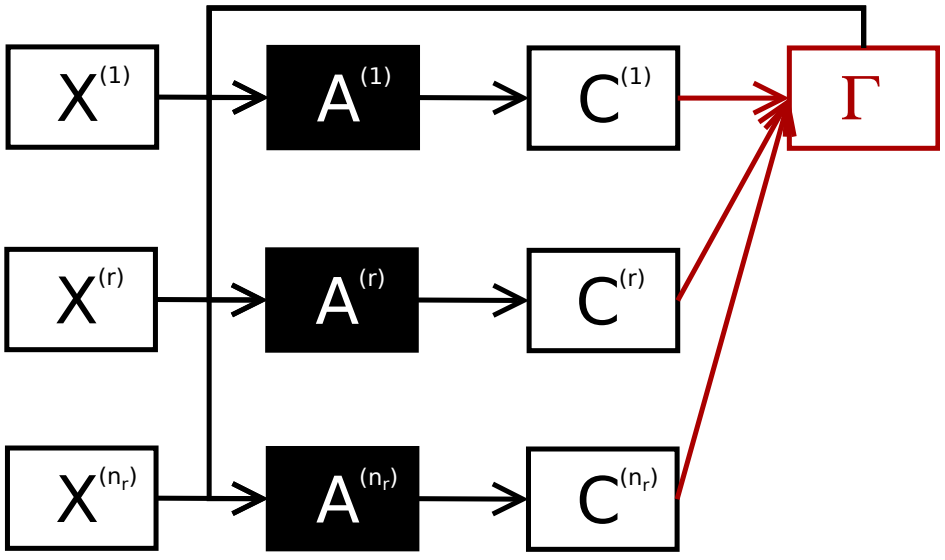
	Objectif		Algorithmes fixés	
	Consensus	Alternative	Non	Oui
CE	•		•	
FT	•		•	
SAMARAH	•		•	
MOCLE	•	•	•	
COALA		•		•
ADFT		•	•	
CAMI		•		•

- pas d'unification des deux problématiques
- algorithmes de *clustering* fixés ou pas de collaboration
- $\Longrightarrow$ proposer une plateforme unificatrice libre d'hypothèses



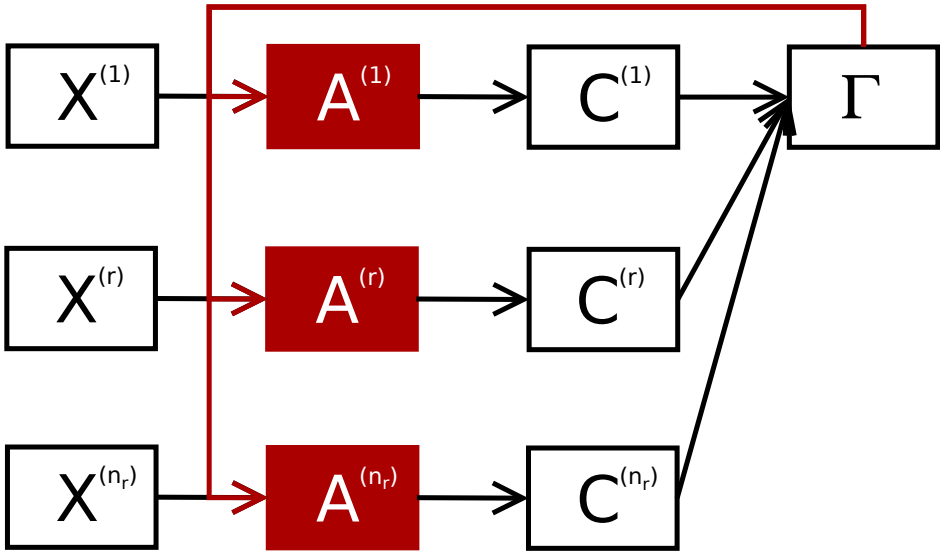
Objectifs

1. Pas d'hypothèses sur l'algorithme de *clustering* utilisé dans chaque vue



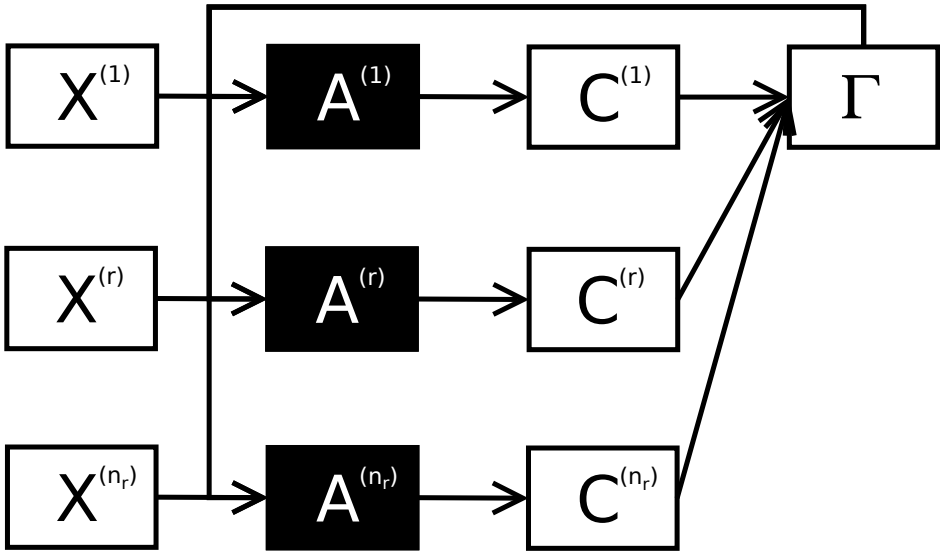
Objectifs

1. Pas d'hypothèses sur l'algorithme de *clustering* utilisé dans chaque vue
2. Échange d'informations pour la recherche d'un consensus



Objectifs

1. Pas d'hypothèses sur l'algorithme de *clustering* utilisé dans chaque vue
2. Échange d'informations pour la recherche d'un consensus
3. Intégration locale des informations pour atteindre le consensus



État de l'art

+ ne spécifient pas les algorithmes de *clustering*
– font collaborer les *résultats* d'algorithmes de *clustering*
– n'exploitent pas la représentation d'origine des individus

Constat

Les approches issues de l'état de l'art :

- font collaborer les *résultats* d'algorithmes de *clustering*
- n'exploitent pas la représentation d'origine des individus

Notre Objectif

- Unifier consensus et alternatives
- Formulation générique
- Données mono-vue et multi-vues

Idée

On s'inspire de CE et SAMARAH

Constat

Les approches issues de l'état de l'art :

- font collaborer les *résultats* d'algorithmes de *clustering*
- n'exploitent pas la représentation d'origine des individus

Notre Objectif

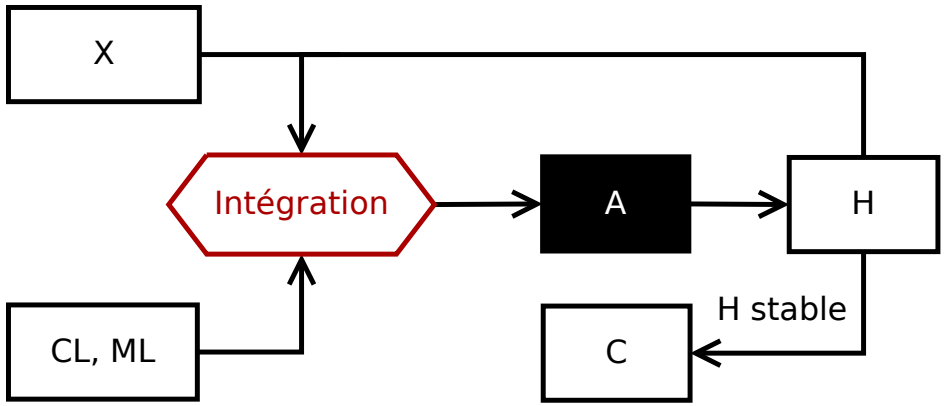
- Unifier consensus et alternatives ⇒ collaboration adaptée
- Formulation générique ⇒ utilisation de AdaUzaBoC
- Données mono-vue et multi-vues ⇒ redondance / complémentarité

Idée

On s'inspire de CE et SAMARAH

Parenthèse sur l'intégration de connaissances

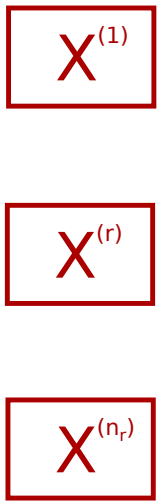
AdaUzaBoC : *boosting* de *clustering* adaptatif par *Uzawa*



Apprentissage itératif de représentations (ACP sous contraintes) :

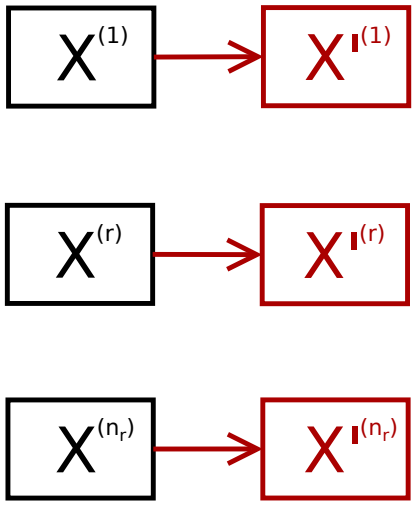
- adéquat à la représentation d'origine (ACP) ;
- conforme aux connaissances expertes (contraintes).

CoBoC : algorithme



Données multi-vues ou mono-vue

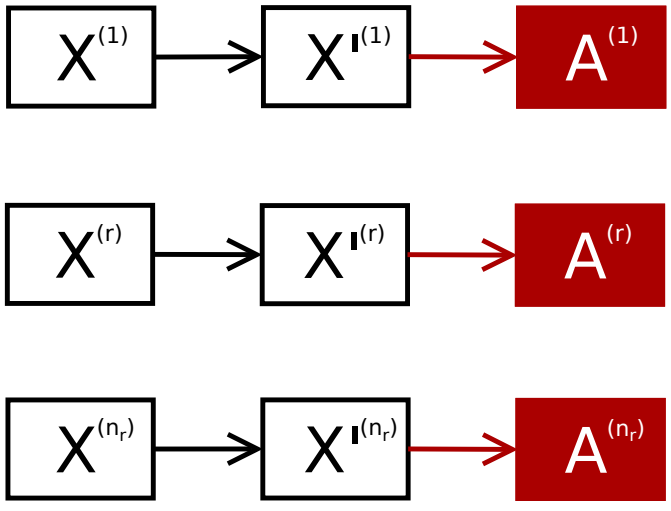
CoBoC : algorithme



Représentation initiale :

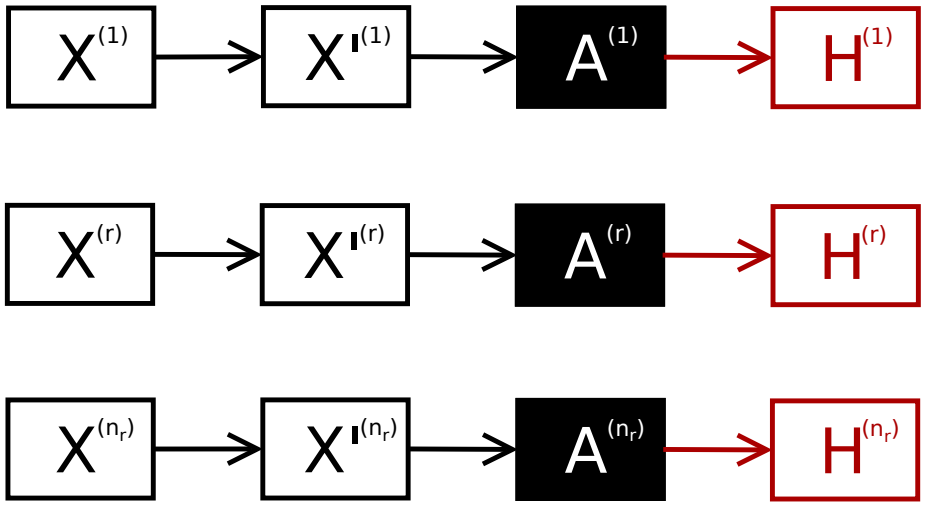
$X'^{(r)} = X^{(r)}$ pour toutes les vues

CoBoC : algorithme



Application d'algorithmes de *clustering* quelconques

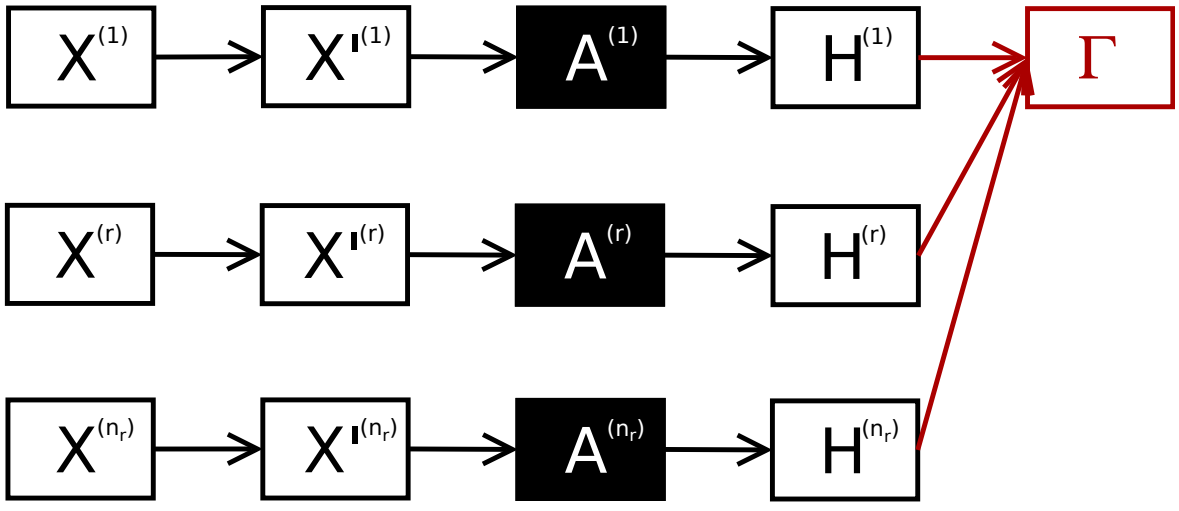
CoBoC : algorithme



Observation des hypothèses proposées par chaque algorithme :

$$H_{ij}^{(r)} = \begin{cases} 1 & A^{(r)} \text{ regroupe } x_i \text{ et } x_j \\ -1 & A^{(r)} \text{ sépare } x_i \text{ et } x_j \end{cases}$$

CoBoC : algorithme

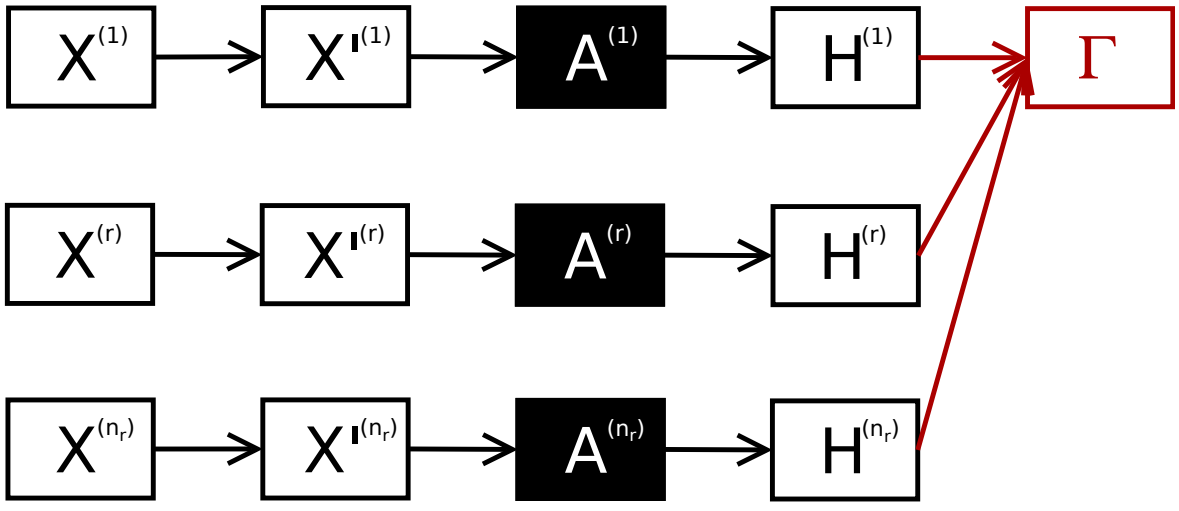


Combinaison des différentes hypothèses :

- confiance globale $\alpha_{ij} \in [-1, 1]$ associée à chaque (x_i, x_j)

$$\alpha_{ij} \approx \begin{cases} 1 & \Rightarrow x_i \text{ et } x_j \text{ sont } \textit{globalement} \text{ regroupés} \\ -1 & \Rightarrow x_i \text{ et } x_j \text{ sont } \textit{globalement} \text{ séparés} \end{cases}$$

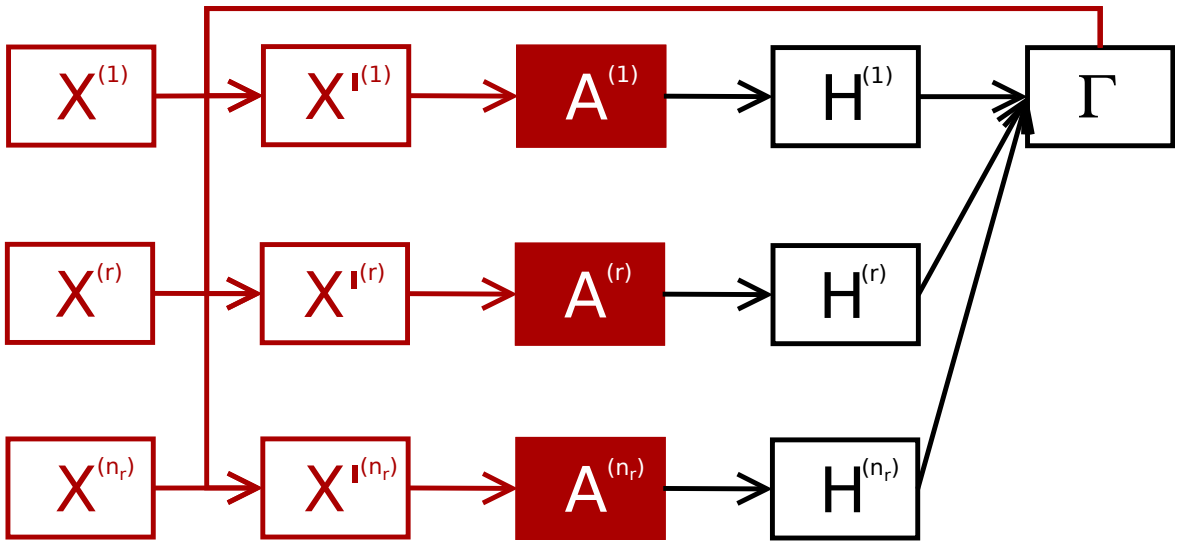
CoBoC : algorithme



Combinaison des différentes hypothèses :

- classement par la confiance α
- sélection de paires d'individus qui deviennent $\mathcal{ML}$ et $\mathcal{CL}$
- deux heuristiques / trois stratégies proposées

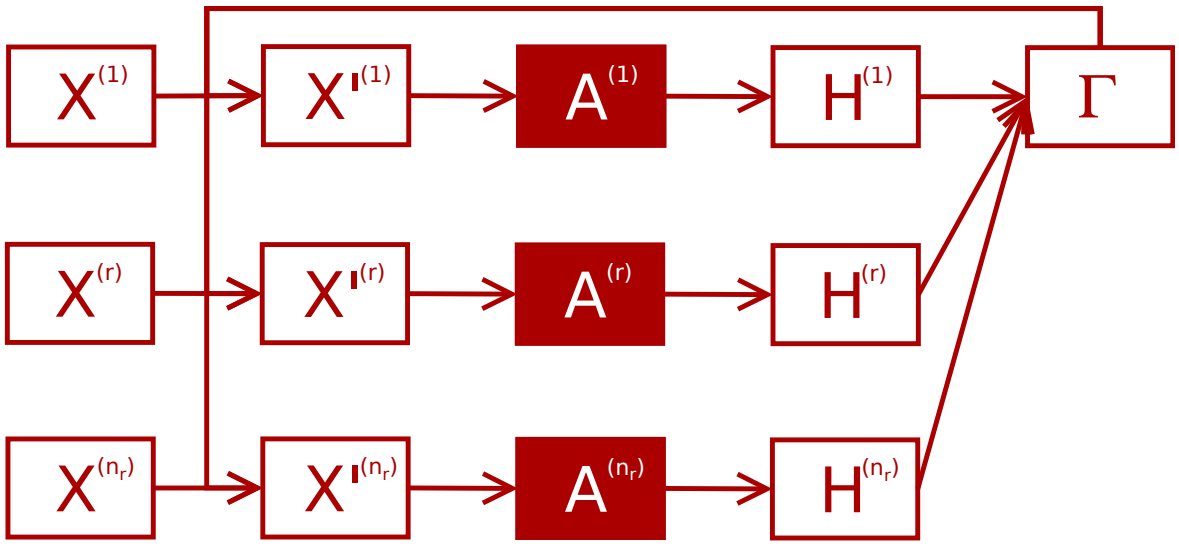
CoBoC : algorithme



Apprentissage de nouvelles représentations :

- par AdaUzaBoC à partir des $\mathcal{ML}$ et $\mathcal{CL}$
- les connaissances doivent permettre de mener à un consensus

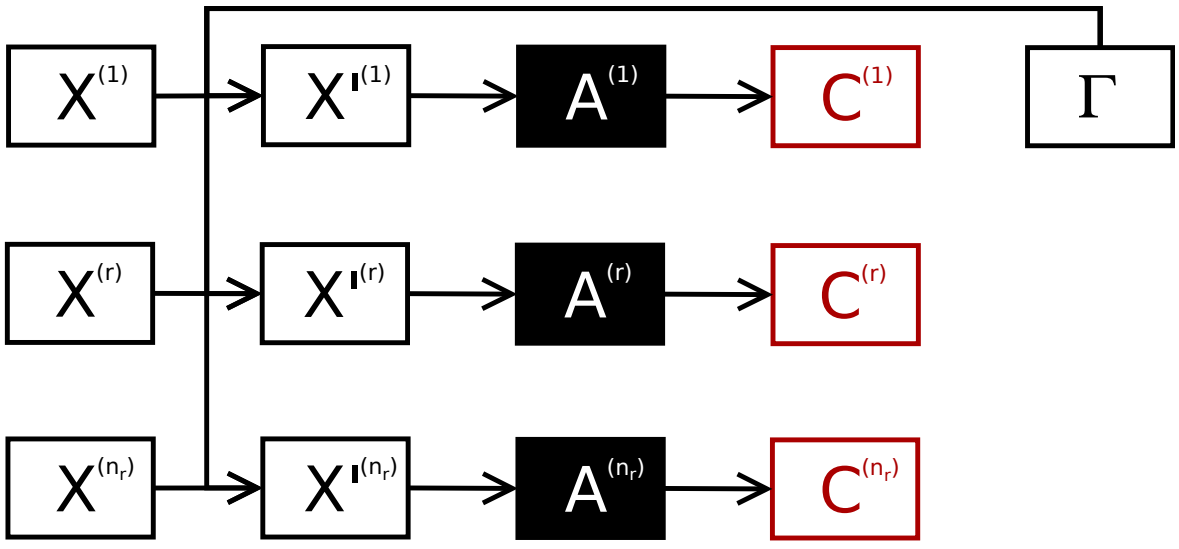
CoBoC : algorithme



Répéter le processus

- jusqu'à un nombre de collaborations donné
- l'ensemble des connaissances est enrichi à chaque étape

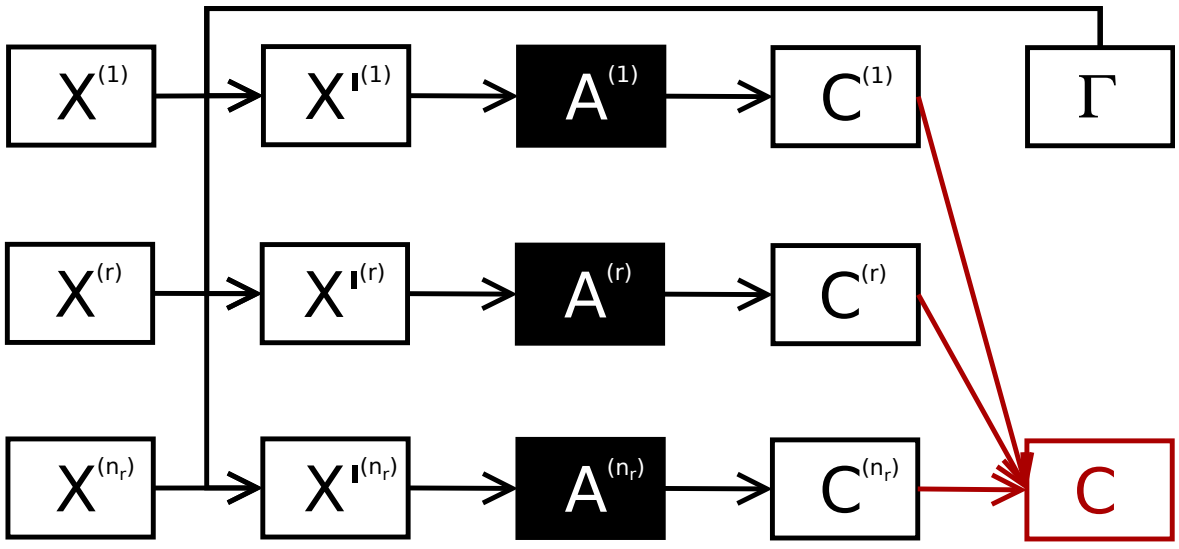
CoBoC : algorithme



Convergence vers un ensemble de *clusterings* :

- répond au problème du consensus de *clusterings* multiple

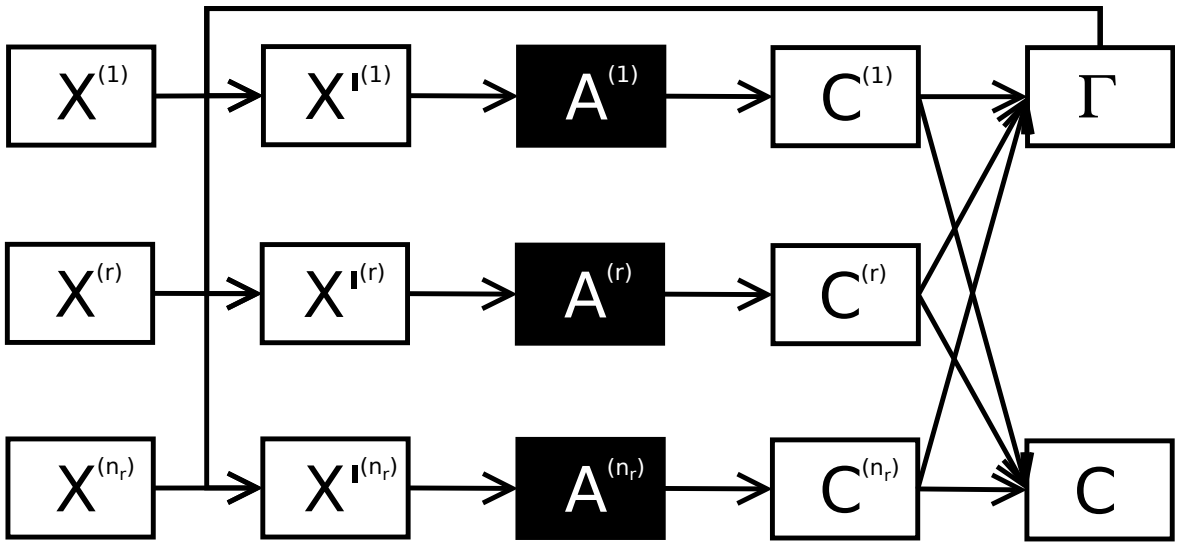
CoBoC : algorithme



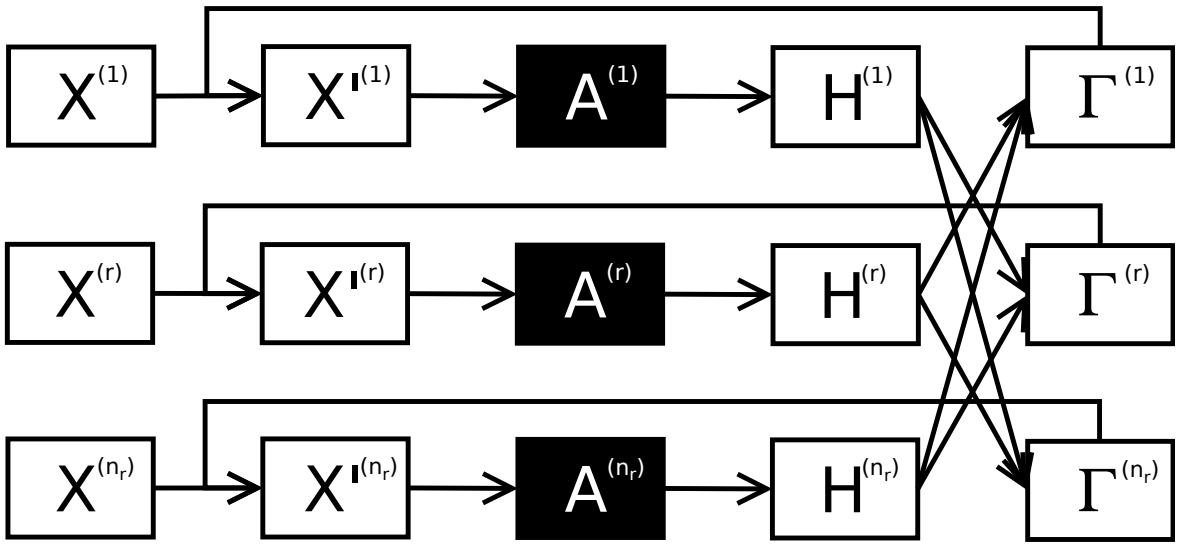
Vote final pour obtenir un unique *clustering* consensus :

- deux fusions ont été proposées
- calcul d'une matrice de similarité (hypothèses, confiances)
- appel un algorithme de *clustering* adapté

CoBoC : algorithme



AlterBoC : algorithme



Données multi-vues ou mono-vue

Des connaissances opposées sont générées

CoBoC

Données UCI

Jeu	individus	attributs/vues	classes	*clusterers*
Parkinson	195	22/1	2	2
Mfeat	2000	>500/6	10	6

Tests sur CoBoC pour la recherche de consensus

consensus : un même ensemble de contraintes est construit pour tous les algorithmes

complémentaire : un ensemble de contraintes par algorithme, pour guider chacun vers une même solution consensus

AlterBoC

Données UCI

Jeu	individus	attributs/vues	classes	*clusterers*
Parkinson	195	22/1	2	2
Mfeat	2000	>500/6	10	6

Tests sur AlterBoC pour la recherche d'alternatives

global : un ensemble de contraintes par algorithme pour guider celui-ci vers une solution différente de celle des autres

complémentaire : un ensemble de contraintes par algorithme pour guider celui-ci vers une solution différente de celle des autres

Protocole expérimental

CoBoC / AlterBoC

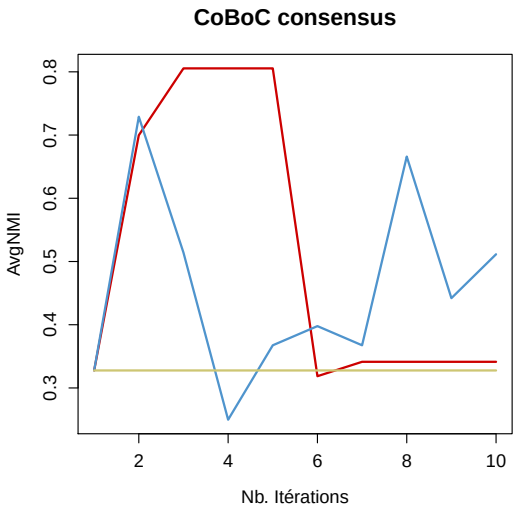
- normalisation des données multi-vues
- nombre d'échanges d'hypothèses T fixé à 10
- m connaissances (équilibrées) supplémentaires à chaque étape
- $m = 1\%$ du nombre de paires d'individus

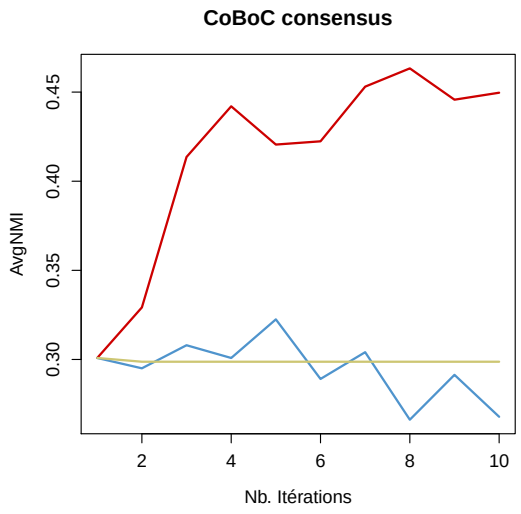
AdaUzaBoC

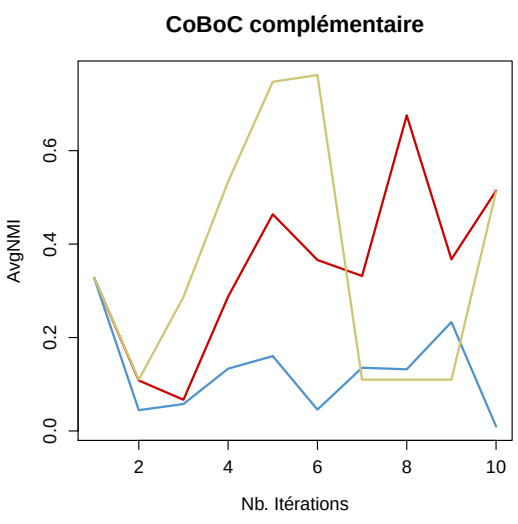
- dimension du sous-espace X' : heuristique
- nombre d'itérations maximal t fixé à 50

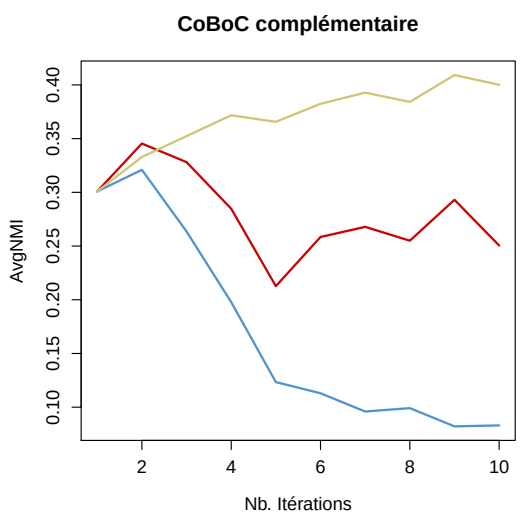
Approche CoBoC paramétrée par t, T, A_i et A_i $\forall$ $1 \leq i \leq n_c$

CoBoC : évaluation interne

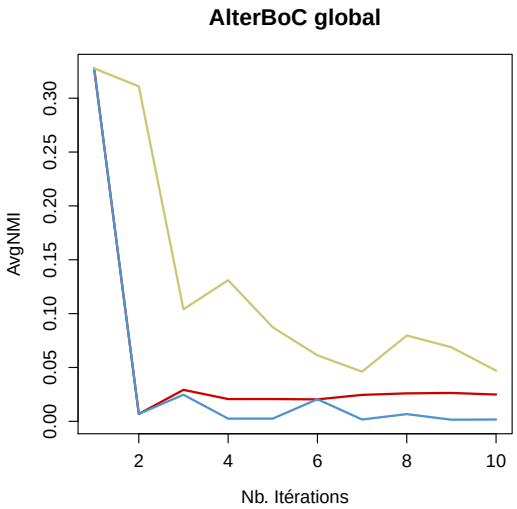


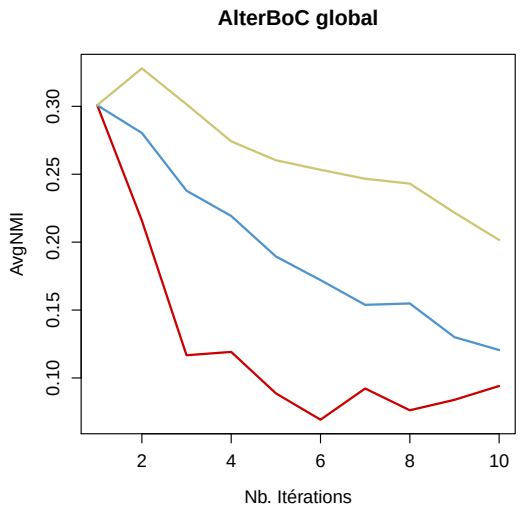


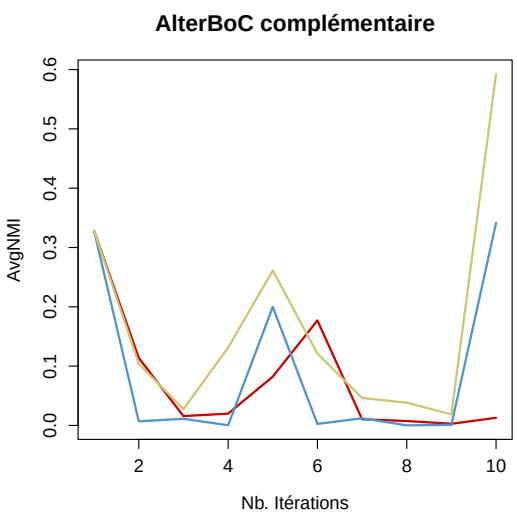


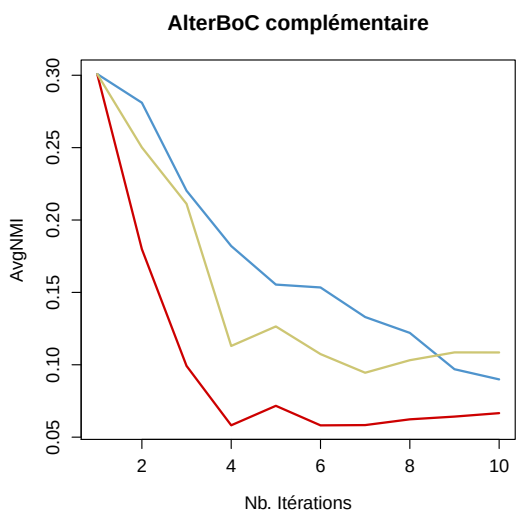


AlterBoC : évaluation interne









Bilan

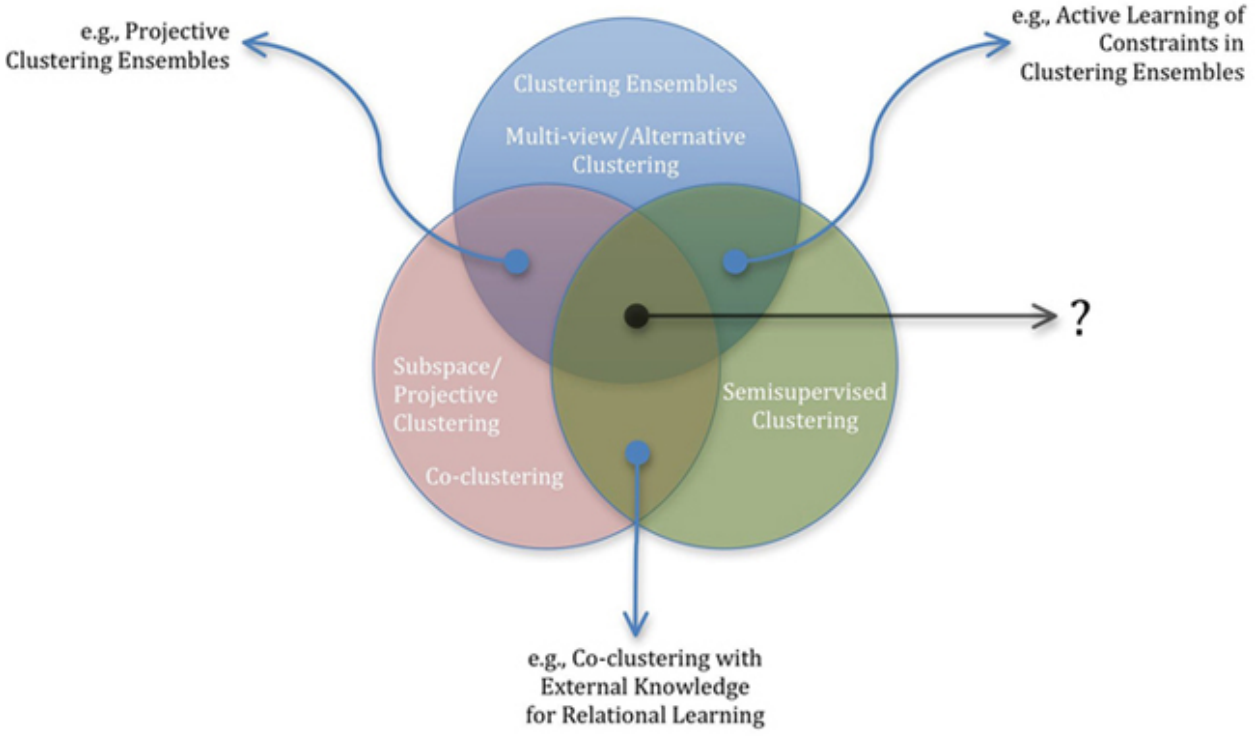
+ Approche pour le consensus CoBoC
+ Approche pour les alternatives AlterBoC
+ Plusieurs heuristiques (2) et plusieurs stratégies (3)
+ Mise en évidence de la nécessité de stratégies non triviales
– Difficulté du paramétrage
– Complexité algorithmique
– Gestion des *outliers*

Perspectives

Pour les approches CoBoC et AlterBoC :

- Réfléchir à des stratégies plus complexes et moins évidentes
- Intégrer des approches d'analyse de tableaux multiples (*ParaFac/CanDecomp*, *Tucker3*, *etc.*)
- Passage à l'échelle sur environnement *Big Data*

Perspectives



Workshop 3Clust, *PAKDD* 2012, Kuala Lumpur

Merci !

Questions ?


Bae, E. and Bailey, J. (2006).
Coala: A novel approach for the extraction of an alternate clustering of high quality and high dissimilarity.
In *ICDM*, pages 53–62. IEEE Computer Society.


Dang, X. H. and Bailey, J. (2010).
Generation of alternative clusterings using the cami approach.
In *SDM*, pages 118–129. SIAM.


Davidson, I. and Qi, Z. (2008).
Finding alternative clusterings using constraints.
In *ICDM*, pages 773–778. IEEE Computer Society.


Faceli, K., de Souto, M. C. P., de Araujo, D. S. A., and de Carvalho, A. C. P. L. F. (2009).
Multi-objective clustering ensemble for gene expression data analysis.
Neurocomputing, 72(13-15):2763–2774.


Guénoche, A. (2011).
Consensus of partitions : a constructive approach.
Adv. Data Analysis and Classification, 5(3):215–229.


Strehl, A. and Ghosh, J. (2003).
Cluster ensembles — a knowledge reuse framework for combining multiple partitions.
J. Mach. Learn. Res., 3:583–617.


Wemmert, C., Gançarski, P., and Korczak, J. J. (2000).
A collaborative approach to combine multiple learning methods.
International Journal on Artificial Intelligence Tools, 9(1):59–78.