OKMED et WOKM : deux variantes de OKM pour la classification recouvrante

Guillaume Cleuziou

Laboratoire d'Informatique Fondamentale d'Orléans (LIFO)
Université d'Orléans
Rue Léonard de Vinci - 45067 Orléans Cedex 2
Guillaume.Cleuziou@univ-orleans.fr

Résumé. Cet article traite de la problématique de la classification recouvrante (overlapping clustering) et propose deux variantes de l'approche OKM : OKMED et WOKM. OKMED généralise k-médoïdes au cas recouvrant, il permet d'organiser un ensemble d'individus en classes non-disjointes, à partir d'une matrice de distances. La méthode WOKM (Weighted-OKM) étend OKM par une pondération locale des classes ; cette variante autorise chaque individu à appartenir à plusieurs classes sur la base de critères différents. Des expérimentations sont réalisées sur une application cible : la classification de textes. Nous montrons alors que OKMED présente un comportement similaire à OKM pour la métrique euclidienne, et offre la possibilité d'utiliser des métriques plus adaptées et d'obtenir de meilleures performances. Enfin, les résultats obtenus avec WOKM montrent un apport significatif de la pondération locale des classes.

1 Introduction

La classification recouvrante (ou *overlapping clustering*) constitue une problématique particulière dans le domaine de la classification non-supervisée (ou *clustering*). Il s'agit d'organiser un ensemble d'individus en classes d'individus similaires en autorisant chaque donnée à appartenir à plusieurs classes. Ce type de schéma correspond à une organisation naturelle des données pour de nombreuses applications. Par exemple, en Recherche d'Information un même document peut porter sur une ou plusieurs thématiques, en Bioinformatique un même gène peut intervenir dans un ou plusieurs processus métaboliques, en Traitement du Langage un même verbe peut satisfaire une ou plusieurs grammaires de sous-catégorisation, etc.

On parle de "problématique" au même titre que la problématique générale de la classification, puisqu'il n'existe pas d'avantage de solution triviale pour extraire des classes d'individus similaires qui soient indiscutables et universelles. De surcroît, la classification recouvrante offre un espace de solutions plus vaste que dans le cas traditionnel, qu'il est donc encore plus difficile d'explorer.

Durant les quatre dernières décennies, quelques solutions ont été proposées spécifiquement pour la classification recouvrante. Dattola (1968) envisagait une approche de type centres mobiles avec affectation multiple des individus déterminée par un seuil. Jardine et Sibson (1971), en introduisant les k-ultramétriques, ont ouvert la voie des recherches fondamentales sur les

hiérarchies recouvrantes : pyramides (Diday (1987)) ou hiérarchies faibles (Bertrand et Janowitz (2003)). Plus récemment, sous la pression des applications en Recherche d'Information ou en Bioinformatique, de nouvelles investigations ont été menées afin d'étendre les modèles de partitionnement (k-moyennes ou CEM) aux cas recouvrants. Ainsi Banerjee et al. (2005) ont proposé le modèle MOC qui généralise CEM et Cleuziou (2008) le modèle OKM qui généralise k-moyennes. Ces deux derniers modèles sont naturellement très proches et diffèrent principalement dans la définition des intersections entre classes et dans la mise en œuvre algorithmique proposée (initialisation et méthode d'affectation en particulier). Une étude comparative plus complète a été proposée par Cleuziou et Sublemontier (2008).

Le modèle commun à OKM et MOC permet d'envisager un large éventail de pistes à explorer tellement les déclinaisons de k-moyennes sont nombreuses. Par exemple, des variantes de k-moyennes ont été proposées pour rechercher le nombre k de classes approprié (D.Pelleg et Moore (2000)), pour limiter le risque d'une solution localement optimale (Likas et al. (2003)) ou encore pour initialiser l'algorithme de façon intelligente (Peña et al. (1999)). Dans la présente étude nous avons choisi d'étudier deux extensions particulières du modèle OKM pour répondre aux priorités dans ce domaine, à savoir la nécessité de diversifier les métriques acceptables par le modèle d'une part et la possibilité d'affecter un même individus à plusieurs classes sur la base de caractéristiques différentes d'autre part.

Nous proposons alors une première variante (OKMED) qui se fonde sur les méthodes de partitionnement autour de médoïdes et qui permet d'organiser un ensemble d'individus décrits par une matrice de distance quelconque en classes recouvrantes d'individus similaires. OKMED nécessite de définir judicieusement la notion de représentant d'une intersection et pose quelques problèmes de complexité théoriques qui peuvent aisément être contournés en pratique. La deuxième contribution de l'étude est la variante pondérée (WOKM) qui vient généraliser le modèle OKM en introduisant une pondération des attributs, locale à chaque classe. Cette approche s'inspire de la version pondérée de k-moyennes proposée par Chan et al. (2004) et plus fondamentalement des distances adaptatives de Diday et Govaert (1977). WOKM semble particulièrement adapté à la classification recouvrante : chaque classe étant "caractérisée" par une pondération différente des attributs, la description d'un individu est considérée différemment d'une classe à une autre et un même individu peut donc naturellement appartenir à plusieurs classes sur la base d'attributs éventuellement différents. Nous montrerons que la transposition du modèle initial pondéré au cas recouvrant n'est pas trivial, nous proposerons des solutions algorithmiques permettant d'assurer la convergences des algorithmes et montrerons l'efficacité de ces choix sur des données réelles.

L'article s'organise en quatre principales sections : la section 2 rappelle le cadre formel des modèles OKM et MOC afin de mieux appréhender les deux sections suivantes qui présentent respectivement les variantes OKMED et WOKM. Avant de conclure, la section 5 présente les expérimentations réalisées sur des données réelles de classification de textes.

2 Cadre formel des modèles MOC et OKM

Le modèle MOC proposé par Banerjee et al. (2005) et le modèle OKM proposé par Cleuziou (2008) reposent tous les deux sur une extension des méthodes d'agrégation autour de centres mobiles au domaine de la classification recouvrante. MOC est initialement formalisé en terme de modèles de mélanges (recouvrants). Cependant, l'optimisation du critère objectif (log-vraisemblance) nécessite dans le cas recouvrant de restreindre le modèle génératif (va-

riances constantes et toutes égales) ainsi que l'algorithme de résolution (CEM plutôt que EM). De facto, MOC peut être vu comme une méthode plus classique d'optimisation d'un critère d'inertie de type moindres carrés.

Soit $\mathcal{X} = \{x_i\}_{i=1}^n$ un ensemble d'individus dans $\mathbb{R}^p$, la fonction objective des modèles MOC et OKM peut s'exprimer de manière unifiée par :

$$\mathcal{J}(\{\pi_c\}_{c=1}^k) = \sum_{x_i \in \mathcal{X}} \|x_i - \phi(x_i)\|^2 \quad (1)$$

Dans ce critère, les $\{\pi_c\}_{c=1}^k$ désignent les k classes recouvrantes et $\phi(x_i)$ le représentant de x_i dans le schéma de classification, appelé "image" de x_i par Cleuziou (2007) . Cette image est déterminée par une combinaison des centres $\{m_c\}_{c=1}^k$ des classes auxquelles x_i appartient : une somme dans le modèle MOC et une moyenne dans le modèle OKM. Soit $A_i = \{m_c | x_i \in \pi_c\}$ l'ensemble des centres de classes d'appartenance de l'individu x_i

$$\phi_{MOC}(x_i) = \sum_{m_c \in A_i} m_c \quad ; \quad \phi_{OKM}(x_i) = \frac{\sum_{m_c \in A_i} m_c}{|A_i|} \quad (2)$$

Ainsi défini, le critère objectif (1) suggère deux remarques :

- Tout d'abord on notera que ce critère doit être vu comme un critère d'inertie au même titre que le critère des moindres carrés utilisé dans k-moyennes ; en effet la fonction objective $\mathcal{J}$ exprime l'inertie des individus $\{x_i\}_{i=1}^n$ par rapport à leur image $\{\phi(x_i)\}_{i=1}^n$ dans la classification.
- Ensuite on observe que dans le cas de partitions, chaque individu ne possède qu'une seule classe d'appartenance ($\forall i,\ |A_i| = 1$) ; pour les deux modèles l'image $\phi(x_i)$ de x_i correspond au centre m_c de la classe d'appartenance de l'individu et la fonction objective prend la forme du critère classique des moindres carrés (somme des distances au centre) ; à ce titre MOC et OKM généralisent k-moyennes.

L'optimisation[1] du critère d'inertie (1) est réalisée en itérant les deux étapes usuelles : calcul des paramètres des classes (ici les centres $\{m_c\}_{c=1}^k$) puis affectation de chaque individu à une ou plusieurs classes. MOC et OKM proposent des heuristiques différentes pour l'initialisation des paramètres et pour l'étape d'affectation multiple qui pose un problème combinatoire.

3 OKMED comme généralisation des k-médoïdes

3.1 Motivation de l'approche et méthodes à base de médoïdes

Les méthodes de classification de type k-médoïdes consistent à agréger les individus autour de représentants de classes choisis parmi les individus eux-mêmes ; on les appelle des *médoïdes* par opposition aux traditionnels *centroïdes* de classe qui sont définis dans l'espace de description des individus mais n'appartiennent pas nécessairement à $\mathcal{X}$.

La méthode PAM (*Partitioning Around Medoids*) proposée par Kaufman et Rousseeuw (1987) est communément admise comme l'algorithme de référence dans ce domaine. PAM construit un partitionnement des individus par itération de deux étapes : affectation de chaque individu au médoïde le plus proche puis mise à jour des médoïdes pour chaque classe.

[1] L'initialisation des paramètres induit une recherche locale et le risque habituel d'aboutir à un optimum local.

Dans la seconde étape, la mise à jour du médoïde d'une classe consiste à rechercher parmi tous les individus de la classe celui qui minimise la somme des distances avec tous les autres individus de la classe.

Les deux principaux avantages de ces méthodes sont d'une part leur robustesse face aux individus atypiques (*outliers*) et d'autre part la possibilité qu'elles offrent d'utiliser diverses métriques puisqu'elles nécessitent uniquement la matrice des distances entre individus en entrée ; c'est précisément ce dernier point qui motive la présente étude. En effet, les modèles recouvrants de base MOC et OKM se limitent pour le moment à une famille réduite de métriques (les divergences de Bregman) et l'extension à d'autres mesures n'est pas triviale.

3.2 Le modèle OKMED

Nous proposons de conserver pour le modèle OKMED le critère objectif (1) du modèle originel OKM, généralisé cette fois à une distance quelconque entre individus. Soient $\mathcal{X} = \{x_i\}_{i=1}^n$ un ensemble d'individus et d une mesure de distance de $\mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}^+$, le critère objectif du modèle OKMED est donné par :

$$\mathcal{J}(\{\pi_c\}_{c=1}^k) = \sum_{x_i \in \mathcal{X}} d^2(x_i - \phi(x_i)) \tag{3}$$

Il s'agira donc à nouveau de minimiser l'inertie des individus par rapport à leur image. La notion d'image quant à elle doit être redéfinie en prenant appui sur des médoïdes de classe plutôt que des centroïdes. Nous définissons alors l'image $\phi_{OKMED}(x_i)$ d'un individu x_i dans la classification $\{\pi_c\}_{c=1}^k$ par l'individu de $\mathcal{X}$ qui minimise la somme des distances avec tous les médoïdes des classes d'appartenance de x_i :

$$\phi_{OKMED}(x_i) = \underset{x_j \in \mathcal{X}}{\arg\min} \sum_{m_c \in A_i} d(x_j, m_c) \tag{4}$$

Notons qu'avec cette nouvelle définition, la recherche d'une image nécessite de parcourir l'ensemble des individus de $\mathcal{X}$. En pratique on pourra se contenter d'effectuer une seule fois cette recherche pour chaque combinaison[2] d'affectations A_i observée.

Il est enfin utile de préciser que dans le cas d'un partitionnement strict, chaque individu x_i étant affecté à une seule classe π_c, l'image $\phi(x_i)$ correspond exactement au médoïde m_c. Ainsi le modèle OKMED, via le critère d'inertie (3), doit effectivement être vu comme une généralisation des modèles de type k-médoïdes.

3.3 L'algorithme OKMED

Dans la lignée des méthodes d'agrégation autour de centres mobiles, nous proposons un algorithme visant à optimiser le critère (3) en deux étapes : affectation des individus et mise à jour des paramètres. Nous donnons Figure 1 la description de l'algorithme.

L'affectation d'un individu à une ou plusieurs classes est réalisée au moyen de la fonction ASSIGN() qui se fonde sur l'heuristique proposée par Cleuziou (2008) . L'adaptation de cette heuristique pour OKMED consistera, pour un individu x_i, à parcourir l'ensemble des médoïdes

[2]Le nombre de combinaisons possibles peut être très grand en théorie, cependant seulement un sous-ensemble de combinaisons est observé en pratique.

OKMED(D,k,t_{max},ϵ)

Entrée : D une matrice de distance $(n \times n)$ sur un ensemble d'individus $\mathcal{X}$, k un nombre de classes, t_{max} : un nombre maximum d'itérations (optionnel), ϵ : un paramètre d'évolution minimale du critère objectif (optionnel).
Sortie : $\{\pi_c\}_{c=1}^{k}$: une classification recouvrante sur $\mathcal{X}$.

1. Tirer aléatoirement k médoïdes $\{m_c^{(0)}\}_{c=1}^{k}$ dans $\mathcal{X}$,
   t=0.
2. Pour chaque individu $x_i \in \mathcal{X}$ calculer les affectations

$$A_i^{(t+1)} = \text{ASSIGN}(x_i, \{m_c^{(t)}\}_{c=1}^{k})$$

   en déduire une classification $\{\pi_c^{(t+1)}\}_{c=1}^{k}$ telle que $\pi_c^{(t+1)} = \{x_i | m_c^{(t)} \in A_i^{(t+1)}\}$
3. Pour chaque classe $\pi_c^{(t+1)}$ successivement, calculer le nouveau médoïde

$$m_c^{(t+1)} = \text{MEDOID}(\pi_c^{(t+1)})$$

4. Si $\{\pi_c^{(t+1)}\}$ différent de $\{\pi_c^{(t)}\}$ ou $t < t_{max}$ ou $\mathcal{J}(\{\pi_c^{(t)}\}) - \mathcal{J}(\{\pi_c^{(t+1)}\}) > \epsilon$, alors $t = t+1$ et aller à l'étape 2 ; Sinon retourner la classification $\{\pi_c^{(t+1)}\}_{c=1}^{k}$.

FIG. 1 – *L'algorithme* OKMED.

$\{m_c^{(t+1)}\}_{c=1}^{k}$ dans un ordre précis (du plus proche au plus éloigné au sens de D) et à affecter x_i à la classe correspondante tant que l'inertie $d(x_i, \phi(x_i))$ diminue. La nouvelle affectation $A_i^{(t+1)}$ ne sera conservée que si elle améliore l'ancienne affectation $A_i^{(t)}$ en terme d'inertie toujours. Cette manière de faire assure la décroissance du critère d'inertie pour cette étape.

La mise à jour des paramètres se résume ici à rechercher pour chaque classe un nouveau représentant ou médoïde parmi les individus de la classe, qui soit meilleur au sens du critère d'inertie. L'heuristique de recherche que nous proposons est formalisée par la fonction MEDOID() (voir Figure 2) et privilégie la recherche d'un médoïde pertinent pour la classification plutôt que du médoïde "optimal" pour le critère objectif. Ceci pour deux raisons :

- d'une part car il est souhaitable de limiter les évaluations de médoïdes potentiels qui sont très coûteuses dans notre modèle recouvrant car elles nécessitent, pour chaque individu de la classe, une recherche d'image tenant compte du nouveau médoïde potentiel.
- d'autre part pour éviter autant que possible de choisir comme médoïde d'une classe un individu qui appartiendrait à de nombreuses autres classes ; si un individu propre (i.e. affecté uniquement) à la classe permet d'améliorer le critère d'inertie, celui-ci sera élu parcequ'il est un bon représentant de classe et parcequ'il suffit à faire décroître le critère.

Chacune des deux étapes - affectations et mise à jour des médoïdes - permet de faire diminuer le critère objectif (3). En notant de plus que l'ensemble des classification recouvrantes de n individus en k classes est fini pour n et k fixés, il en découle la convergence de l'algorithme OKMED. La classification obtenue correspond à un optimum local du critère objectif, la méthode étant en effet sensible à l'initialisation.

MEDOID(π_c)

Entrée : π_c une classe d'individus de $\mathcal{X}$.
Sortie : m_c le médoïde de la classe π_c.

1. Calculer l'inertie des individus de π_c :

$$\mathcal{J}(\pi_c) = \sum_{x_i \in \pi_c} d^2(x_i - \phi(x_i))$$

2. **Pour** p **allant de** 1 **à** k **faire :**
Pour chaque $x_j \in \pi_c$ **tel que** $|A_j| = p$ **faire :**
calculer les images $\phi(x_i)'$ avec $m_c = x_j$ pour tout $x_i \in \pi_c$
calculer la nouvelle inertie pour π_c

$$\mathcal{J}'(\pi_c) = \sum_{x_i \in \pi_c} d^2(x_i - \phi'(x_i))$$

si $\mathcal{J}'(\pi_c) < \mathcal{J}(\pi_c)$ retourner x_j (nouveau médoïde de π_c)

FIG. 2 – *Mise à jour des médoïdes de classe.*

4 WOKM pour une pondération locale des classes

4.1 Motivation et modèle de référence

Prenons l'exemple de la classification de documents textuels décrits par des vecteurs de fréquences de mots. Si l'objectif est d'organiser les textes de façon thématique, il est naturel d'imaginer que certains documents soient mono-thématiques (vocabulaire d'un seul thème) et d'autres pluri-thématiques (vocabulaires de plusieurs thèmes). Plutôt que d'avoir à choisir une seule classe, et risquer de perdre une partie importante de l'information du document, la classification recouvrante offre la possibilité d'affecter le document à plusieurs classes. Pour autant, dans les modèles vus précédemment (MOC et OKM), un document qui contiendrait le vocabulaire d'un thème verrait ses chances d'être affecté à la classe thématique correspondante diminuées s'il contenait également les vocabulaires associés à des thèmes différents.

L'idée des modèles avec pondération locale des classes, vise justement à éviter le phénomène précédent, en permettant à l'individu d'être affecté à une classe sur la base des descripteurs importants pour la dite classe. Ainsi, la présence du vocabulaire d'un thème suffirait à décider de l'appartenance du document à la classe associée à ce thème. Ce type de modèle est donc particulièrement approprié lors de la recherche de classes recouvrantes.

Nous proposons ici d'étendre le modèle des k-moyennes pondéré proposé par Chan et al. (2004) au cas recouvrant. Ce modèle généralise le critère des moindres carrés utilisé dans k-moyenne par une pondération des variables, différente pour chaque classe. Soit $\mathcal{X} = \{x_i\}_{i=1}^n$ un ensemble d'individus dans $\mathbb{R}^p$, le critère s'exprime ainsi :

$$\mathcal{J}(\{\pi_c\}_{c=1}^k) = \sum_{c=1}^{k} \sum_{x_i \in \pi_c} \sum_{v=1}^{p} \lambda_{c,v}^{\beta} |x_{i,v} - m_{c,v}|^2 \text{ avec } \forall c, \sum_{v=1}^{p} \lambda_{c,v} = 1 \qquad (5)$$

Dans (5), les $\{\lambda_{c,v}\}$ représentent les poids associés à chaque variable pour chaque classe, et β est un paramètre (> 1) permettant de régler l'influence de la pondération dans le modèle. C'est sur cette base de travail que nous proposons le modèle WOKM qui généralise à la fois les modèles OKM et k-moyennes pondéré.

4.2 Le modèle WOKM

Intégrer la pondération locale des classes dans le critère d'inertie (1) des modèles recouvrants n'est pas trivial. En effet, l'inertie mesure la dispersion des individus vis à vis de leur image plutôt que de leur représentant de classe. Il s'agit donc de se poser en premier la question de définir l'image d'un individu dans un environnement de classes avec pondération. Nous proposons de définir l'image de x_i par la moyenne pondérée des centres des classes de x_i :

$$\phi_{WOKM}(x_i) = (\phi_1(x_i), \ldots, \phi_p(x_i)) \text{ avec } \phi_v(x_i) = \frac{\sum_{m_c \in A_i} \lambda_{c,v}^{\beta} m_{c,v}}{\sum_{m_c \in A_i} \lambda_{c,v}^{\beta}} \quad (6)$$

Cette définition assure d'une part la généricité du modèle et d'autre part une bonne intuition pour la notion d'image. Par ailleurs, l'image d'un individu x_i modélise en quelque sorte un point de $\mathbb{R}^p$ représentatif de l'intersection des classes de A_i. Puisque chaque classe π_c est caractérisée par un vecteur de poids λ_c, il convient de proposer une pondération des intersections et donc, par analogie, de proposer un vecteur de poids γ_i pour les images $\phi(x_i)$. On pose alors :

$$\gamma_{i,v} = \frac{\sum_{m_c \in A_i} \lambda_{c,v}}{|A_i|} \quad (7)$$

Il en découle le critère objectif suivant pour le modèle WOKM :

$$\mathcal{J}(\{\pi_c\}_{c=1}^k) = \sum_{x_i \in \mathcal{X}} \sum_{v=1}^{p} \gamma_{i,v}^{\beta} |x_{i,v} - \phi_v(x_i)|^2 \quad (8)$$

Ce critère est soumis à la contrainte $\forall c,\ \sum_{v=1}^{p} \lambda_{c,v} = 1$ sur les pondérations locales des classes, encapsulées dans la définition des poids $\{\gamma_{i,v}\}$. On note la généricité du modèle en observant que :

- dans le cas d'affectations strictes, si $x_i \in \pi_c$ alors $\phi_v(x_i) = m_{c,v}$ et $\gamma_{i,v} = \lambda_{c,v}$; le critère se ramène alors au critère (5) utilisé dans les k-moyennes pondéré.
- dans le cas de pondérations uniformes ($\forall c,\ \forall v,\ \lambda_{c,v} = 1/p$), $\gamma_{i,v} = 1/p$ et $\phi_{WOKM}(x_i) = \phi_{OKM}(x_i)$; le critère se ramène au critère (1) utilisé dans OKM, à une constante près.

4.3 L'algorithme WOKM

L'optimisation du critère (8) se traduit algorithmiquement par l'itération de trois étapes : affectation, mise à jour des centres et mise à jour des poids (cf. Figure 3).

L'étape d'affectation (ASSIGN) procède de façon similaire aux algorithmes OKM et OKMED, par affectation d'un individu à ses classes proches tant que $\sum_{v=1}^{p} \gamma_{i,v}^{\beta} |x_{i,v} - \phi_v(x_i)|^2$ diminue. La mise à jour des centres de classes (CENTROID) peut être réalisée sur chaque classe successivement en considérant les autres centres fixés ; dans ce contexte on peut montrer[3] que le centre optimal m_c^* pour la classe π_c est donné par le centre de gravité du nuage de points

[3] La preuve n'est pas présentée ici faute de place ; ce résultat s'obtient par minimisation d'une fonction convexe.

WOKM($\mathcal{X}$,k,t_{max},ϵ)

Entrée : $\mathcal{X}$ un ensemble d'individus dans $\mathbb{R}^p$, k un nombre de classes, t_{max} un nombre maximum d'itérations (optionnel), ϵ un paramètre d'évolution minimale du critère objectif (optionnel).

Sortie : $\{\pi_c\}_{c=1}^k$: une classification recouvrante sur $\mathcal{X}$.

1. Tirer aléatoirement k centres $\{m_c^{(0)}\}_{c=1}^k$ dans $\mathbb{R}^p$ ou dans $\mathcal{X}$,
   Initialiser les poids $\{\lambda_{c,v}^{(0)}\}$ uniformément ($\lambda_{c,v}^{(0)} = 1/p$),$t = 0$.
2. Pour chaque individu $x_i \in \mathcal{X}$ calculer les affectations
$$A_i^{(t+1)} = \text{ASSIGN}(x_i, \{m_c^{(t)}\}_{c=1}^k)$$
   en déduire une classification $\{\pi_c^{(t+1)}\}_{c=1}^k$ telle que $\pi_c^{(t+1)} = \{x_i | m_c^{(t)} \in A_i^{(t+1)}\}$
3. Pour chaque classe $\pi_c^{(t+1)}$ successivement, calculer le nouveau centre
$$m_c^{(t+1)} = \text{CENTROID}(\pi_c^{(t+1)})$$
4. Pour chaque classe $\pi_c^{(t+1)}$ successivement, calculer la nouvelle pondération
$$\lambda_{c,\cdot} = \text{WEIGHTING}(\pi_c^{t+1})$$
5. Si $\{\pi_c^{(t+1)}\}$ différent de $\{\pi_c^{(t)}\}$ ou $t < t_{max}$ ou $\mathcal{J}(\{\pi_c^{(t)}\}) - \mathcal{J}(\{\pi_c^{(t+1)}\}) > \epsilon$, alors $t = t+1$ et aller à l'étape 2 ; Sinon retourner la classification $\{\pi_c^{(t+1)}\}_{c=1}^k$.

FIG. 3 – *L'algorithme* WOKM.

$\{(\hat{x}_i^c, w_i) | x_i \in \pi_c\}$. La notation $\hat{x}_i^c$ désigne le centre de la classe π_c qui permettrait à l'image de l'individu x_i d'être confondue avec x_i lui-même ($\forall v,\ |x_{i,v} - \phi_v(x_i)| = 0$) et w_i désigne le vecteur de pondération associé et défini par : $w_{i,v} = \frac{\gamma_{i,v}^{\beta}}{\left(\sum_{m_l \in A_i} \lambda_{l,v}^{\beta}\right)^2}$.

Enfin la troisième étape (WEIGHTING), la mise à jour des vecteurs de poids $\{\lambda_c\}_{c=1}^k$, revient à résoudre un problème d'optimisation sous contrainte ($\sum_{v=1}^p \lambda_{c,v} = 1$) ; contrairement au modèle non recouvrant de Chan et al. (2004), le caractère recouvrant de notre modèle ne permet pas la mise à jour de chaque vecteur λ_c de façon indépendante ; le théorème proposé par Bezdek (1981) pour la classification floue n'assure pas l'optimalité de la solution dans ce contexte. Néanmoins, nous utiliserons une heuristique qui s'inspire de ce théorème et qui consiste, pour chaque classe indépendamment, à :

1. calculer une nouvelle pondération $\lambda_{c,v}$ de la classe π_c en évaluant la variance des individus propres à la classe sur chaque variable :
$$\lambda_{c,v} = \frac{\left(\sum_{\{x_i \in \pi_c |\ |A_i|=1\}} (x_{i,v} - m_{c,v})^2\right)^{1/(1-\beta)}}{\sum_{u=1}^p \left(\sum_{x_i \in \pi_c |\ |A_i|=1} (x_{i,u} - m_{c,u})^2\right)^{1/(1-\beta)}}$$
2. conserver cette pondération seulement si elle améliore le critère d'inertie globale du modèle.

Les choix d'affectation et de mise à jour des paramètres ont chaque fois été effectués de manière à assurer la décroissance du critère objectif et donc la convergence de WOKM.

5 Expérimentations

Nous présentons une série d'expérimentations préliminaires visant à observer le comportement des deux variantes OKMED et WOKM. Le premier jeu de données utilisé (Iris) est familier des chercheurs du domaine et permet de se faire une première idée sur les performances d'une méthode de classification ; le second (Reuters) correspond d'avantage aux applications ciblées dans cette étude puisqu'il s'agit de documents textuels multi-labels.

Pour chaque expérience, l'évaluation proposée consiste à comparer la classification obtenue avec la classification de référence (labels) en terme de mesures de précision, rappel et F-Score. Ces indices[4] sont ceux utilisés et détaillés par Cleuziou (2007) pour OKM et Banerjee et al. (2005) pour MOC.

5.1 Expérimentations sur la base Iris

La base Iris (base de l'UCI repository) comporte 150 individus dans $\mathbb{R}^4$ organisés en trois classes de mêmes tailles, dont l'une (*setosa*) est connue pour être séparée des deux autres.

Les valeurs présentées dans le tableau Tab.1 résultent d'une moyenne sur 500 exécutions des six méthodes avec les mêmes conditions initiales et avec $k = 3$.

	Précision	Rappel	F-Score	Affectations
k-moyennes	0.75	0.82	0.78	1.00
k-médoïdes	0.75	0.84	0.79	1.00
k-moyennes pondéré	0.85	0.89	0.86	1.00
OKM	0.57	0.98	0.72	1.40
OKMED	0.61	0.88	0.71	1.16
WOKM	0.62	0.98	0.76	1.32

TAB. 1 – *Comparaisons des performances des modèles sur Iris.*

Les résultats sur les méthodes non recouvrantes sont données à titre indicatif ; il est en effet normal que les méthodes recouvrantes obtiennent des résultats inférieurs puisque le jeu de données n'est pas multi-labels.

On note en premier lieu que OKMED obtient un F-Score sensiblement égal à celui de OKM ; dans la mesure où on observe le même phénomène sur leur analogue non-recouvrant, ce résultat vient conforter expérimentalement le fait que OKMED généralise k-médoïdes. On remarquera également, via l'indice d'affectations, que OKMED génère moins de recouvrements que OKM ; ce phénomène s'explique naturellement par le fait que la recherche d'images est limitée aux individus de $\mathcal{X}$ dans OKMED et élargie à l'ensemble des points de $\mathbb{R}^p$ dans OKM.

Enfin, la supériorité de la version pondérée de k-moyennes est également observée dans les versions recouvrantes. Ceci vient confirmer empiriquement le modèle WOKM en tant que généralisation du modèle de Chan et al. (2004) et surtout valide notre intérêt pour les approches avec pondération locale des classes.

[4] Aussi l'indice "affectations" correspondant au nombre moyen de classes auxquelles chaque individu appartient.

5.2 Expérimentations sur les données Reuters

La seconde série d'expérimentations est réalisée sur le corpus Reuters traditionnellement utilisé comme benchmark en recherche d'information. Nous nous limitons à un sous-ensemble de 300 documents multi-labels et décrits par des vecteurs de fréquences sur un vocabulaire constitué des 500 mots les plus pertinents au sens de l'indice tfxidf.

Afin d'illustrer la possibilité offerte par OKMED d'avoir recours à des métriques autres que la distance euclidienne, nous observons (Figure 4) les performances de OKM, de OKMED avec la distance euclidienne puis avec la divergence de Kullback-Leibler (ou I-Divergence), en faisant varier le nombre k de classes.

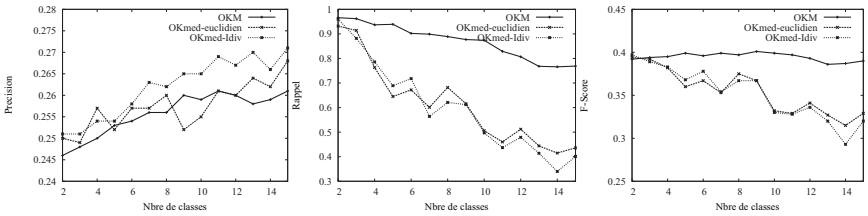

FIG. 4 – OKMED *avec différentes distances.*

Nous pouvons ainsi remarquer que OKMED présente un comportement stable pour des distances différentes et surtout que l'utilisation de la I-Divergence, reconnue performante pour comparer des textes, permet d'obtenir des précisions meilleures et inaccessibles par la distance euclidienne. Ce dernier résultat valide l'intérêt du modèle OKMED.

Enfin, les courbes présentées en Figure 5 précisent l'influence des modèles de pondération locale en particulier pour la classification recouvrante.

Si la pondération locale n'apporte pas de changement significatif dans les modèles non-recouvrants (k-moyennes $\approx$ k-moyennes pondéré), ceci n'est pas vrai dans les modèles recouvrants. En effet, l'apport espéré par le modèle WOKM vis à vis de son analogue non pondéré OKM, se traduit empiriquement par :

1. une limitation des recouvrements ; le nombre moyen d'affectations plus faible ;
2. une réduction du rappel ; ce qui est une conséquence directe de l'observation précédente ;
3. une amélioration significative de la précision.

D'une manière générale, la pondération locale introduite dans WOKM semble venir corriger le modèle OKM en limitant les affectations multiples "parasites".

6 Conclusion et perspectives

Nous avons exposé dans cette étude deux contributions dans le domaine de la classification recouvrante. La première s'inspire des techniques de partitionnement par agrégation autour de médoïdes en proposant le modèle OKMED ; ce dernier permet d'organiser un ensemble d'individus en classes recouvrantes d'individus similaires, sur la base uniquement d'une matrice de distance. La seconde contribution propose, à travers l'algorithme WOKM, d'introduire une pondération locale des classes dans les modèles de classification recouvrante.

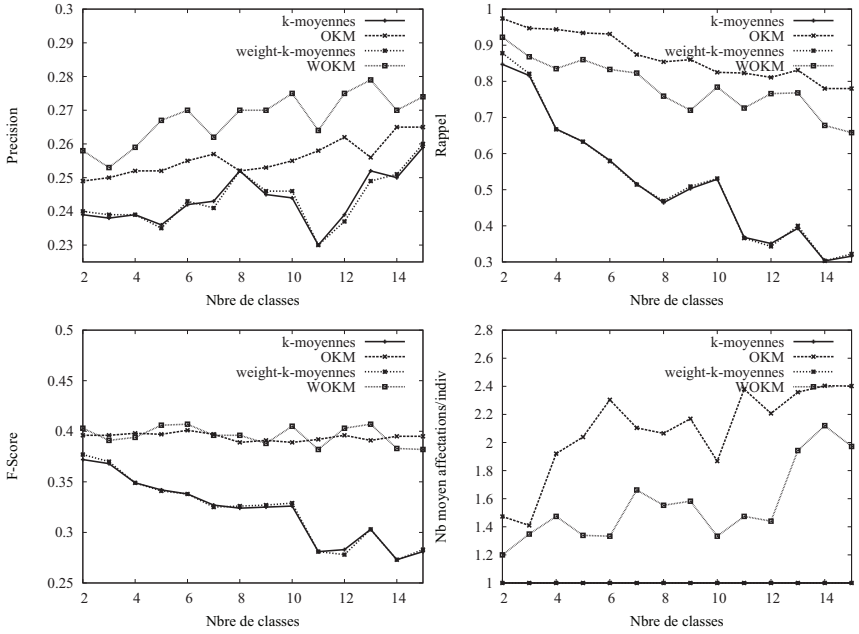

FIG. 5 – *Influence de la pondération locale des classes.*

Ces deux apports se présentent comme des généralisations à la fois des méthodes de partitionnement strict (k-moyennes, k-médoïdes et k-moyennes pondéré) et des méthodes de recouvrement (OKM et MOC). Nous avons justifié les critères objectifs associés, proposé des algorithmes et heuristiques d'optimisation de ces critères puis validé ces méthodes par des expérimentations préliminaires sur des jeux de données adaptés.

Nous chercherons à confirmer les bonnes propriétés observées sur ces méthodes par des expérimentations supplémentaires mettant en jeu d'autres corpus textuels, et d'autres domaines (e.g. Bioinformatique). Nous envisagerons de poursuivre l'enrichissement de cette famille de méthodes de classification recouvrante en explorant d'autres variantes pertinentes telles que les cartes auto-organisatrices recouvrantes, la classification recouvrante à base de noyaux, etc. Cependant, de façon plus directement liée à cette étude, les deux variantes proposées posent les bases d'une approche combinant les bienfaits des deux modèles au sein d'un algorithme d'agrégation autour de petits ensembles de médoïdes capturant la forme des clusters.

Références

Banerjee, A., C. Krumpelman, J. Ghosh, S. Basu, et R. J. Mooney (2005). Model-based overlapping clustering. In *KDD '05 : Proceeding of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining*, New York, NY, USA, pp. 532–537. ACM Press.

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.

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

Chan, E. Y., W.-K. Ching, M. K. Ng, et J. Z. Huang (2004). An optimization algorithm for clustering using weighted dissimilarity measures. *Pattern Recognition 37*(5), 943–952.

Cleuziou, G. (2007). Okm : une extension des k-moyennes pour la recherche de classes recouvrantes. In *EGC'2007*, Volume 2, Namur, Belgique. Revue des Nouvelles Technologies de l'Information, Cépaduès-Edition.

Cleuziou, G. (2008). An Extended Version of the k-Means Method for Overlapping Clustering. In *19th Conference on Pattern Recognition ICPR'08 (to appear)*.

Cleuziou, G. et J.-H. Sublemontier (2008). étude comparative de deux approches de classification recouvrante : Moc vs. okm. In *8èmes Journées Francophones d'Extraction et de Gestion des Connaissances*, Volume 2. Revue des Nouvelles Technologies de l'Information, Cépaduès-Edition.

Dattola, R. (1968). A fast algorithm for automatic classification. Technical report, Report ISR-14 to the National Science Foundation, Section V, Cornell University, Department of Computer Science.

Diday, E. (1987). Orders and overlapping clusters by pyramids. Technical report, INRIA num.730, Rocquencourt 78150, France.

Diday, E. et G. Govaert (1977). Classification avec distances adaptatives. *RAIRO 11*(4), 329–349.

D.Pelleg et A. Moore (2000). X-means : Extending k-means with efficient estimation of the number of clusters. In *Proceedings of the Seventeenth International Conference on Machine Learning*, San Francisco, pp. 727–734. Morgan Kaufmann.

Jardine, N. et R. Sibson (1971). *Mathematical Taxonomy*. London : John Wiley and Sons Ltd.

Kaufman, L. et P. J. Rousseeuw (1987). Clustering by means of medoids. *In Dodge, Y. (Ed.) Statistical Data Analysis based on the L1 Norm*, 405–416.

Likas, A., N. Vlassis, et J. Verbeek (2003). The global k-means clustering algorithm. *Pattern Recognition 36*, 451–461.

Peña, J., J. Lozano, et P. Larrañaga (1999). An empirical comparison of four initialization methods for the k-means algorithm. *Pattern Recognition Letters 20*(50), 1027–1040.

Summary

This paper deals with overlapping clustering and presents two extensions of the approach OKM denoted as OKMED and WOKM. OKMED generalizes the well known k-medoid method to overlapping clustering and help in organizing data with distance matrices as input. WOKM (Weighted-OKM) proposes a model with local weighting of the classes; this variant is suitable for overlapping clustering since a single data can matches with multiple classes according to different characteristics. On text clustering, we show that OKMED has a behavior similar to OKM but offers to use metrics other than euclidean distance. Then we observe significant improvement using the weighted extension of OKM.