Clustering multi-vues : une approche centralisée

Jacques-Henri Sublemontier, Guillaume Cleuziou,
Matthieu Exbrayat, Lionel Martin

Laboratoire d'Informatique Fondamentale d'Orléans
Université d'Orléans
B.P. 6759 - 45067 ORLEANS Cedex 2
{prenom.nom}@univ-orleans.fr

Résumé. Nous abordons dans ce papier le problème de la classification non-supervisée multi-vues, *i.e.* où les données peuvent être décrites par plusieurs ensembles de variables ou par plusieurs matrices de proximités. De nombreux domaines d'applications sont concernés, tels la Recherche d'Information, la Biologie, la Chimie et le Marketing. L'objet de cet axe de recherche est de proposer un cadre théorique et méthodologique permettant la découverte d'une classification réalisant un consensus entre les organisations émanant de toutes les vues. Il convient alors de combiner les informations de chacune des vues par l'intermédiaire d'un processus de fusion consistant à identifier l'accord entre les vues et à réduire le conflit. Plusieurs stratégies de fusion peuvent être appliquées, en amont, en aval, ou pendant le processus de classification. Nous présentons les différentes solutions de fusion envisageables suivant différents contextes applicatifs, puis nous nous focalisons sur des techniques dites centralisées. Nous proposons une approche de classification non supervisée floue qui généralise différentes solutions de fusion et nous présentons une extension à noyaux de cette approche, permettant le traitement de données hétérogènes. Nous montrons l'apport théorique et expérimental de cette approche sur des jeux de données *benchmarks* synthétiques et réels.

1 Introduction

La complexité toujours croissante des données constitue un défi pour la communauté fouille de données. Cette complexité peut concerner plusieurs aspects tels que la taille du jeu de donnée, la complexité des attributs, la temporalité ou plus généralement la multiplicité des données. Parmi les avancées récentes réalisées dans la communauté fouille de données, nous nous intéresserons dans ce travail à la fouille de données multi-vues. Pour de telles données, chaque individu peut-être décrit simultanément par plusieurs vues, et chaque vue apporte un éclairage différent sur l'organisation des individus.

1.1 Des données réelles multi-vues

Les données ayant de multiples représentations (vues) sont assez communes dans les domaines scientifiques, économiques et sociaux tels que la biologie, la chimie, la médecine, le marketing ou l'étude des réseaux sociaux. Par exemple, les biologistes considèrent simultanément l'activité de gènes (leur mesure d'expression), leur profil phylogénétique et leur localisation, dans le but de detecter des intéractions ou des corégulations de gènes (Yamanishi et al., 2004). En ce qui concerne la chimie, des molécules peuvent être décrites à la fois par des empreintes *fingerprints* et leur structure spatiale (Labute, 1998). Des données médicales peuvent inclure (Martin et al., 2006) des images rayon-X, des rapports médicaux, des analyses biologiques.
Dans le contexte du marketing, des informations sur un même ensemble de clients sont disponibles à partir de différentes bases de données (banque, magasin, administration, *etc.*). La construction et la fouille de réseaux sociaux suppose d'assembler plusieurs données (*e.g* e-mails, collaborations, appartenances à des organisations, *etc*.).
D'un point de vue transversal, les vues multiples se retrouvent également fréquemment pour l'analyse de données textuelles en linguistique computationelle et en recherche d'information. Les applications de recherche d'information traitent, par exemple, de pages web décrites selon plusieurs angles (texte, liens, structure, *etc*.) et plusieurs niveaux de descriptions du texte peuvent également être considérés pour l'analyse en linguistique computationelle (lexique, morphologie, syntaxe, *etc*.) (Cleuziou et Poudat, 2007).

1.2 Contextes et motivations

La présente étude cible la tâche spécifique de clustering. Des travaux récents ont souligné le fait que le clustering multi-vues pouvait être d'intérêt pour différents contextes, tels que :

La **réutilisation de connaissances** consistant à utiliser plusieurs clusterings différents exécutés chacun sur un même jeu de données et considérés comme une connaissance sur ce jeu de donnée. On peut alors mettre en place un nouveau processus combinant simultanément ces différentes connaissances. Chaque résultat de clustering est alors considéré comme une vue pour un processus de clustering multi-vues.
Le clustering de données multiples peut alors concerner autant la temporalité (resp. spatialité) lorsqu'un même ensemble d'individus est observé à des instants (resp. sur des sites) différents que la multiplicité lorsque l'on traite différents ensembles d'individus (*e.g.* différentes sociétés) (Strehl et Ghosh, 2003).

Le **calcul distribué** intervient quand les données ou les descriptions sont réparties sur plusieurs sites géographiques [1] avec l'impossibilité de les collecter en un site unique, ceci pour des raisons techniques ou de confidentialité (Strehl et Ghosh, 2003; Pedrycz, 2002).

L'**amélioration de la qualité et de la robustesse** peut être un objectif recherché par une approche de clustering multi-vues. La diversité des descriptions fournies par les différentes vues peut aider à l'apprentissage de meilleures solutions de clustering, qui satisfont simultanément chaque espace de représentation (vue). Des résultats expérimentaux présentés dans la dernière partie de la présente étude tendent à renforcer cette hypothèse.

[1] Nous ne considérons ici que la distribution verticale, dans le sens où la description de chaque objet est divisée en fragment sur des sites distants. Chaque site connait alors pour chaque objet, un sous-ensemble de l'ensemble d'attributs.

1.3 Stratégie pour le clustering multi-vues

Plusieurs approches de classification supervisée et semi-supervisée, apprenant des modèles à partir de plusieurs vues, ont été conçues ces dernières années (Blum et Mitchell, 1998; Ghani, 2002; Ganchev et al., 2008). Ces approches cherchent à tirer parti de la multiplicité des données en utilisant notamment une notion de désaccord entre les différentes vues comme un pseudo superviseur qui complémente les labels disponibles.
Le processus peut alors être transposé dans des situations complètement non supervisées, en utilisant l'accord entre les vues comme un pseudo superviseur qui complémente les mesures de qualités usuelles (nous rappellerons juste que de telles mesures de qualité guident les processus de clustering simple-vue de telle sorte que des objets similaires tendent à appartenir à un même groupe (Jain et al., 1999).
Dans le cas particulier du clustering multi-vues, nous distinguons trois stratégies de combinaison selon que cette dernière est réalisée avant, pendant, ou après le processus de clustering.

Stratégie de concaténation. La première stratégie (appelée aussi fusion *a priori* dans la suite), consiste en une concaténation des vues au sein d'une seule, soit directement en juxtaposant les ensembles d'attributs, ou indirectement en combinant les matrices de proximités issues de chaque vue. Par exemple, dans (Heer et Chi, 2002), les auteurs combinent la description lexicale et contextuelle des pages web dans une matrice de similarité obtenue par combinaison linéaire de similarités entre vecteurs de termes, puis entre contenus hypertextuels. Dans (Yamanishi et al., 2004), une proximité entre gènes est représentée par une matrice noyau obtenue par la somme pondérée de plusieurs noyaux, chacun étant obtenu par un processus specifique sur les expressions, localisations, ainsi que les profils phylogénétiques et compatibilités chimiques.

La stratégie de concaténation est une solution naturelle qui mène en pratique vers de bons résultats. Malheureusement, comme nous l'avons mentionné auparavant, une telle concaténation est impossible dans certains contextes : lorsque les données proviennent de différentes sources avec des restrictions de confidentialité, ou de plusieurs sites géographiques avec des limitations de stockage ou de bande passante. Enfin, une telle concaténation est inenvisageable dans le cas général de la grande dimensionnalité, dans lequel tout processus de classification devient inefficace (phénomène bien connu de la concentration des normes).

Stratégie distribuée. Connue aussi comme la problématique *clustering ensembles* ou de fusion *a posteriori*, cette stratégie procède par un clustering local et indépendant dans chaque vue, puis par la recherche d'une solution représentant un consensus sur l'ensemble des clusterings obtenus. Par exemple, (Strehl et Ghosh, 2003) proposent plusieurs heuristiques visant à trouver un meilleur consensus selon un critère objectif correspondant à une moyenne d'information mutuelle normalisée entre chaque paire de vues. [(Long et al., 2008)] définissent une mesure de qualité basée sur une I-divergence généralisée associée à un problème d'optimisation. (Fred, 2001) propose un algorithme de type vote basé sur une matrice de coassociation qui résume plusieurs résultats de clustering. Un panorama plus complet des différentes techniques de *clustering ensembles* est disponible dans (Reza et al., 2009).

Le principal inconvénient de ces méthodes repose sur le fait qu'elle ne reconsidèrent pas les résultats de clusterings donnés en entrée du processus de recherche de consensus. Néanmoins, (Forestier et al., 2008) ont proposé une stratégie collaborative qui modifie les clusterings existants pour les faire converger vers un résultat unique. Il est alors possible d'envisager une

troisième stratégie (centralisée) où les clusterings ne sont pas donnés en entrée du processus de collaboration, mais plutôt appris durant ce processus.

Stratégie centralisée. Cette dernière stratégie vise à utiliser la description multiple des données simultanément dans le but d'en extraire des motifs cachés. Cela représente un défi important et requiert de modifier profondément le processus de clustering. Autant que nous le sachons, deux principales méthodes centralisées ont été proposées dans la litérature, d'abord par (Pedrycz, 2002), puis par (Bickel et Scheffer, 2005). Toutes deux présentent une amélioration significative par rapport aux solutions de stratégie distribuée et de concaténation.

Dans ce papier, nous présentons (Section 2) quelques limitations des approches centralisées existantes, concernant aussi bien leur applicabilité que leurs définitions théoriques. Nous proposons ensuite un nouvel algorithme de clustering multi-vues issue des approches floues (Section 3) de telle sorte que les trois stratégies de combinaison peuvent être instanciées selon l'ajustement d'un seul paramètre. Nous montrons également comment notre approche peut être étendue (par l'astuce du noyau) pour prendre en compte une plus grande variété de données (Section 4). De plus, nous proposons une première comparaison expérimentale entre les approches centralisées, basée à la fois sur des données synthétiques et des données réelles.

2 Etat de l'art des approches centralisées

Dans cette section, nous introduisons la notation nécessaire et présentons brièvement les deux approches centralisées de la litérature : "*Collaborative Fuzzy Clustering*" and "*CoEM*" proposés respectivement par Pedrycz et Bickel&Sheffer.

2.1 Notations

Dans la suite nous considérons un ensemble de données $X = \{x_1, \ldots, x_n\}$ à organiser en K groupes. Les données sont décrites sur un ensemble de vues R telles que N_r est la dimensionalité de la vue r et $x_{i,r}$ dénote le vecteur d'attributs x_i de la vue r. Dans le cadre du clustering flou, on utilise $u_{i,k,r} \in [0, 1]$ pour modéliser le degré d'appartenance dans la vue r de l'objet $x_{i,r}$ au groupe k, $c_{k,r}$ correspond au centre du groupe k et $d_r(x_i, c_k)$ dénote la distance euclidienne entre l'objet x_i et c_k dans la vue r. De manière analogue, dans le contexte des modèles génératifs, $P(k|x_{i,r}, \Theta_r)$ est la probabilité *a posteriori* que x_i ait été généré par la composante k dans la vue r et Θ_r correspond aux paramètres des lois du mélange dans la vue r.

2.2 *Collaborative Fuzzy Clustering*

(Pedrycz, 2002) utilise le modèle des k-moyennes floues (Bezdek, 1981) et en dérive une variante collaborative (CoFC) pour le contexte multi-vues. La collaboration entre les vues concerne seulement les degrés d'appartenances $\{u_{i,k,r}\}$; dans ce cas la confidentialité est satisfaite et les coûts de stockage et de bande passante sont fortements réduits. Un terme d'inertie locale (à minimiser) est défini (1) par une inertie des centres flous locaux (vue r et correspondant au premier terme), pénalisé par un désaccord avec les autres vues r' (second terme). Le désaccord entre les vues r et r' est pondéré *via* $\alpha_{r,r'}$ (fixé) qui modélise une information connue *a priori* à propos d'une collaboration souhaitée entre les deux vues r et r'.

$$Q_{CoFC}(r) = \sum_{k=1}^{K} \sum_{x_i \in X} u_{i,k,r}^2 d_r(x_i, c_k)^2 \quad (1)$$
$$+ \sum_{r'=1}^{|R|} \alpha_{r,r'} \sum_{k=1}^{K} \sum_{x_i \in X} (u_{i,k,r} - u_{i,k,r'})^2 d_r(x_i, c_k)^2$$

Notons que la version originale des k-moyennes floues contient un paramètre $\beta > 1$ qui est ici fixé à 2 dans le modèle CoFC, pour des raisons de convergence de l'algorithme d'optimisation associé au problème. De ce point de vue, CoFC est moins général que k-moyennes floues ; nous montrerons par la suite l'importance du paramètre β et son influence sur la qualité de la solution obtenue. Enfin, la minimisation de $Q_{CoFC}(r)$ passe par une mise à jour non intuitive des paramètres, en particulier, la mise à jour d'un degré d'appartenance $u_{i,k,r}$ dépend seulement des $u_{j,k,r'}$ ($j \neq i$) des autres vues (et non de la vue courante).

2.3 CoEM

Bickel&Sheffer ont proposé dans (Bickel et Scheffer, 2005) un processus très similaire à CoFC mais construit dans le cadre génératif des modèles de mélanges. Ils proposent une variante collaborative (CoEM) de l'algorithme EM communément utilisé pour le clustering (Dempster et al., 1977). La fonction objective à maximiser (3), combine les log-vraisemblances locales $Q_{EM}(r)$ (2) de toutes les vues avec un terme de désaccord Δ (4) :

$$Q_{EM}(r) = E[\log P(X, Z|\Theta_r)|X, \Theta_r^t] \quad (2)$$

$$Q_{CoEM} = \sum_{r=1}^{|R|} Q_{EM}(r) - \eta\Delta \quad (3)$$

Le critère simple vue (2) dans la vue r utilise les variables X des données (de cette même vue), les variables cachées Z, les paramètres d'optimisation Θ_r et les estimateurs Θ_r^t à l'étape précédente.

Dans (3), η ajuste la contribution du désaccord Δ dans le processus d'optimisation. Ce désaccord est proche d'une divergence de Kullback-Leibler entre les distributions de probalités *a posteriori* sur toutes les paires de vues :

$$\Delta = \frac{1}{|R|-1} \times \quad (4)$$
$$\sum_{r \neq r'} \sum_{x_i \in X} \sum_{k=1}^{K} P(k|x_{i,r}, \Theta_r^t) \log \frac{P(k|x_{i,r}, \Theta_r)}{P(k|x_{i,r'}, \Theta_{r'})}$$

Les auteurs donnent une formulation de (3)[2] par une somme de log-vraisemblances sur chaque vue où les probabilités *a posteriori* sont obtenues par des moyennes pondérées des

[2] en introduisant le désaccord sous la somme sur chaque vue

probabilités *a posteriori* locales. Malheureusement, en utilisant la nouvelle expression (cette fois intuitive) pour le calcul des probabilités *a posteriori* le critère ne peut pas être maximiser dans son ensemble sauf en annulant la contribution du désaccord ($\eta \to 0$) et ainsi en ne tenant plus compte de la collaboration à travers les itérations. Notons qu'une affectation finale au groupe est obtenue à partir des paramètres du modèle collaboratif appris.

Le modèle CoFKM (*Collaborative Fuzzy K-means*) que nous proposons dans cette étude enrichit les approches présentées ici et offre une solution aux problèmes pratiques et théoriques rencontrés. Pour palier au problème de la convergence CoFKM se positionne dans le cadre flou de Pedrycz ; un nouveau terme de désaccord (inspiré de CoEM) est proposé pour rendre le modèle plus simple à paramétrer et le processus d'apprentissage plus intuitif. Enfin, le paramètre η utilisé par Bickel&Sheffer pour assurer la convergence est conservé dans CoFKM car il permet d'instancier ce modèle dans les trois stratégies de fusion : concaténation, distribuée et centralisée.

3 Le modèle CoFKM

L'approche que nous proposons ici est une extension des k-moyennes floues FKM (Bezdek, 1981). Nous essayons d'obtenir dans chaque vue une organisation spécifique, mais, comme dans CoEM, nous introduisons un terme de pénalisation visant à réduire le désaccord entre les différentes vues.

Rappelons que le but de FKM (Bezdek, 1981) est d'optimiser un critère d'inertie pondérée :

$$Q_{FKM} = \sum_{k=1}^{K} \sum_{x_i \in X} u_{i,k}^{\beta} d(x_i, c_k)^2 \tag{5}$$

avec $\forall x_i \in X$, $\Sigma_{k=1}^{K} u_{i,k} = 1$, où les variables du problème sont les centres de groupes (c_k) et les degrés d'appartenances des objets x_i aux groupes ($u_{i,k}$). Les valeurs optimales de ces paramètres sont données par :

$$c_k = \frac{\sum_{x_i \in X} u_{i,k}^{\beta} x_i}{\sum_{x_i \in X} u_{i,k}^{\beta}} \quad , \quad u_{i,k} = \frac{d(x_i, c_k)^{2/(1-\beta)}}{\sum_{k=1}^{K} d(x_i, c_k)^{2/(1-\beta)}}$$

3.1 Critère optimisé

Etant donné un ensemble R de vues, on note $Q_{FKM}(r)$ le critère associé à la vue $r \in R$. Dans chaque vue, les objets sont décrits par un vecteur de $\mathbb{R}^{N_r}$, où N_r est la dimensionalité de la vue r.

Nous proposons une approche collaborative, basée sur FKM, visant à minimiser $Q_{FKM}(r)$ dans chaque vue, et pénalisant le désaccord entre les paires de vues. Le critère à minimiser peut alors être réécrit :

$$Q_{CoFKM} = \left(\sum_{r\in R} Q_{FKM}(r)\right) + \eta\Delta \quad (6)$$

$$= \sum_{r\in R}\sum_{x_i\in X}\sum_{k=1}^{K} u_{i,k,r}^{\beta} d_r(x_i,c_k)^2 + \eta\Delta \quad (7)$$

où les vues sont normalisées pour avoir des inerties comparables dans toutes les vues. Cette normalisation est réalisée de la manière suivante :

- chaque variable est réduite (variance unitaire pour toutes les variables),
- un poids égal à $N_r^{-1/2}$ est associé à chaque variable appartenant à la vue r.

Dans la définition précédente de Q_{CoFKM}, Δ est un terme de désaccord : lorsque les organisations obtenues dans toutes les vues sont semblables, ce terme doit être égal à zéro. Nous proposons :

$$\Delta = \frac{1}{|R|-1}\sum_{r\neq r'}\sum_{x_i\in X}\sum_{k=1}^{K}(u_{i,k,r'}^{\beta} - u_{i,k,r}^{\beta}) d_r(x_i,c_k)^2$$

Dans cette expression, on somme les différences entre les organisations obtenues des vues r et r', pour tout couple de vues (r,r'). L'expression précédente peut-être écrite comme une somme sur les paires (r,r') telles que $r > r'$:

$$\Delta = \frac{1}{|R|-1}\sum_{r>r'}\sum_{x_i\in X}\sum_{k=1}^{K}(u_{i,k,r'}^{\beta} - u_{i,k,r}^{\beta})(d_r(x_i,c_k)^2 - d_{r'}(x_i,c_k)^2)$$

Le terme de désaccord est fait pour pénaliser notre critère. Il peut être considéré comme une divergence entre les organisations puisque plus $(u_{i,k,r}^{\beta} - u_{i,k,r'}^{\beta})$ est petit, plus faible est le désaccord. La normalisation que nous appliquons implique que $d_{r'}(x_i,c_k)$ et $d_r(x_i,c_k)$ sont comparables. $d_r(x_i,c_k)$ étant inversement proportionnel à $u_{i,k,r}$, on peut considérer le terme $(d_{r'}(x_i,c_k) - d_r(x_i,c_k))$ comparable à $(u_{i,k,r} - u_{i,k,r'})$, ainsi, le désaccord peut-être vu comme une distance entre les organisations locales $(u_{i,k,r})$ et $(u_{i,k,r'})$.

L'avantage est que notre terme de désaccord a le même ordre de grandeur que l'inertie locale, ainsi la somme de ces expressions peut être considérée comme un critère global cohérent Q_{CoFKM} (6).

Finalement, Q_{CoFKM} peut-être écrit :

$$Q_{CoFKM} = \sum_{r\in R}\sum_{x_i\in X}\sum_{k=1}^{K} u_{i,k,r,\eta} d_r(x_i,c_k)^2 \quad (8)$$

où

$$u_{i,k,r,\eta} = (1-\eta)u_{i,k,r}^{\beta} + \frac{\eta}{|R|-1}(\sum_{\overline{r}} u_{i,k,\overline{r}}^{\beta}) \quad (9)$$

Ce critère intuitif correspond alors à une inertie pondérée où, dans chaque vue, $u_{i,k,r,\eta}$ est une moyenne pondérée des degrés d'appartenances usuels $(u_{i,k,r'}^{\beta})$ provenant de chaque vue.

3.2 Résolution du problème d'optimisation

Comme dans le cas de FKM, notre but est de trouver une solution minimisant le critère global (8) et satisfaisant $\Sigma_k u_{i,k,r} = 1$ pour toute vue r et pour tout objet x_i. Pour résoudre ce problème d'optimisation sous contraintes, on considère le Lagrangien associé :

$$L(C, U, \lambda) = Q_{CoFKM} + \sum_{r \in R} \sum_{x_i \in X} \lambda_{r,i}(1 - \sum_{k=1}^{K} u_{i,k,r})$$

où C est une matrice contenant les centres dans chaque vue, U est la matrice contenant les degrés d'appartenances de chaque objet à chaque groupe et dans chaque vue, et λ est le vecteur des multiplicateurs de Lagrange. Si (C^*, U^*) est un optimum (local), $\nabla L(C^*, U^*) = 0$ est une condition nécessaire, *i.e.* les dérivées partielles par rapport aux variables du problème $u_{i,k,r}$ et $c_{k,r}$ s'annulent.

Ces dérivées partielles sont :

$$\begin{aligned}
\frac{\partial L}{\partial u_{i,k,r}} &= (1-\eta)\beta u_{i,k,r}^{\beta-1} d_r(x_i, c_k)^2 \\
&\quad + \frac{\eta}{|R|-1}\Big(\sum_{\overline{r}} \beta u_{i,k,r}^{\beta-1} d_{\overline{r}}(x_i, c_k)^2\Big) - \lambda_{r,i} \\
\frac{\partial L}{\partial c_{k,r}} &= -2 \sum_{x_i \in X} u_{i,k,r,\eta}(x_{i,r} - c_{k,r})
\end{aligned}$$

Comme pour FKM, nous proposons un algorithme alternant deux étapes d'optimisation :

- calcul des centres $c_{k,r}$ à partir des degrés $u_{i,k,r}$;
- calcul des degrés $u_{i,k,r}$ à partir des centres $c_{k,r}$.

Les équations $\frac{\partial L}{\partial c_{k,r}} = 0$ et $\frac{\partial L}{\partial u_{i,k,r}} = 0$ impliquent respectivement :

$$\begin{aligned}
c_{k,r} &= \frac{\sum_{x_i \in X} u_{i,k,r,\eta} x_{i,r}}{\sum_{x_i \in X} u_{i,k,r,\eta}} \qquad (10) \\
u_{i,k,r} &= \Big(\frac{\lambda_{r,i}}{\beta}\Big)^{1/(\beta-1)} \Big((1-\eta) d_r(x_i, c_k)^2 \\
&\quad + \frac{\eta}{|R|-1} \sum_{\overline{r}} d_{\overline{r}}(x_i, c_k)^2\Big)^{1/(1-\beta)}
\end{aligned}$$

En sommant les termes $u_{i,k,r}$ sur k et en utilisant la contrainte $\Sigma_{k=1}^{K} u_{i,k,r} = 1$, on obtient :

$$u_{i,k,r} = \frac{\Big((1-\eta) d_r(x_i, c_k)^2 + \frac{\eta}{|R|-1} \sum_{\overline{r}} d_{\overline{r}}(x_i, c_k)^2\Big)^{1/(1-\beta)}}{\sum_{k=1}^{K} \Big((1-\eta) d_r(x_i, c_k)^2 + \frac{\eta}{|R|-1} \sum_{\overline{r}} d_{\overline{r}}(x_i, c_k)^2\Big)^{1/(1-\beta)}} \qquad (11)$$

Algorithme 1 CoFKM

Entrée : Ensemble d'objets X, Nombre de groupes K, Nombre de vues $|R|$
Initialiser les mêmes K centres de groupes pour toutes les vues $r \in R$.
repeat
  for $i = 1; r = 1; k = 1$ **to** $|X|; |R|; K$ **do**
    maj $u_{i,k,r}$ en utilisant (11)
  end for
  for $i = 1; r = 1; k = 1$ **to** $|X|; |R|; K$ **do**
    maj $u_{i,k,r,\eta}$ en utilisant (9)
  end for
  for $r = 1; k = 1$ **to** $|R|; K$ **do**
    maj $c_{k,r}$ en utilisant (10)
  end for
until convergence
affecter x_i au groupe C_k en utilisant la règle (12)

Par conséquent, à chaque étape, nous calculons la valeur optimale de $c_{k,r}$ (resp. $u_{i,k,r}$) pour des valeurs fixées de $u_{i,k,r}$ (resp. $c_{k,r}$). Ainsi, par cet algorithme, la décroissance du critère (8) est garantie, ce qui assure la convergence (vers un optimum local).

3.3 Règle d'affectation

La méthode proposée assure l'obtention d'un optimum local du critère global. Cependant, elle produit dans chaque vue, des centres de groupes et des degrés d'appartenances potentiellement différents. Ainsi, dans le but d'obtenir un résultat de clustering unique, nous devons fusionner ces résultats locaux. Nous proposons de construire une partition des objets en utilisant une règle d'affectation qui assigne un seul groupe à chaque objet. Cette règle consiste à calculer, pour chaque objet et chaque groupe, un degré d'appartenance global, correspondant à une moyenne géométrique des degrés d'appartenances locaux :

$$\hat{u}_{i,k} = \sqrt[|R|]{\prod_{r \in R} u_{i,k,r}} \tag{12}$$

l'objet x_i est affecté au groupe k maximisant $\hat{u}_{i,k}$.

Cette règle requiert l'association d'un même groupe dans plusieurs vues. Nous considérons ici qu'un groupe est identifié par un même index $k \in [1..K]$ dans toutes les vues. La consistance de cette identification est assurée par la façon dont sont initialisées et mises à jour les variables. L'initialisation consiste à choisir aléatoirement un objet comme centre de tous les groupes de même index. Ainsi, pour tous k, les centres $c_{k,r}$ correspondent à toutes les vues du même objet. L'algorithme est présenté dans Algorithme 1.

3.4 Positionnement théorique

Nous montrons ici que notre modèle CoFKM est une généralisation à la fois de FKM appliqué à la concaténation des vues (fusion *a priori*), et d'un simple modèle *a posteriori* où

FKM est appliqué indépendamment dans chaque vue. Considérons les expressions de $c_{k,r}$ et $u_{i,k,r}$ proposées dans les équations (10) et (11), où les termes correspondant aux différentes vues sont pondérés de manière équivalente, *i.e.* $(1-\eta) = \frac{\eta}{|R|-1}$ ou $\eta = \frac{|R|-1}{|R|}$.

Dans ce cas,

$$u_{i,k,r,\eta} = \frac{1}{|R|} u_{i,k,r}^{\beta} + \frac{1}{|R|} \sum_{\overline{r}} u_{i,k,\overline{r}}^{\beta} = \frac{1}{|R|} \sum_{r'} u_{i,k,r'}^{\beta}$$

$u_{i,k,r,\eta}$ ne dépend pas de r, et l'expression de $c_{k,r}$ correspond exactement à celle de FKM appliqué à la concaténation des vues.

De la même manière, $d_r(x_i, c_k)^2 + \sum_{\overline{r}} d_{\overline{r}}(x_i, c_k)^2$ est la distance entre x_i et c_k associée à la concaténation des vues, l'expression de $u_{i,k,r}$ correspond alors exactement à celle de FKM dans le cas de la concaténation.

Finalement, on peut voir CoFKM comme une généralisation de FKM appliquée à la concaténation des vues, où l'on peut forcer l'obtention d'une solution correspondant à un consensus en choisissant une valeur η plus petite que $\frac{(|R|-1)}{|R|}$.

Les expériences montrent que le signe du terme de désaccord dépend de la valeur de η. Quand $\eta > \frac{(|R|-1)}{|R|}$, le terme de désaccord est négatif, ce qui n'est pas espéré pour un terme de désaccord. Pour cette raison, nous proposons de choisir $0 \leq \eta \leq \frac{(|R|-1)}{|R|}$.

Considérons maintenant le modèle CoFKM avec $\eta = 0$, le critère Q_{CoFKM} peut alors être réécrit comme une somme sur toutes les vues des critères FKM classiques :

$$\begin{aligned} Q_{CoFKM_{\eta=0}} &= \left(\sum_{r \in R} Q_{FKM(r)} \right) \\ &= \sum_{r \in R} \sum_{x_i \in X} \sum_{k=1}^{K} u_{i,k,r}^{\beta} d_r(x_i, c_k)^2 \end{aligned}$$

Ce critère est la somme des inerties locales, qui sont optimisées de manière indépendantes par l'algorithme FKM. La fusion *a posteriori* est réalisée par notre règle d'affectation. Notre modèle collaboratif CoFKM est alors une généralisation de la fusion *a posteriori*, en choisissant $\eta = 0$.

Nous comparons maintenant notre modèle à l'approche CoEM. L'inconvénient théorique majeur de CoEM réside en la non convergence de l'algorithme associé. Pour assurer cette convergence, (Bickel et Scheffer, 2005) proposent de faire décroître le paramètre η jusqu'à 0, ce qui correspond à l'optimisation du critère local indépendamment dans toutes les vues. CoEM peut ainsi être vu comme une approche en deux temps : durant la première phase ($\eta > 0$) les paramètres sont estimés dans le but d'accroître le consensus mais sans garanties de convergence ; lors de la seconde phase ($\eta = 0$) le critère global converge par convergence locale dans toutes les vues, mais le terme de pénalité n'est pas considéré. Notre modèle est défini de telle sorte que quelquesoit la valeur de η, la convergence est assurée puisque le critère global décroît à chaque étape de l'algorithme.

4 Variante à noyaux

4.1 Le modèle CoKFKM

Le modèle CoFKM généralise le modèle classique des k-moyennes floues mais se voit toujours restreint à l'utilisation de la métrique euclidienne. En particulier, ce modèle ne s'applique que dans le cas où les données sont décrites par des vecteurs d'attributs numériques. Nous introduisons ici une formalisation pour une version à noyau du modèle CoFKM, dans le but de trouver un clustering où l'ensemble des objets est décrit par plusieurs matrices de similarités. On peut alors tenter d'obtenir dans chaque vue une organisation spécifique en utilisant une matrice de similarité semi-définie positive, tout en réduisant le désaccord entre les vues pour obtenir un clustering consensus.

Plusieurs travaux ont montré comment utiliser l'astuce du noyau dans des approches de clustering classiques. Cela permet de découvrir des groupes qui sont non-linéairement séparables dans l'espace de description d'origine, par une projection implicite des objets dans un espace de plus grande dimension (ϕ est la fonction de projection non linéaire). Cette astuce a été appliquée par exemple dans les k-moyennes floues standard, où la fonction objective devient :

$$Q_{KFKM} = \sum_{k=1}^{K} \sum_{x_i \in X} u_{i,k}^{\beta} d(\phi(x_i), c_k)^2$$

Les valeurs optimales des paramètres c_k et $u_{i,k}$ sont données par :

$$c_k = \frac{\sum_{x_i \in X} u_{i,k}^{\beta} \phi(x_i)}{\sum_{x_i \in X} u_{i,k}^{\beta}} \tag{13}$$

$$u_{i,k} = \frac{d(\phi(x_i), c_k)^{2/(1-\beta)}}{\sum_{k=1}^{K} d(\phi(x_i), c_k)^{2/(1-\beta)}}$$

Il a été montré que même si les centres optimaux ne peuvent pas être calculés (car ϕ est en général inconnue), on peut optimiser ce critère grâce à l'utilisation d'une matrice noyau. Soit K une matrice noyau donnée (définie positive) représentant le produit scalaire des vecteurs dans l'espace de projection (*i.e* $K_{ij} = \langle \phi(x_i), \phi(x_j) \rangle$), alors en utilisant l'expression (13) on peut calculer le carré de la distance euclidienne entre $\phi(x_i)$ et c_k :

$$d(\phi(x_i), c_k)^2 = K_{ii} - 2\frac{\sum_{x_j \in X} u_{j,k}^{\beta} K_{ij}}{\sum_{x_j \in X} u_{j,k}^{\beta}} + \frac{\sum_{x_j \in X} \sum_{x_l \in X} u_{j,k}^{\beta} u_{l,k}^{\beta} K_{jl}}{(\sum_{x_j \in X} u_{j,k}^{\beta})^2}$$

Les centres sont implicitement déplacés dans l'espace de projection lors du calcul des nouvelles distances (dépendantes des nouveaux estimateurs des degrés d'appartenances). On

Algorithme 2 CoKFKM

Entrée : Ensemble d'objets X, Nombre de groupes K, Nombre de vues $|R|$, Matrices noyaux $\{K_i\}_{1\leq|R|}$ (un noyau pour chaque vue)
repeat
  for $i=1; r=1; k=1$ **to** $|X|; |R|; K$ **do**
    maj $u_{i,k,r}$ en utilisant (18)
  end for
  for $i=1; r=1; k=1$ **to** $|X|; |R|; K$ **do**
    maj $u_{i,k,r,\eta}$ en utilisant (9)
  end for
  for $r=1; k=1; i=1$ **to** $|R|; K; |X|$ **do**
    maj $d_r(\phi(x_i), c_k)$ en utilisant (17)
  end for
until convergence
affecter x_i au groupe C_k en utilisant la règle (12)

peut alors transposer ce résultat au modèle CoFKM pour obtenir une version à noyau de la manière suivante :

$$Q_{CoKFKM} = \left(\sum_{r\in R} Q_{KFKM}(r)\right) + \eta\Delta \tag{14}$$

$$= \sum_{r\in R}\sum_{x_i\in X}\sum_{k=1}^{K} u_{i,k,r}^{\beta} d_r(\phi(x_i), c_k)^2 + \eta\Delta \tag{15}$$

avec

$$\Delta = \frac{1}{|R|-1}\sum_{r\neq r'}\sum_{x_i\in X}\sum_{k=1}^{K}(u_{i,k,r'}^{\beta} - u_{i,k,r}^{\beta}) d_r(\phi(x_i), c_k)^2$$

A l'instar de CoFKM, CoKFKM peut également être exprimé :

$$Q_{CoKFKM} = \sum_{r\in R}\sum_{x_i\in X}\sum_{k=1}^{K} u_{i,k,r,\eta} d_r(\phi(x_i), c_k)^2 \tag{16}$$

où

$$d_r(\phi(x_i), c_k)^2 = K_{r_{ii}} - 2\frac{\sum_{x_j\in X} u_{j,k,r,\eta}^{\beta} K_{r_{ij}}}{\sum_{x_j\in X} u_{j,k,r,\eta}^{\beta}} + \frac{\sum_{x_j\in X}\sum_{x_l\in X} u_{j,k,r,\eta}^{\beta} u_{l,k,r,\eta}^{\beta} K_{r_{jl}}}{(\sum_{x_j\in X} u_{j,k,r,\eta}^{\beta})^2} \tag{17}$$

et $u_{i,k,r,\eta}$ est obtenu de la même manière qu'en (9).

Finalement, les paramètres optimaux $u_{i,k,r}$ sont obtenus de la même manière que pour CoFKM (11) mais le carré de la distance euclidienne dans l'espace de description d'origine est remplacé par le carré de la distance euclidienne dans l'espace de projection, c'est à dire :

$$u_{i,k,r} = \frac{((1-\eta)d_r(\phi(x_i),c_k)^2 + \frac{\eta}{|R|-1}\sum_{\overline{r}} d_{\overline{r}}(\phi(x_i),c_k)^2)^{1/(1-\beta)}}{\sum_{k=1}^{K}((1-\eta)d_r(\phi(x_i),c_k)^2 + \frac{\eta}{|R|-1}\sum_{\overline{r}} d_{\overline{r}}(\phi(x_i),c_k)^2)^{1/(1-\beta)}} \quad (18)$$

L'algorithme est présenté dans Algorithme 2. La version à noyaux CoKFKM généralise complètement CoFKM. En effet, il suffit de choisir comme matrices noyaux dans toutes les vues les matrices des produits scalaires entre les objets dans l'espace de description d'origine. En d'autres termes, il suffit de choisir $K_{r_{ij}} = \langle \phi(x_i), \phi(x_j) \rangle = \langle x_i, x_j \rangle$. Ainsi on a bien :

$$\begin{aligned} d_r(\phi(x_i),c_k)^2 &= \langle x_i,x_i,\rangle_r - 2\frac{\sum_{x_j\in X} u^{\beta}_{j,k,r,\eta}\langle x_i,x_j,\rangle_r}{\sum_{x_j\in X} u^{\beta}_{j,k,r,\eta}} + \frac{\sum_{x_j\in X}\sum_{x_l\in X} u^{\beta}_{j,k,r,\eta}u^{\beta}_{l,k,r,\eta}\langle x_j,x_l,\rangle_r}{(\sum_{x_j\in X} u^{\beta}_{j,k,r,\eta})^2} \\ &= d_r(x_i,c_k)^2 \end{aligned}$$

avec $\langle x_i, x_j, \rangle_r$ correspondant au produit scalaire entre x_i et x_j dans la vue r. Le critère optimisé correspond exactement à celui de CoFKM ; l'intérêt de CoKFKM réside dans la possibilité d'utiliser différentes matrices noyaux plus adaptées aux données.

4.2 Complexités

Nous présentons maintenant le calcul des complexités en temps des approches CoFKM et CoKFKM. L'objectif est ici de voir ce que l'on perd à utiliser CoKFKM (plus général) par rapport à CoFKM.

L'algorithme 1 (CoFKM) se décompose en trois étapes :

1. Calcul des degrés d'appartenances $u_{i,k,r}$. (équation (11))
   On doit calculer pour chaque x_i,k et r une somme pondérée sur les vues $r \in R$ de distances. La distance dans une vue r se calculant en $\mathcal{O}(N_r)$, le calcul d'un $u_{i,k,r}$ est alors en $\mathcal{O}(|R|.N_r)$. L'étape de mise à jour complète des degrés a pour complexité au pire des cas $\mathcal{O}(K.N_r.|R|^2.|X|)$.

2. Calcul des degrés collaboratifs $u_{i,k,r,\eta}$. (équation(9))
   Il suffit de calculer pour chaque x_i,k et r une somme pondérée sur les vues. La mise à jour de tous les $u_{i,k,r,\eta}$ se fait ainsi en $\mathcal{O}(K.|R|^2.|X|)$.

3. Calcul des centres $c_{k,r}$. (équation(10))
   Il suffit de calculer pour chaque k et r une moyenne pondérée sur les objets. La mise à jour de tous les $c_{k,r}$ a un coût de $\mathcal{O}(K.|R|.|X|)$.

La complexité à l'issue des trois étapes devient $\mathcal{O}(K.|R|.|X|(1 + |R| + N_r.|R|))$. Soit t le nombre d'itérations de l'algorithme, et en admettant que $1 + |R|$ soit négligeable devant

$N_r.|R|$, après simplifications, on considerera pour simplifier la complexité de CoFKM suivante : $\mathcal{O}(t.K.|R|.|X|(N_r.|R|))$.

Dans le cas de l'algorithme 2 (CoKFKM), des trois étapes de calculs, seule la dernière change, puisqu'il n'est pas possible de calculer explicitement les centres dans l'espace de projection. Ceux-ci sont déplacés implicitement pendant le recalcul des distances. De ce fait ces distances sont désormais stockées en mémoire, ce qui n'était pas nécéssaire dans CoFKM. Ainsi :

1. La mise à jour des degrés d'appartenances est moins coûteuse : $\mathcal{O}(K.|R|^2.|X|)$.
2. Le coût de mise à jour des degrés collaboratifs est inchangé : $\mathcal{O}(K.|R|^2.|X|)$.
3. La mise à jour des distances aux centres (17) se réalise en $\mathcal{O}(K.|R|.|X|^2)$.

La complexité à l'issue des trois étapes devient de l'ordre de $\mathcal{O}(K.|R|.|X|(|X| + 2.|R|))$. Après simplifications, la complexité de CoKFKM devient : $\mathcal{O}(t.K.|R|.|X|(|X| + |R|))$.

Si on émet les hypothèses suivantes (largement vérifiées dans les cas concrets d'applications)

- $|X| >> |R|$
- $N_r >> |R|$

alors les complexités des deux approches à comparer deviennent :

- $\mathcal{O}(t.K.|R|.|X|.N_r)$ pour CoFKM ;
- $\mathcal{O}(t.K.|R|.|X|.|X|)$ pour CoKFKM.

En d'autres termes, si le nombre d'objets est beaucoup plus grand que la dimensionnalité, alors l'approche CoFKM est moins complexe et plus rapide d'éxécution. En revanche, dans le cas de la malédiction de la dimensionnalité, où le nombre d'attributs est beaucoup plus grand que le nombre d'objets, l'approche à noyau devient moins complexe, et se justifie alors comme une variante efficace.

5 Résultats expérimentaux

Nous avons conduit des expériences et validé notre approche sur plusieurs jeux de données. Le premier est *multiple features*[3] disponible sur le dépôt *UCI Machine Learning*. Le second est un jeu de données artificiel utilisé dans (Strehl et Ghosh, 2003) [4]. Le dernier jeu de données est WebKB utilisé dans (Bickel et Scheffer, 2005). Nous avons observé une forte amélioration comparée à CoEM, CoFC et également à FKM et EM appliqués à chaque vue séparément.

[3] *http ://archive.ics.uci.edu/ml/*

[4] ce jeu s'appelle *2D2K* , disponible en libre téléchargement sur *http ://strehl.com/*

5.1 Jeux de données et Méthodologie d'évaluation

Le jeu de données *multiple features* correspond à un ensemble de 2000 chiffres manuscrits (images numérisées) décrites dans six vues differentes correspondant à six techniques d'encodage d'images (fou, fac, kar, mor, pix, zer). Dix classes sont à retrouver (de 0 à 9),avec 200 objets par classe.

Le jeu *2D2K* contient 1000 objets générés par un mélange de deux gaussiennes avec des matrices de covariances diagonales. Trois vues sont construites artificiellement, comme proposé dans (Strehl et Ghosh, 2003). Deux classes sont à retrouver.

Le jeu de donnée WebKB est un jeu de donnée réel correspondant à une collection de 4501 pages web académiques regroupées manuellement en six classes. Deux vues sont disponibles : la première contient le texte de chaque page web et la seconde correspond au texte de tous les liens entrants.

Dans le but de faire contribuer de manière équitable toute les vues, nous avons fait quelques prétraitement statistiques classiques. Les variables ont été centrées, réduites, et pondérées (cf. section 3.1).

L'évaluation d'un résultat de clustering est toujours un problème ouvert, car on ne connaît pas toujours le vrai étiquetage des objets. Cependant, lorsque toutes les étiquettes de classes sont disponibles, on peut utiliser un critère d'évaluation externe mesurant l'adéquation entre la classification obtenue par l'algorithme de clustering et la classification de référence. Nous avons choisi ici de mesurer la qualité des approches comparées par trois différentes mesures bien connues : La F-mesure, l'entropie moyenne, et l'information mutuelle normalisée.

La F-mesure[5] combine precision et rappel sur les paires d'objets ayant la même étiquette dans la classification de référence, ainsi que les paires d'objets appartenant au même groupe dans le clustering obtenu :

$$\text{Précision} = \frac{\text{Nb. Liens (paires) Identifiés Correctement}}{\text{Nb. Liens (paires) Identifiés}}$$

$$\text{Rappel} = \frac{\text{Nb. Liens (paires) Identifiés Correctement}}{\text{Nb. Liens (paires) Corrects}}$$

$$\text{F-mesure}(\beta) = \frac{(\beta^2+1) \times \text{Précision} \times \text{Rappel}}{\beta^2 \times \text{Précision} + \text{Rappel}}$$

L'entropie moyenne utilise les étiquettes de classes pour calculer la moyenne de l'impureté de touts les groupes :

$$\text{AvgEnt} = \sum_{k=1}^{K} \frac{n_k}{n} \left(-\sum_{c=1}^{C} p_{ck} \times log(p_{ck}) \right)$$

où K est le nombre de groupes, C est le nombre de classes. n_k et n sont respectivement le nombre d'objets dans le groupe k et le nombre total d'objets. p_{ck} correspond à la proportion d'objets de classe c dans le groupe k.

Finalement, l'information mutuelle normalisée quantifie l'information statistique partagée entre deux distributions, ici, les distributions des étiquettes de groupes et des étiquettes de classes :

[5] β=1 Dans les expériences.

$$\text{NMI} = \frac{2}{n} \sum_{k=1}^{K} \sum_{c=1}^{C} n_{kc} \times log_{K \times C} \left(\frac{n_{kc} \times n}{n_k \times n_c} \right)$$

où n_{kc} est le nombre d'objets simultanément dans le groupe k et dans la classe c.

5.2 Expériences

Nous avons conduit différentes expériences dans le but d'une part, de justifier l'intérêt des approches collaboratives centralisées comparées aux approches *a posteriori* et par concaténation. Puis de montrer empiriquement le gain de performance obtenu par rapport à ces techniques. De plus nous montrons que notre approche se positionne au mieux par rapport aux autres approches de la litérature. Enfin, nous montrons l'intérêt et le potentiel de la variante à noyaux sur un jeu de donnée réel (WebKB).

Les résultats que nous avons obtenu correspondent à une moyenne de 20 exécutions pour *multiple features*, 100 exécutions pour *2D2K* et 10 exécutions pour WebKB. Les différentes méthodes ont été comparées chaque fois avec la même initialisation. Les paramètres de notre modèle sont fixé à $\beta = 1.25$ (valeur couramment employée) et $\eta = \frac{|R|-1}{2\times|R|}$, ce qui correspond à notre heuristique entre les versions *a priori* et *a posteriori* de CoFKM.

5.2.1 Critères internes

Un premier objectif justifiant l'intérêt des approches centralisées concerne la stabilité de la qualité du clustering final au regard de chacune des vues. L'idée est ici d'observer si le clustering obtenu à l'issue du processus collaboratif est bon sur chacune des vues. Une telle observation confirmerait l'idée qu'une bonne solution globale peut être obtenue tout en assurant que toutes les vues s'accordent pour conforter la qualité de cette solution. Voici comment nous avons procédé :

- on compare les critères internes (inerties) obtenues par CoFKM et ses variantes *a priori* et *a posteriori*,
- on observe les valeurs de ses critères dans chacune des vues, et ceci à la fois avant et après la règle d'affectation (12).

L'objectif visé est qu'une solution consensus soit bonne sur toutes les vues (stable) au sens du critère interne avant la règle d'affectation, et que cette règle ne détériore pas trop cette stabilité.

Les figures 1 et 2 confirment nos intuitions sur les approches collaboratives. On remarque dans les deux cas qu'au sens du critère interne, et avant fusion CoFKM permet d'apprendre une solution meilleure que celle de sa variante concaténée (*a priori*) et surtout l'écart entre les inerties locales est plutôt faible dans le cas collaboratif (ce qui traduit la stabilité de la solution sur toutes les vues). La version *a posteriori* est celle qui optimise localement les inerties (sans collaborations), elle se positionne comme une référence (avant fusion). En revanche, si l'on observe l'impact de la règle d'affectation (permettant d'obtenir un clustering unique pour toutes les vues), l'approche sans collaborations se détériore complètement. Le résultat de référence après fusion est la concaténation qui reste inchangée puisque le degré d'appartenance d'un objet aux groupes est le même dans toutes les vues (avant ou après la règle). Nous constatons que CoFKM devient sensiblement équivalent à sa variante concaténée.

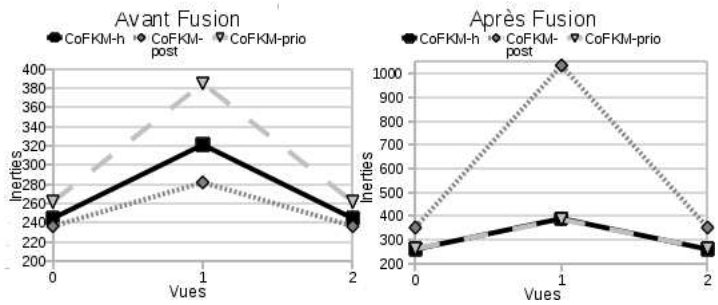

FIG. 1 – *Comparaisons des valeurs de critère interne dans chaque vue avant et après fusion (règle d'affectation) pour CoFKM et ses variantes* a priori *et* a posteriori *pour 2D2K.*

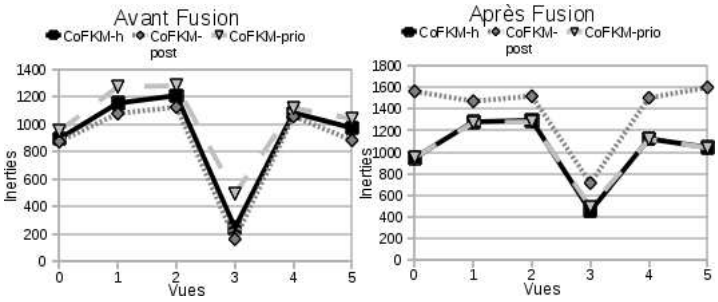

FIG. 2 – *Comparaisons des valeurs de critère interne dans chaque vue avant et après fusion (règle d'affectation) pour CoFKM et ses variantes* a priori *et* a posteriori *pour multiple features.*

5.2.2 Evaluation externe de CoFKM

Les tableaux 1, 3, 5 montrent les résultats obtenus pour le jeu de données *multiple features*. Les tableaux 2, 4, 6 montrent les résultats obtenus pour le jeu de données *2D2K*.

Nous avons observé que l'estimation des paramètres d'un modèle de mélange gaussien général était inefficace, nous avons donc choisi d'instancier ce modèle en trois modèles parsimonieux :

- (vs1) matrices de variances/covariances $\sigma_k.I$,
- (vs2) matrices $\sigma.I$ (le même σ pour toutes les composantes du mélange),
- (vs3) matrices diagonales.

On peut observer pour *multiple features* dans le tableau 1 que CoFKM surpasse les approches de l'état de l'art CoEM et CoFC. Nous avons cherché à savoir pourquoi CoFC-*vue* fonctionnait si mal pour ce jeu de données. Notre conclusion est que le choix figé du paramètre de flou semble être la cause de ce résultat dégénéré. Les résultats obtenus pour *2D2K* (Tableau 2) sont différents, cette fois CoFKM fait un peu moins bien que CoEM pour un mélange de

	% F-score	AvgEnt	NMI
CoFKM	**91.95 ± 0.00**	**0.29 ± 0.00**	**0.91 ± 0.00**
CoEM(vs1)	39.81 ± 5.34	1.61 ± 0.13	0.52 ± 0.04
CoEM(vs2)	82.80 ± 4.44	0.50 ± 0.09	0.85 ± 0.03
CoEM(vs3)	74.96 ± 5.42	0.72 ± 0.12	0.78 ± 0.04
CoFC-mor	31.22 ± 0.03	2.45 ± 0.00	0.26 ± 0.00

TAB. 1 – *Comparaison des modèles collaboratifs CoFKM, CoEM et CoFC pour multiple features.*

	% F-score	AvgEnt	NMI
CoFKM	94.18 ± 0.00	0.18 ± 0.00	0.82 ± 0.00
CoEM(vs1)	93.85 ± 1.09	0.18 ± 0.02	0.82 ± 0.02
CoEM(vs2)	**95.12 ± 0.00**	**0.15 ± 0.00**	**0.85 ± 0.00**
CoEM(vs3)	66.62 ± 0.00	1.00 ± 0.00	0.00 ± 0.00
CoFC-v2	94.17 ± 0.00	0.19 ± 0.00	0.81 ± 0.00

TAB. 2 – *Comparaison des modèles collaboratifs CoFKM, CoEM et CoFC pour 2D2K.*

gaussiennes sphériques (matrices $\sigma.I$), ce qui est très proche du critère objectif de FKM. Nous pouvons également nous référer à (Strehl et Ghosh, 2003) pour les résultats d'approches de type *cluster ensembles* pour *2D2K* et noter que dans ce cas CoFKM améliore également les heuristiques de *cluster ensembles* également.

	% F-score	AvgEnt	NMI
CoFKM	**92.01 ± 0.00**	**0.29 ± 0.00**	**0.91 ± 0.00**
CoFKMpost	55.72 ± 4.28	1.21 ± 0.12	0.64 ± 0.04
CoEMpost(vs1)	27.46 ± 9.01	2.53 ± 0.54	0.24 ± 0.16
CoEMpost(vs2)	57.20 ± 5.22	1.18 ± 0.14	0.65 ± 0.04
CoEMpost(vs3)	45.64 ± 5.21	1.54 ± 0.15	0.54 ± 0.05
FKMconcat	90.42 ± 3.44	0.33 ± 0.07	0.90 ± 0.02
EMconcat(vs1)	32.51 ± 6.68	1.77 ± 0.25	0.47 ± 0.08
EMconcat(vs2)	77.90 ± 5.72	0.56 ± 0.12	0.83 ± 0.04
EMconcat(vs3)	60.10 ± 5.53	1.04 ± 0.14	0.69 ± 0.04

TAB. 3 – *Comparaison entre CoFKM, et les variantes* a priori *et* a posteriori *pour multiple features.*

	% F-score	AvgEnt	NMI
CoFKM	94.18 ± 0.00	0.18 ± 0.00	0.82 ± 0.00
CoFKMpost	86.28 ± 13.27	0.34 ± 0.27	0.66 ± 0.27
CoEMpost(vs1)	80.43 ± 14.21	0.45 ± 0.29	0.55 ± 0.29
CoEMpost(vs2)	86.60 ± 14.69	0.32 ± 0.29	0.68 ± 0.29
CoEMpost(vs3)	85.47 ± 13.36	0.36 ± 0.27	0.64 ± 0.27
FKMconcat	**96.27 ± 0.00**	**0.13 ± 0.00**	**0.87 ± 0.00**
EMconcat(vs1)	93.18 ± 8.22	0.19 ± 0.15	0.81 ± 0.15
EMconcat(vs2)	**96.27 ± 0.00**	**0.13** ± 0.00	**0.87** ± 0.00
EMconcat(vs3)	96.07 ± 0.00	0.14 ± 0.00	0.86 ± 0.00

TAB. 4 – *Comparaison entre CoFKM, et les variantes* a priori *et* a posteriori *pour 2D2K.*

On observe que CoFKM améliore les solutions *a posteriori* de CoFKM et CoEM, et réussi à améliorer les variantes *a priori* de FKM et EM, voir les tableaux 3 et 4. La performance de CoFKM pour *multiple features* est un résultat très prometteur, l'objectif n'étant pas de surpasser la fusion *a priori*.

	% F-score	AvgEnt	NMI
CoFKM	**92.01 ± 0.00**	**0.29 ± 0.00**	**0.91 ± 0.00**
FKM-pix	70.41 ± 2.93	0.88 ± 0.06	0.74 ± 0.02
EM(vs1)-mor	38.20 ± 3.48	1.71 ± 0.15	0.48 ± 0.05
EM(vs2)-pix	63.38 ± 5.68	1.01 ± 0.13	0.70 ± 0.04
EM(vs3)-fac	63.78 ± 5.64	0.99 ± 0.13	0.70 ± 0.04

TAB. 5 – *Comparaison entre CoFKM, et les solutions simple-vues pour multiple features.*

	% F-score	AvgEnt	NMI
CoFKM	**94.18 ± 0.00**	**0.18 ± 0.00**	**0.82 ± 0.00**
FKM-v1	85.32 ± 5.88	0.40 ± 0.19	0.60 ± 0.19
EM(vs1)-2d2kv2	79.74 ± 4.26	0.50 ± 0.07	0.50 ± 0.07
EM(vs2)-2d2kv1	85.12 ± 5.82	0.40 ± 0.19	0.60 ± 0.19
EM(vs3)-2d2kv1	82.80 ± 6.41	0.44 ± 0.20	0.56 ± 0.20

TAB. 6 – *Comparaison entre CoFKM, et les solutions simple-vues pour 2D2K.*

Finalement, on observe pour ces deux jeux de données une nette amélioration sur les approches simple-vues (tableaux 5 et 6), ce qui est un résultat connu, et la principale motivation des approches multi-vues.

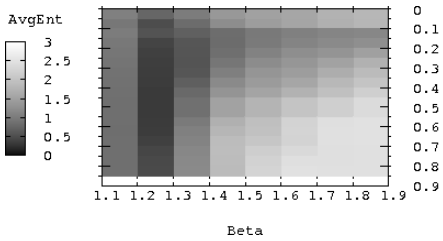

FIG. 3 – *Influence des paramètres η et β dans CoFKM pour multiple features.*

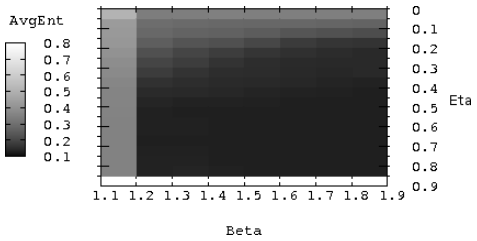

FIG. 4 – *Influence des paramètres η et β dans CoFKM pour 2D2K.*

Nous avons également testé l'influence des paramètres η and β sur la qualité des résultats. La courbe (Fig.3) présente le comportement de CoFKM pour le jeu de données *multiple features*. On peut noter qu'une valeur appropriée pour β devrait être proche de 1.2, la valeur $\beta = 2$ donne de très mauvais résultats, comme expliqué auparavent pour le résultat dégénéré de CoFC qui impose cette valeur. On peut aussi noter que notre choix heuristique de $\eta = \frac{|R|-1}{2\times|R|} = \frac{5}{12}$ donne de bons résultats (Fig.5). La courbe (Fig.4) indique que l'on peut choisir n'importe qu'elle valeur de β au delà de $\beta = 1.1$ et on peut observer que l'heuristique pour $\eta = \frac{|R|-1}{2\times|R|} = \frac{1}{3}$ donne des résultats corrects.

5.2.3 Apports de la variante à noyaux

Nous avons également étudié l'apport de l'extension à noyaux sur une partie du jeu de données WebKB (les 100 premiers individus). Ce jeu de données réel est assez difficile à traiter, puisqu'il réunit un certain nombre de conditions néfastes pour les approches de classifications usuelles. En effet, la dimensionnalité est très élevée comparée aux nombre d'objets (documents) disponibles. De plus, dans la vue représentant le contenu des liens entrants, beaucoup d'objets n'ont pas de descriptions et enfin les tailles des classes sont déséquilibrées. Nous avons comparé le modèle CoFKM avec CoEM pour un mélange de gaussiennes, puis avec l'extension CoKFKM en choisissant comme matrices de similarités, les cosinus entre les documents, considérés comme plus efficace sur les données textuelles que les produits scalaires classiques. Les algorithmes CoKFKM, CoFKM et CoEM ont été modifiés pour prendre en compte notamment les descriptions vides de la plupart des objets. En effet, lorsqu'un objet n'a pas de descriptions dans une vue, on ne l'intègre pas dans la définition des centres (directement, ou par le calcul des distances aux centres, dans le cas de la version à noyaux).

La figure 6 montre l'évolution de l'entropie moyenne en fonction du nombre de groupes demandé. Nous notons que CoEM se comporte mieux que CoFKM, mais l'apport le plus significatif concerne l'utilisation de matrices noyaux cosinus, ce qui n'est pas disponible dans CoEM. Les résultats obtenus par CoKFKM sont sensiblements équivalents à ceux obtenus par concaténation avec un modèle FKM à noyau classique. La version accelérée de CoKFKM correspond au fait que dans la définition des distances aux centres (17) nous ne tenons pas comptes de tous les objets, mais seulement d'un pourcentage réduit pour réduire un peu la complexité. Par exemple, si l'on veut calculer $d_r(x_i, c_k)^2$, nous ne considérerons que les $nb = \frac{|X|}{K}$ objets

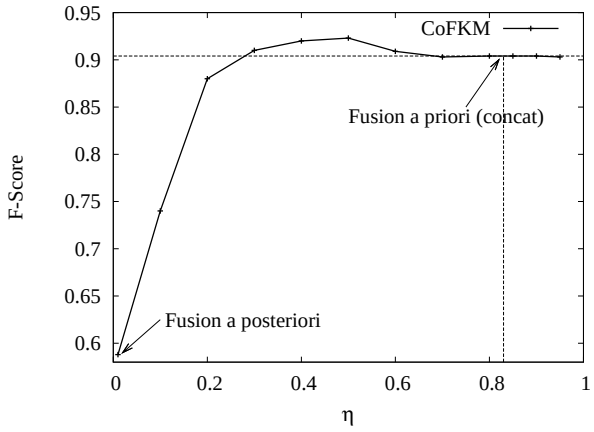

FIG. 5 – *CoFKM pour différentes valeurs de η.*

x_j les plus représentatifs du groupe k, *i.e* ceux correspondants aux nb valeurs $u_{j,k,r,\eta}$ les plus élevées. Nous constatons que cette accélération laisse entrevoir des perspectives sur l'extension à noyaux.

6 Conclusion et Perspectives

Nous avons centré cette étude sur la problématique de classification non-supervisée (ou clustering) sur des données complexes multi-vues. Nous avons alors présenté les différentes alternatives proposées dans la littérature et choisi de poursuivre l'amorce proposée par (Pedrycz, 2002) puis (Bickel et Scheffer, 2005) pour les méthodes de clustering collaboratif. Le modèle CoFKM que nous avons défini présente de bonnes propriétés puisqu'il généralise différentes solutions de fusion, permet de lui associer une solution algorithmique efficace (convergente) et se compose de peu de paramètres (moins sensible au paramétrage). Nous avons également proposé une variante à noyaux de notre modèle pour permettre le traitement à partir de plusieurs matrices de similarités. Nous présentons également une première comparaison expérimentale entre les méthodes multi-vues centralisées sur des données synthétiques et réelles. Les résultats viennent confirmer l'apport du modèle et de son extension à noyaux comparé aux approches existantes.

Parmi les orientations futures de ce travail nous envisageons de proposer d'autres formalisations du désaccord, de permettre de prendre en compte un nombre de groupes différents dans chaque vue, d'étudier le signe de ces termes relativement au nombre de représentations et de permettre un traitement semi-supervisé de ces données.

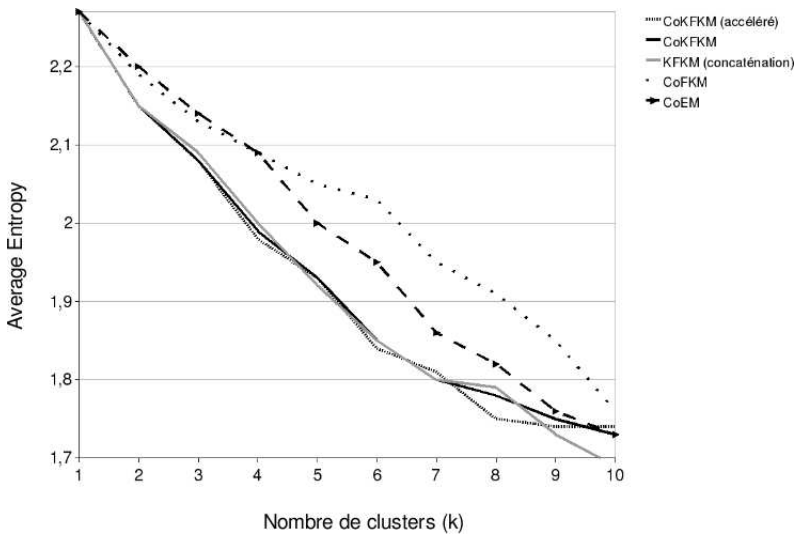

FIG. 6 – *Tests comparatifs entre CoKFKM, CoFKM et CoEM.*

Références

Bezdek, J. C. (1981). Pattern Recognition with Fuzzy Objective Function Algorithms. *Plenum Press, New York*.

Bickel, S. et T. Scheffer (2005). Estimation of mixture models using co-EM. In *16th European Conference on Machine Learning ECML 2001*, Volume 3720 of *Lecture Notes in Artificial Intelligence*, pp. 35–46. Springer.

Blum, A. et T. Mitchell (1998). Combining labeled and unlabeled data with co-training. In *COLT : Proceedings of the Workshop on Computational Learning Theory, Morgan Kaufmann Publishers.*

Cleuziou, G. et C. Poudat (2007). On the impact of lexical and linguistic features in genre and domsain-based text categorization. In *Proceedings of the Eighth International Conference on Intelligent Text Processing and Computational Linguistic*, Mexico City, Mexico, Australia, pp. 599–610.

Dempster, A., N. Laird, et D. Rubin (1977). Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of Royal Statistical Society B 39*, 1–38.

Forestier, G., C. Wemmert, et P. Gancarski (2008). Multi-source images analysis using collaborative clustering. *EURASIP Journal on Advances in Signal Processing - Special issue on Machine Learning in Image Processing 2008*.

Fred, A. (2001). Finding consistent clusters in data partitions. In *In Proc. 3d Int. Workshop on Multiple Classifier*, pp. 309–318. Springer.

Ganchev, K., J. Graça, J. Blitzer, et B. Taskar (2008). Multi-view learning over structured and non-identical outputs. In D. A. McAllester et P. Myllymäki (Eds.), *UAI*, pp. 88–96. AUAI

Press.

Ghani, R. (2002). Combining labeled and unlabeled data for multiclass text categorization. In *In Proceedings of the International Conference on Machine Learning*, pp. 187–194.

Heer, J. et E. H. Chi (2002). Mining the Structure of User Activity using Cluster Stability. In *proceedings of the Web Analytics Workshop, SIAM Conference on Data Mining*.

Jain, A. K., M. N. Murty, et P. J. Flynn (1999). Data clustering : a review. *ACM Computing Surveys 31*(3), 264–323.

Labute, P. (1998). Quasar-cluster : A different view of molecular clustering. *Chemical Computing Group, Inc.*.

Long, B., P. S. Yu, et Z. M. Zhang (2008). A general model for multiple view unsupervised learning. In *SDM*, pp. 822–833. SIAM.

Martin, C., H. grosse Deters, et T. W. Nattkemper (2006). Fusing biomedical multi-modal data for exploratory data analysis. In *ICANN 2006, Part II, LNCS 4132*, pp. 798–807.

Pedrycz, W. (2002). Collaborative fuzzy clustering. *Pattern Recogn. Lett. 23*(14), 1675–1686.

Reza, G., S. Md. Nasir, I. Hamidah, et M. Norwati (2009). A survey : Clustering ensembles techniques. *Proceedings of World Academy of Science, Engineering and Technology 38*, 644–653.

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

Yamanishi, Y., J. p. Vert, et M. Kanehisa (2004). Protein network inference from multiple genomic data : a supervised approach. *Bioinformatics 20*(1), i363–i370.

Summary

This paper deals with the clustering for multi-view data, *i.e.* objects described by several sets of variables or proximity matrices. Many important domains or applications such as Information Retrieval, biology, chemistry and marketing are concerned by this problematic. The aim of this data mining research field is to propose a theoretical and methodological framework allowing the search of clustering patterns that perform a consensus between the patterns from different views. This requires to merge information from each view by performing a fusion process that identifies the agreement between the views and solves the conflicts. Various fusion strategies can be applied, occurring either before, after or during the clustering process. We draw our inspiration from the existing algorithms based on a centralized strategy. We propose a fuzzy clustering approach that generalizes the three fusion strategies and present a kernel extension allowing a better applicability. We show that our approach outperforms the main existing multi-view clustering algorithm both on synthetic and real datasets.