Généralisation des k-moyennes pour produire des recouvrements ajustables

Chiheb-Eddine Ben N'Cir*, Guillaume Cleuziou**,***, Nadia Essoussi*

*LARODEC, ISG Tunis, Université de Tunis, Tunisie
chiheb.benncir@isg.rnu.tn
nadia.essoussi@isg.rnu.tn
**LIFO, Université d'Orléans, France
***GREYC, Université de Caen Basse-Normandie, France
guillaume.cleuziou@univ-orleans.fr

Résumé. La recherche de groupes non-disjoints à partir de données non-étiquetées est une problématique importante en classification non-supervisée. La classification recouvrante (Overlapping clustering) contribue à la résolution de plusieurs problèmes réels qui nécessitent la détermination de groupes qui se chevauchent. Cependant, bien que les recouvrements entre groupes soient tolérés voire encouragés dans ces applications, il convient de contrôler leur importance. Nous proposons dans ce papier des généralisations de k-moyennes offrant le contrôle et le paramétrage des recouvrements. Deux principes de régulation sont mis en place, ils visent à contrôler les recouvrements relativement à leur taille et à la dispersion des classes. Les expérimentations réalisées sur des jeux de données réelles, montrent l'intérêt des principes proposés.

1 Introduction

La classification non-supervisée est une tâche importante dans l'exploration de données non-étiquetées, elle vise à les organiser en groupes (ou classes) contenant des données similaires. Cette technique est utilisée avec succès dans de nombreux domaines d'application tels que le marketing et la recherche d'information. Cependant, dans plusieurs de ces applications, les données s'organisent naturellement en groupes non-disjoints nécessitant donc de l'émergence de groupes qui se chevauchent. Le domaine de recherche correspondant à cette problématique est la classification recouvrante (*overlapping clustering*), étudiée à travers différentes approches au cours du dernier demi-siècle (Shepard et Arabie, 1979; Diday, 1987; Banerjee et al., 2005; Cleuziou, 2008; Depril et al., 2008; Fellows et al., 2011).

Le clustering recouvrant trouve ses applications dans de nombreux domaines nécessitant qu'un individu appartienne à plusieurs classes. Par exemple, en analyse des réseaux sociaux, un acteur peut appartenir à plusieurs communautés (Tang et Liu, 2009; Wang et al., 2010; Fellows et al., 2011) ; en classification de vidéos, chaque entrée peut potentiellement avoir plusieurs genres différents (Snoek et al., 2006) ; en détection d'émotions, une pièce de musique peut engendrer plusieurs émotions (Wieczorkowska et al., 2006), dans les systèmes de recherche

d'information, un document peut aborder plusieurs thématiques (Gil-García et Pons-Porrata, 2010; Pérez-Suárez et al., 2013), etc.

Contrairement aux méthodes floues, la modélisation des recouvrements suppose que chaque observation peut avoir une appartenance totale à plusieurs groupes simultanément (sans recours à un degré d'appartenance). Quelle que soit l'approche utilisée, clustering hiérarchique ou partitionnement, les algorithmes existants produisent des groupes sans possibilité de contrôle sur la taille et la qualité des recouvrements. Bien que la méthode devrait idéalement révéler une classification qui convient le plus aux structures et formes sous-jacentes des données, un tel objectif n'est généralement pas unique et la nature des recouvrements doit être utilisée comme un paramètre dans le processus de classification. Nous proposons dans cette étude, partant de l'algorithme bien connu k-moyennes, deux nouvelles méthodes R_1-OKM et R_2-OKM permettant l'ajustement des recouvrements de clusters selon deux principes de régulation à savoir : le nombre et la dispersion des groupes concernés.

La suite de l'article est organisée ainsi : la Section 2 donne un bref aperçu des différentes approches de classification recouvrante et en particulier les algorithmes OKM et ALS. La Section 3 présente la motivation liée au contrôle des recouvrements ainsi qu'une description des deux principes de régulation que nous proposons pour ajuster les recouvrements. Les deux dernières sections 4 et 5 exposent respectivement les résultats expérimentaux obtenus puis les conclusions et les perspectives de l'étude.

2 Méthodes de classification recouvrante

La tâche de classification recouvrante a été étudiée et partiellement résolue durant les trente dernières années par une série d'études et de propositions de deux types : des solutions heuristiques ou théoriques. Nous appelons heuristiques les solutions qui consistent soit à modifier les sorties d'approches usuelles de clustering (e.g. k-moyennes ou k-moyennes flou) telles que proposé dans Lingras et West (2004) ou Zhang et al. (2007), soit à proposer un nouveau processus intuitif de construction de classes recouvrantes telles que l'algorithme CBC (*Clustering by Committee*) proposé par Pantel et Lin (2002) ou encore POBOC (*Pole-Based Overlapping Clustering*) proposé par Cleuziou et al. (2004). Ces deux types de contributions peuvent conduire à des résultats pertinents sans toutefois s'appuyer sur des modèles théoriques, limitant de fait la possibilité de les améliorer ou de les généraliser.

Les études théoriques sont, en revanche, des extensions de modèles usuels de classification non-supervisée, tels que les approches hiérarchiques, à base de modèles de mélanges ou de graphes. Les variantes recouvrantes des hiérarchies sont les pyramides (Diday, 1987) et plus généralement des hiérarchies faibles (Bertrand et Janowitz, 2003) ; elles visent à améliorer la correspondance entre l'indice de distance induit par la structure et la mesure de dissimilarité initiale. Cependant, les structures pseudo-hiérarchiques recouvrantes sont soient restrictives en terme de configuration des recouvrements, soient complexes à générer et à visualiser.

Plus récemment, les modèles de mélanges recouvrants ont été introduits (Banerjee et al., 2005; Heller et Ghahramani, 2007; Fu et Banerjee, 2008; Cleuziou et Sublemontier, 2008) ; ils sont motivés par la modélisation de processus biologiques et se fondent sur l'hypothèse que chaque observation est le résultat d'un mélange de lois et de combinaisons additives (Banerjee et al., 2005) ou multiplicatives (Heller et Ghahramani, 2007; Fu et Banerjee, 2008) de ces lois. Ce formalisme probabiliste permet de considérer non seulement des distributions gaussiennes

mais peut se généraliser à toute loi exponentielle ; en revanche, les modèles génératifs ne sont pas paramétrables et n'autorisent pas le contrôle des recouvrements notamment.

Nous concentrons notre étude sur un autre type de méthodes recouvrantes, celles formalisées par des critères objectifs et résolues de manière itérative. Deux types de modélisations des recouvrements ont été proposées et se réfèrent à deux hypothèses différentes :

- le **modèle additif** a été introduit initialement par Shepard et Arabie (1979), réutilisé par Mirkin (1987) et Depril et al. (2008) et formalisé en terme de modèle de mélange par Banerjee et al. (2005). Il se fonde sur l'hypothèse que les données situées à l'intersection de plusieurs clusters résultent d'une addition des caractéristiques de chacun des clusters ; les recouvrements sont alors modélisés par la somme des profils de ces clusters. Le modèle additif a été appliqué avec succès dans divers domaines tels que le marketing, l'expression de gènes et la psychologie pour lesquels il semble effectivement adapté de modéliser les données multi-classes par une combinaison additive des caractéristiques de chaque classe.
- le **modèle géométrique** a été introduit par Cleuziou (2008) puis réutilisé par BenN'Cir et al. (2010). Il modélise les recouvrements comme une combinaison barycentrique des profils de clusters. Ce modèle est basé sur un raisonnement géométrique dans l'espace (Euclidien) où les intersections de clusters correspondent à des recouvrements au sens spatial. Ce modèle a attesté expérimentalement sur des données textuelles et multi-média par exemple.

Nous détaillons dans la suite de cette section les algorithmes ALS (Depril et al., 2008) et OKM (Cleuziou, 2008) et leur modèle additif et géométrique respectivement.

Étant donnée une matrice $X = (x_1, \ldots, x_N)^T$ définissant N observations dans $\mathbb{R}^M$, l'algorithme ALS (*Alternating Least Square*) consiste à minimiser la fonction objective suivante :

$$J_{ALS}(A, P) = \sum_{i=1}^{N} \sum_{j=1}^{M} \left(x_{i,j} - \sum_{k=1}^{K} a_{i,k} p_{k,j} \right)^2 , \tag{1}$$

avec K le nombre de composantes (clusters) attendues, A une matrice binaire $(N \times K)$ représentant les affectations ($a_{i,k} = 1$ ssi x_i appartient au k^{eme} cluster) et P une matrice réelle $(K \times M)$ définissant les K profils de clusters dans $\mathbb{R}^M$. En se basant sur le modèle additif, ALS optimise une somme d'erreurs locales, propres à chaque observation x_i, et définies par la distance Euclidienne entre l'observation et la somme des profils des clusters auxquels x_i appartient.

De même, la fonction objective utilisée dans OKM (*Overlapping K-Means*) est basée sur les erreurs locales, mais diffère dans la manière de combiner les profils de clusters. En se basant sur un modèle géométrique, OKM évalue l'erreur locale à chaque donnée x_i par la moyenne (barycentre) des profils :

$$J_{OKM}(A, P) = \sum_{i=1}^{N} \sum_{j=1}^{M} \left(x_{i,j} - \frac{\sum_{k=1}^{K} a_{i,k} p_{k,j}}{\sum_{k=1}^{K} a_{i,k}} \right)^2 . \tag{2}$$

Les deux critères objectifs peuvent également s'exprimer sous forme matricielle comme suit :

$$J_{ALS} = \|X - AP\|_F^2 \tag{3}$$

$$J_{OKM} = \|X - SP\|_F^2, \tag{4}$$

avec $\|.\|_F$ la norme de Frobenius, et S définie pour chaque entrée par $s_{i,k} = \frac{a_{i,k}}{\sum_l a_{i,l}}$. La minimisation des critères objectifs (3) et (4) est réalisée par itération classique de deux étapes, décrites dans l'Algorithme générique 1 :

1. L'étape d'affectation correspond à un problème d'optimisation discret, résolu dans ALS par l'évaluation de toutes les affectations possibles 2^K pour chaque observation x_i ou bien par la relaxation du problème en utilisant une heuristique telle que dans OKM.
2. L'étape de mise à jour des profils peut être réalisée de manière optimale en utilisant la pseudo-inverse de A ou S tel que proposé dans ALS ou bien à travers une mise à jour successive des profils pour éviter l'inversion de la matrice qui est généralement coûteuses (utilisé dans OKM).

Algorithm 1 *Réallocations dynamiques pour la classification recouvrante*

ENTRÉE X : ensemble de données dans $\mathbb{R}^M$.
$J(.)$: une fonction objective.
K : nombre de clusters attendu.
SORTIE A : matrice binaire d'affectation et P : matrice des profils de clusters
Initialiser les profils P
répéter
1: Calculer les nouvelles affectations A sachant P
2: Calculer les nouveaux profils P sachant A
Tant que $J(A, P)$ diminue
Retourner les affectations finales A.

Par la suite, le problème de la régulation des recouvrements est considéré pour le modèle géométrique sans perte de généralité puisque les principes de régulation proposés peuvent être appliqués aux modèles additifs.

3 Ajustement des recouvrements

Dans un processus d'extraction de connaissances, l'utilisateur ou l'expert préfère généralement avoir les moyens d'interagir avec le système afin d'explorer plusieurs alternatives concernant le nombre de clusters ou la métrique ou encore afin de contrôler le degré de "flou" du clustering. De même, dans un processus de recherche de classes recouvrantes, l'expert préfèrera un système lui permettant d'ajuster l'importance des recouvrements relativement à ses attentes et à ses connaissances sur les données.

Pour rendre possible la régulation des recouvrements pour le modèle géométrique nous formalisons deux principes de régulation qui sont basés respectivement sur le nombre de clusters

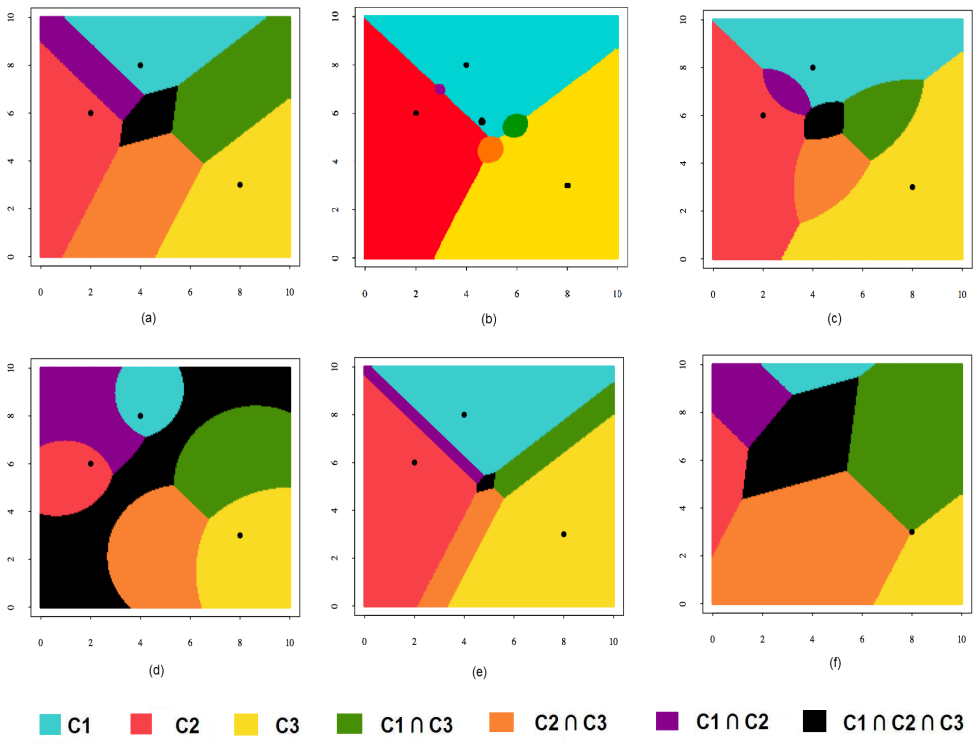

FIG. 1 – *Exemple en 2 dimensions de cellules de Voronoï résultant du modèle* OKM *(a), du nouveau modèle* R_1*-*OKM *avec* $\alpha = 5$ *(b),* $\alpha = 1$ *(c) et* $\alpha = -1$ *(d) et du nouveau modèle* R_2*-*OKM *avec* $\lambda = 2$ *(e) et* $\lambda = -0.5$ *(f).*

concernés et la dispersion des profils des groupes concernés. Pour visualiser ces principes de régulation, nous visualisons les zones de recouvrements avec des cellules de Voronoï dans un espace à deux dimensions dans un contexte de trois clusters. La Figure 1.(a) montre les cellules de Voronoï construites avec le modèle OKM en utilisant trois profils de classes de coordonnées respectives $(4, 8)$, $(2, 6)$ et $(8, 3)$. Chaque zone de couleur est une cellule de Voronoï qui représente une classe ou une intersection possible (recouvrement) de classes. Par exemple toutes les données situées dans la zone en *jaune* seront affectées uniquement à la classe 3 avec le modèle OKM, tandis que les données situées dans la zone en *vert* seront attribuées à l'intersection des classes 1 et 3 ($bleu \wedge jaune \rightarrow vert$) et la zone en *noir* illustre l'intersection des trois classes.

L'idée directrice de notre étude est donc de gérer différemment les classes et les combinaisons de classes de manière à limiter ou favoriser les affectations multiples aux situations où l'amélioration induite est réellement significative.

3.1 Régulation des recouvrements par le nombre de clusters (R_1-OKM)

Nous proposons une première modélisation consistant à introduire une pondération, ajustable avec un paramètre α, visant à réguler chaque erreur locale relativement au nombre d'affectations de l'individu. Le nouveau modèle s'exprime par le critère objectif suivant :

$$J_{R_1-OKM}(A,P) = \sum_i \left[\sum_k a_{i,k}\right]^{\alpha} \sum_j \left(x_{i,j} - \frac{\sum_k a_{i,k} p_{k,j}}{\sum_k a_{i,k}}\right)^2, \tag{5}$$

avec $\sum_k a_{i,k}$ le nombre de classes auxquelles l'objet x_i appartient et $\alpha \in \mathbb{R}$ un paramètre contrôlant les recouvrements :

1. $\alpha = 0$ annule la pondération sur les tailles de combinaison conduisant ainsi au modèle OKM originel,
2. $\alpha > 0$ pénalise les affectations à des combinaisons larges tant que α augmente jusqu'à aboutir à un modèle équivalent à k-moyennes ($\alpha \to +\infty$),
3. $\alpha < 0$ favorise les recouvrements tant que α diminue jusqu'à aboutir à un modèle totalement recouvrant avec des données affectées à toute les classes ($\alpha \to -\infty$).

Les Figures 1.(b), 1.(c) et 1.(d) illustrent les nouvelles cellules de Voronoï résultant du modèle R_1-OKM avec $\alpha = 5$, $\alpha = 1$ et $\alpha = -1$ respectivement. Ces figures montrent visuellement que les zones de recouvrements sont plus larges avec $\alpha < 0$ et sont de plus en plus réduites avec $\alpha > 0$.

3.2 Régulation des recouvrements par la dispersion des profils (R_2-OKM)

Contrairement au modèle de régulation précédent qui se base sur le nombre de combinaisons de classes, le second modèle se base sur la dispersion des profils de classes afin de contrôler l'importance des recouvrements. L'hypothèses sous-jacente au modèle R_2-OKM est que les recouvrements doivent être d'autant plus pénalisés qu'ils mettent en jeu des classes distantes les unes des autres. Inversement, les recouvrements sont autorisés (voire encouragés) pour les classes dont les profils sont plus proches.

Afin de formaliser ce principe de régulation, nous proposons de favoriser ou pénaliser l'erreur locale à chacune des données x_i en utilisant le critère de dispersion suivant, quantifiant (le carré de) la distance moyenne de x_i avec ses profils de classes :

$$\frac{\sum_k a_{i,k} \sum_j (x_{i,j} - p_{k,j})^2}{\sum_k a_{i,k}}. \tag{6}$$

Le nouveau critère objectif du modèle R_2-OKM se présente alors comme suit :

$$J_{R_2-OKM}(A,P) = \sum_i \sum_j \left[\left(x_{i,j} - \frac{\sum_k a_{i,k} p_{k,j}}{\sum_k a_{i,k}} \right)^2 + \lambda \frac{\sum_k a_{i,k} (x_{i,j} - p_{k,j})^2}{\sum_k a_{i,k}} \right], \tag{7}$$

avec λ un paramètre permettant de contrôler le critère de dispersion et jouant un rôle similaire à α dans le premier modèle : $\lambda = 0$ annule le principe de régulation conduisant à un modèle identique à OKM, $\lambda > 0$ limite les recouvrements alors que $\lambda < 0$ les favorise comme l'illustrent les Figures 1.(e) et 1.(f) pour des valeurs de λ égales à 2 et -0.5 respectivement.

3.3 Processus d'optimisation

L'algorithme d'optimisation pour R_1-OKM et R_2-OKM suit le processus général de réaffectations décrit dans l'Algorithme 1. La stratégie d'affectation du modèle OKM (Cleuziou, 2008) demeure valide, mais en utilisant les nouveaux critères objectifs J_{R_1-OKM} et J_{R_2-OKM}.

La mise à jour des profils est effectuée successivement pour chaque cluster et nécessite la dérivation de nouveaux profils $P^*_{k,.}$ garantissant la convergence des modèles proposés. Étant donnée la matrice A des affectations, les profils optimaux pour R_1-OKM et R_2-OKM sont obtenus par dérivation des critères (5) et (7) respectivement :

$$p^*_{k,j} = \frac{\sum_i a_{i,k} \left(\sum_l a_{i,l}\right)^{\alpha-2} p^i_{k,j}}{\sum_i a_{i,k} \left(\sum_l a_{i,l}\right)^{\alpha-2}} \tag{8}$$

$$p^*_{k,j} = \frac{\sum_i a_{i,k} \left[\left(\sum_l a_{i,l}\right)^{-2} p^i_{k,j} + \lambda \left(\sum_l a_{i,l}\right)^{-1} x_{i,j}\right]}{\sum_i a_{i,k} \left[\left(\sum_l a_{i,l}\right)^{-2} + \lambda \left(\sum_l a_{i,l}\right)^{-1}\right]}, \tag{9}$$

avec P^i_k le profil de la classe P_k idéal vis-à-vis de l'individu x_i, c'est-à-dire tel que l'erreur locale pour x_i est égale à zéro : $p^i_{k,j} = x_{i,j} \sum_l a_{i,l} - \sum_{l \neq k} a_{i,l} p_{l,j}$.

Nous avons proposé et formalisé deux principes de régulation des recouvrements dans le cadre des modèles géométriques ainsi que leur mise en oeuvre algorithmique. Notons que R_1-OKM et R_2-OKM sont deux généralisations de l'algorithme k-moyennes : si les affectations sont restreintes à un seul groupe ou si les paramètres (α et λ) sont suffisamment pénalisant, le critère objectif est équivalent au critère des moindres carrés utilisé par k-moyennes. La convergence des méthodes proposées est assurée par un algorithme à faible coût en terme de complexité : $O(TNKlogK)$ où T est le nombre d'itérations.

4 Expérimentation

L'évaluation des méthodes de classification non-supervisée est connue pour être une tâche difficile puisqu'il n'existe pas, par définition, de classification exacte sur laquelle se fonder pour mesurer la qualité d'un clustering. La plupart des méthodes de classification recouvrantes ont été évaluées en utilisant la "F-mesure" combinant précision et rappel sur les paires d'individus (Banerjee et al., 2005; Cleuziou, 2008; Fu et Banerjee, 2008). Cependant, dans le contexte recouvrant, la "F-mesure" ignore la multiplicité des affectations. Amigó et al. (2009) ont proposé une extension de leur métrique "BCubed" à la classification recouvrante qui permet d'affiner le calcul des mesures de *précision* et de *rappel* en intégrant cette multiplicité. De la même manière que Suárez et al. (2013), nous avons décidé d'utiliser la mesure ajustée "F-BCubed" pour une évaluation plus fine des classifications recouvrantes.

TAB. 1 – *Statistiques des jeux de données utilisés.*

Jeu de données	Domaine	Nb. observations	Dimensionalité	Nb. labels	Taux recouv.
Scene	Image	2407	294	6	1.07
EachMovie	Video	75	3	3	1.14
Music emotion	Musique	593	72	6	1.86
Yeast	Biologie	2417	103	14	4.23

Nous avons mené des expérimentations sur différents domaines qui motivent la recherche de classifications recouvrantes : classification de vidéos (Eachmovie [1]), détection d'émotions dans des extraits musicaux (Music Emotion [2]), classification de paysages naturels (Scene [3]) et clustering de gènes (Yeast [4]). Pour chaque jeu de données, une classification de référence est proposée, avec une ou plusieurs étiquettes de classe pour chaque individu. Ces benchmarks ont été sélectionnés en raison de la diversité de leur domaine d'application, de leur taille (de 75 à 2417), de leur dimensionnalité (de 3 à 294), de leur nombre de classes (de 3 à 14) et de leur taux de recouvrement (de 1 à 4.23). Le taux de recouvrement correspond au nombre moyen de labels par individu ; dans les schémas générés par la suite, ce taux correspondra au nombre moyen d'affectations par individu. Le Tableau 1 présente la description statistique de jeux de données. Les classifications de référence sont donc recouvrantes et nous cherchons à savoir dans quelle mesure les nouveaux modèles proposés permettent de retrouver des organisations naturelles qui s'apparenteraient à ces classifications de références.

Dans nos expérimentations, nous avons confronté les deux méthodes de régulation (R_1-OKM et R_2-OKM) avec cinq algorithmes existants : MOC et ALS ayant le même modèle additif mais utilisant différentes stratégies d'optimisation ; OKM ayant un modèle géométrique ; k-moyennes flou seuillé et k-moyennes simple. Les méthodes MOC, ALS, OKM et k-moyennes ne proposent pas de paramètre de contrôle des recouvrements tandis que les trois autres algorithmes sont paramétrés par α (pour R_1-OKM), λ (pour R_2-OKM) et σ (seuil utilisé dans k-moyennes flou) :

- pour α nous avons détecté par dichotomie dans l'intervalle $[\![0; 10]\!]$ la plus petite valeur α_{min} menant à une classification non-recouvrante, puis nous avons testé une centaine de valeurs réparties uniformément sur $[0, \alpha_{min}]$ et gardé trois valeurs correspondant à des sauts significatifs dans les taux de recouvrement.
- pour λ, sept valeurs ont été considérées : $(0, 0.125, 0.25, 0.5, 1, 2, 5)$.
- enfin pour σ, le seuil d'affectation sur les degrés d'appartenance a été réglé manuellement de manière à explorer une plage de valeurs, bornée par les situations extrêmes (sans recouvrements vs. totalement recouvrantes).

1. cf.http ://www.grouplens.org/node/76.
2. cf.http ://mlkd.csd.auth.gr/multilabel.html
3. cf.http ://mlkd.csd.auth.gr/multilabel.html
4. cf.http ://mlkd.csd.auth.gr/multilabel.html

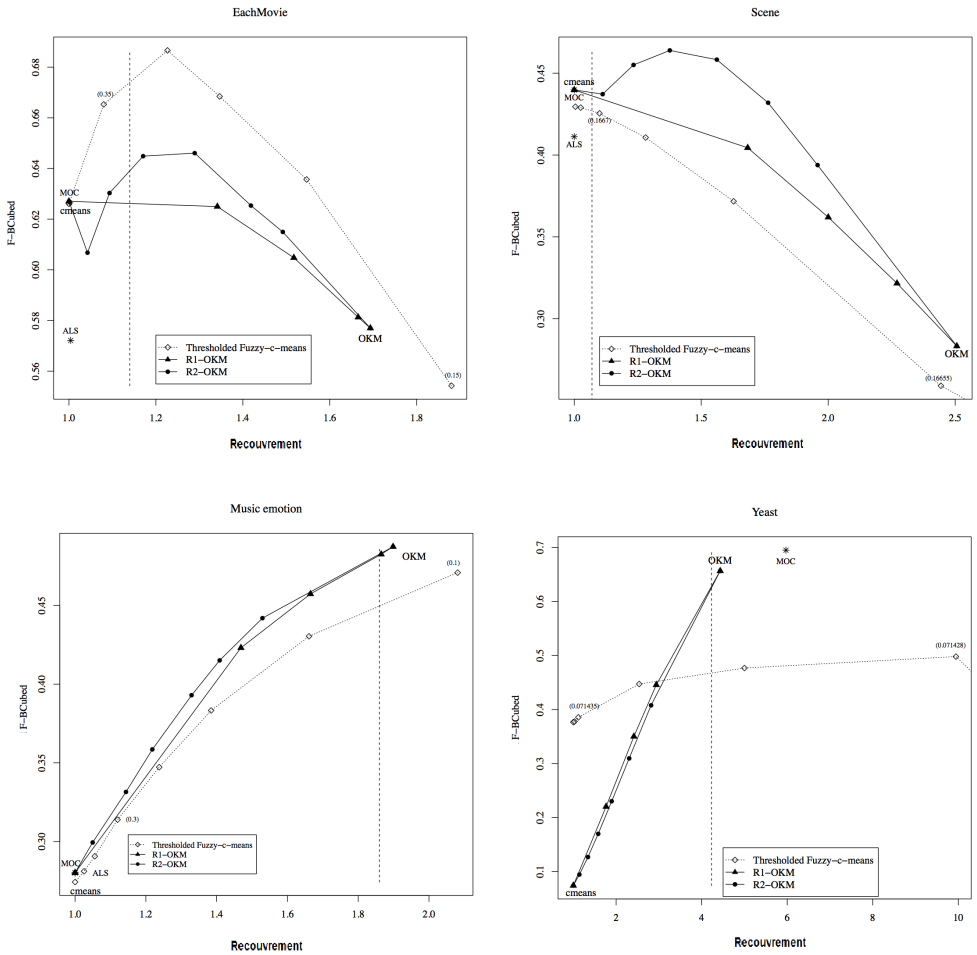

FIG. 2 – *Positionnement relatif des nouveaux modèles de régulation des recouvrements par rapport aux méthodes existantes sur 4 jeux de données.*

La taille des recouvrements a une grande influence sur la qualité des mesures de performance et sur la structure des classes attendues. Ainsi, plutôt que de fournir des tableaux de valeurs, nous avons positionné les scores (F-BCubed) de chaque méthode relativement au taux de recouvrement. La Figure 2 présente ces positionnements : chaque point sur la figure est obtenu par une moyenne sur dix exécutions de chaque algorithme dans les mêmes conditions initiales (initialisation des profils de clusters) et pour un nombre de classes égal au nombre d'étiquettes de la référence.

Les méthodes MOC, ALS et OKM étant sans contrôle possible sur les recouvrements, un unique clustering est produit pour chaque initialisation, ce qui se traduit par un seul point moyen sur les figures. En revanche, pour les trois autres algorithmes, les scores obtenus pour

des paramétrages consécutifs ont été reliés afin d'observer les tendances. Les lignes verticales en pointillés indiquent le taux de recouvrement de la classification de référence.

Les principales observations que nous pouvons relever de ces résultats sont :

- que les modèles additifs (MOC et ALS) ne parviennent pas à construire de recouvrements entre les classes pour EachMovie, Music emotion et Scene. Par contre, sur des données biologiques (Yeast), ces méthodes se caractérisent par une bonne qualité de classification et parviennent même à construire de larges recouvrements.
- la capacité des principes de régulation proposés à produire des recouvrements adaptés à la structure sous-jacente des données : sur EachMovie et Scene, caractérisés par de faibles recouvrements, les méthodes existantes produisent des recouvrements larges conduisant à affaiblir leur score ; l'ajustement des recouvrements offre un moyen d'améliorer significativement la qualité des classifications générées.
- les performances de k-moyennes flou seuillé sont rapidement limitées lorsque le nombre de clusters augmente. Cette méthode ne parviens pas à égaler les résultats de R_1- ou R_2-OKM sur les jeux de données Music emotion, Scene et Yeast avec 6, 6 et 14 clusters respectivement.

5 Conclusion

Nous avons proposé dans cette étude deux modèles généralisant k-moyennes afin de produire des classifications recouvrantes avec un contrôle sur la taille des recouvrements. Ces deux modèles offrent une régulation par le nombre et la dispersion des profils des clusters concernés, conduisant à des schémas de classification plus appropriés.

Pour compléter cette étude nous envisageons, dans une version plus étendue, de proposer une analyse détaillée ainsi qu'une évaluation experte sur un jeu de données réel lié au domaine de la recherche d'information. Les nouveaux principes de régulation des recouvrements, introduits dans ce présent travail, ne sont pas limités aux modèles géométriques et pourront être transposés aux modèles additifs (type ALS).

Références

Amigó, E., J. Gonzalo, J. Artiles, et F. Verdejo (2009). A comparison of extrinsic clustering evaluation metrics based on formal constraints. *Inf. Retr. 12*(5), 613.

Banerjee, A., C. Krumpelman, S. Basu, R. J. Mooney, et J. Ghosh (2005). Model based overlapping clustering. In *International Conference on Knowledge Discovery and Data Mining*, Chicago, USA, pp. 532–537. SciTePress.

BenN'Cir, C., N. Essoussi, et P. Bertrand (2010). Kernel overlapping k-means for clustering in feature space. In *International Conference on Knowledge discovery and Information Retrieval KDIR*, Valencia, SPA, pp. 250–256. SciTePress Digital Library.

Bertrand, P. et M. F. Janowitz (2003). The k-weak hierarchical representations : an extension of the indexed closed weak hierarchies. *Discrete Applied Mathematics 127*(2), 199–220.

Cleuziou, G. (2008). An extended version of the k-means method for overlapping clustering. In *International Conference on Pattern Recognition ICPR*, Florida, USA, pp. 1–4. IEEE.

Cleuziou, G., L. Martin, et C. Vrain (2004). PoBOC : an Overlapping Clustering Algorithm. Application to Rule-Based Classification and Textual Data. In R. López de Mántaras and L. Saitta, IOS Press (Ed.), *Proceedings of the 16th European Conf. on Artificial Intelligence*, Valencia, Spain, pp. 440–444.

Cleuziou, G. et J.-H. Sublemontier (2008). Étude comparative de deux approches de classification recouvrante : Moc vs. okm. In *EGC*, pp. 667–678.

Depril, D., I. Van Mechelen, et B. Mirkin (2008). Algorithms for additive clustering of rectangular data tables. *Computational Statistics and Data Analysis 52*(11), 4923–4938.

Diday, E. (1987). Orders and overlapping clusters by pyramids. Technical Report 730, INRIA, France.

Fellows, M. R., J. Guo, C. Komusiewicz, R. Niedermeier, et J. Uhlmann (2011). Graph-based data clustering with overlaps. *Discrete Optimization 8*(1), 2–17.

Fu, Q. et A. Banerjee (2008). Multiplicative mixture models for overlapping clustering. In *Proceedings of the 8th IEEE International Conference on Data Mining*, Washington, DC, USA, pp. 791–796.

Gil-García, R. et A. Pons-Porrata (2010). Dynamic hierarchical algorithms for document clustering. *Pattern Recogn. Lett. 31*(6), 469–477.

Heller, K. et Z. Ghahramani (2007). A nonparametric bayesian approach to modeling overlapping clusters. *Journal of Machine Learning Research 2*, 187–194.

Lingras, P. et C. West (2004). Interval set clustering of web users with rough k-means. *J. Intell. Inf. Syst. 23*(1), 5–16.

Mirkin, B. G. (1987). Method of principal cluster analysis. *Automation and Remote Control 48*, 1379–1386.

Pantel, P. et D. Lin (2002). Discovering word senses from text. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, Edmonton, Alberta, Canada, pp. 613–619. ACM Press.

Pérez-Suárez, A., J. F. Martínez-Trinidad, J. A. Carrasco-Ochoa, et J. E. Medina-Pagola (2013). An algorithm based on density and compactness for dynamic overlapping clustering. *Pattern Recognition 46*(11), 3040–3055.

Shepard, R. N. et P. Arabie (1979). Additive clustering - representation of similarities as combinations of discrete overlapping properties. *Psychol. Rev. 86*(2), 87–123.

Snoek, C. G. M., M. Worring, J. C. van Gemert, J.-M. Geusebroek, et A. W. M. Smeulders (2006). The challenge problem for automated detection of 101 semantic concepts in multimedia. In *Proceedings of the 14th annual ACM international conference on Multimedia*, MULTIMEDIA '06, New York, USA, pp. 421–430. ACM.

Suárez, A. P., J. F. M. Trinidad, J. A. Carrasco-Ochoa, et J. E. Medina-Pagola (2013). An algorithm based on density and compactness for dynamic overlapping clustering. *Pattern Recognition 46*(11), 3040–3055.

Tang, L. et H. Liu (2009). Scalable learning of collective behavior based on sparse social dimensions. In *Proceedings of the 18th ACM conference on Information and knowledge management*, pp. 1107–1116.

Wang, X., L. Tang, H. Gao, et H. Liu (2010). Discovering overlapping groups in social media. In *Proceedings of the 2010 IEEE International Conference on Data Mining*, pp. 569–578.

Wieczorkowska, A., P. Synak, et Z. Ras (2006). Multi-label classification of emotions in music. In *Intelligent Information Processing and Web Mining*, Volume 35 of *Advances in Soft Computing*, pp. 307–315.

Zhang, S., R.-S. Wang, et X.-S. Zhang (2007). Identification of overlapping community structure in complex networks using fuzzy c-means clustering. *Physica A : Statistical Mechanics and its Applications 374*(1), 483–490.

Summary

Looking for non-disjoint groups from unlabeled data is an important issue in clustering referred as Overlapping Clustering. The resolution of this issue contributes to solve many real problems requiring the determination of overlapping groups. However, although overlaps between groups are tolerated in such applications, it is necessary to control their importance in order to fit the true structures of such data. We propose two new models, based on k-means, for controlling and setting the overlaps. Two principles of regulation are proposed and aim to control overlaps as regard to the number and the dispersal of the cluster concerned. Experiments performed on different multi-labeled benchmarks show the effectiveness of the proposed principles.