OKM : une extension des k-moyennes pour la recherche de classes recouvrantes

Guillaume Cleuziou

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

Résumé. Dans cet article nous abordons le problème de la classification (ou clustering) dans le but de découvrir des classes avec recouvrements. Malgré quelques avancées récentes dans ce domaines, motivées par des besoins applicatifs importants (traitements des données multimédia par exemple), nous constatons l'absence de solutions théoriques à ce problème. Notre étude consiste alors à proposer une nouvelle formulation du problème de classification par partitionnement, adaptée à la recherche d'un recouvrement des données en classes d'objets similaires. Cette approche se fonde sur la définition d'un critère objectif de qualité d'un recouvrement et d'une solution algorithmique visant à optimiser ce critère. Nous proposons deux évaluations de ce travail permettant d'une part d'appréhender le fonctionnement global de l'algorithme sur des données simples (vitesse de convergence, visualisation des résultats) et d'autre part d'évaluer quantitativement le bénéfice d'une telle approche sur une application de classification de documents textuels.

1 Introduction

La classification automatique (ou *clustering*) est un domaine d'étude situé à l'intersection de deux thématiques de recherches majeures que sont l'analyse de données et l'apprentissage automatique. Ce domaine est en perpétuelle évolution du fait de l'apparition constante de nouveaux besoins portant à la fois sur la quantité ou la nature des données à traiter (numériques, symboliques, spatiales, histogrammes, etc.) que sur le type de classification attendue (partition, hiérarchie, schéma flou, etc.).

Nombreuses sont les approches proposées afin d'organiser, de résumer ou de simplifier un ensemble de données à l'aide d'une structure de laquelle il est possible de faire émerger des classes d'objets similaires au sens d'un critère de proximité défini ou plus généralement au regard des propriétés que ces objets partagent. Il est de coutume de structurer ces approches en différentes catégories mutuellement non-exclusives (voir Jain et al. (1999)) comme par exemple, pour ne citer que les principales, les approches hiérarchiques, par partitionnement ou encore les modèles de mélanges.

Les approches par partitionnement, dont l'algorithme des k-moyennes (MacQueen, 1967) en est l'un des plus célèbre représentant, consiste le plus souvent à construire une collection de classes disjointes formant une partition des données par optimisation d'un critère objectif.

Ce critère étant généralement choisi de façon à minimiser la variance intra-classe (les objets à l'intérieur d'une classe doivent tous être assez similaires) et/ou à maximiser la variance inter-classes (les classes doivent être séparées les unes des autres).

Les approches hiérarchiques aboutissent en revanche à une collection de classes emboîtées, que l'on peut représenter par un arbre ou plus généralement un graphe dont les arêtes modélisent une relation d'inclusion. Généralement agglomératifs (parfois divisifs) les algorithmes proposés procèdent par fusions successives de classes similaires et les plus utilisés restent sans nul doutes les méthodes agglomératives hiérarchiques (liens simple, complet, moyen ou critère de Ward) présentées dans Sneath et Sokal (1973).

Enfin, pour les approches de classification par mélange de lois, le problème est posé de façon à maximiser la vraisemblance d'un modèle faisant l'hypothèse que les données sont des observations d'un mélange de densités. Le modèle est caractérisé par le nombre de lois supposées et leurs paramètres. La méthode EM par exemple, proposée par Dempster et al. (1977) constitue une solution algorithmique incontournable à ce genre de problème d'optimisation.

Si la plupart des domaines d'application trouvent dans cette pluralité d'approches des réponses satisfaisantes aux besoins exprimés, des domaines récents nécessitent d'adapter voire de reposer la problématique de la classification et d'y adjoindre une solution algorithmique efficace. Nous nous intéressons ici au problème de la classification de données en classes non-disjointes (également dites empiétantes ou recouvrantes)[1]. Ce type d'approche vise à structurer les données en une collection de classes telle que chaque objet puisse appartenir à plusieurs classes, correspondant alors à une organisation naturelle pour des données par exemple multimédia (texte, image et/ou vidéo) ou encore biologiques (gènes) (Banerjee et al., 2005).

Nous présenterons tout d'abord en Section 2 un ensemble de pistes proposées (approches pyramidales, classification floue, etc.) pour répondre plus ou moins directement aux besoins exprimés par les domaines d'application mentionnés ci-dessus. Nous montrerons alors qu'aucune de ces approches ne constitue une solution globale pour le problème posé. Nous tenterons en Section 3 d'en proposer une nouvelle formalisation et d'y adjoindre une première solution algorithmique en nous inspirant de l'algorithme simple et efficace des k-moyennes. L'approche ainsi présentée sera ensuite observée dans son fonctionnement puis évaluée sur deux jeux de données réelles dont la collection de textes Reuters (Section 4). Enfin nous proposerons en guise de conclusion, un ensemble de perspectives à cette étude, visant à positionner le problème de classification avec recouvrements comme sous-domaine à part entière des recherches menées en classification automatique.

2 Problématique de la classification avec recouvrements

2.1 Des solutions partielles

Une première voie de recherches conduisant à une structuration des données en classes empiétantes réside dans les techniques de classification pyramidales initiées par Diday (1984). Cependant, l'ensemble des recouvrements envisageables par une telle structure se limite aux collections de classes telles que chaque classe s'intersecte avec au plus deux autres classes.

Plusieurs autres structures hiérarchiques ont été proposées par la suite afin d'étendre les schémas atteignables de façon à approcher l'ensemble de tous les recouvrements possibles ; en

[1] En Anglais : *Overlapping clustering*.

particulier les hiérarchies faibles puis les k-hiérarchies faibles (Bertrand et Janowitz, 2003). Cependant deux points restent à déplorer : d'une part il n'existe pas aujourd'hui de méthode algorithmique permettant de construire de telles structures hiérarchiques, et d'autre part l'ensemble des recouvrements atteints - bien que très largement étendu - reste limité aux collections de classes vérifiant la propriété suivante : "l'intersection de $(k+1)$ classes arbitraires peut être réduite à l'intersection de k de ces classes".

Un second axe de recherches a été assez fortement étudié ces dernières années et consiste soit à adapter des algorithmes existants (k-moyennes et sa variante floue ou encore EM) ou à développer de nouvelles méthodologies spécifiées pour la recherche d'un "bon" recouvrement des données en classes d'objets similaires. Dans cette dernière classe de méthodes on peut citer les algorithmes des k-moyennes axiales (Lelu, 1994) et CBC (*Clustering By Committee*) développé par Pantel (2003) tous deux motivés par l'application aux données textuelles (mots ou documents) ou encore l'algorithme plus général POBOC (*Pole-Based Overlapping Clustering*) proposé par Cleuziou et al. (2004).

De façon globale, qu'il s'agisse d'algorithmes nouveaux ou simplement adaptés, toutes ces méthodes consistent, en une ou plusieurs itérations, à rechercher des centres auxquels sont affectés les objets. Ces centres peuvent être des points de l'espace (k-moyennes mais aussi EM[2]), des axes (k-moyennes axiales) ou encore des petits ensembles d'objets appelés *committee* dans CBC et *Pole* dans POBOC. Quelque soit la forme prise par ces centres, l'algorithme permettant de les obtenir ne prend pas en considération le fait que les classes finales formeront un recouvrement et pourront ainsi contenir des objets communs. Par exemple les méthodes d'agrégation autour des centres mobiles (k-moyennes et k-moyennes axiales) déterminent un centre après affectation de chaque objet à un seul de ces centres ; à l'inverse les variantes floues de ce type de méthodes considèrent systématiquement que tous les objets doivent participer à la définition de chaque centre ; l'une des hypothèses utilisées dans l'algorithme EM vise à considérer que chaque objet est une observation de l'une (et une seule) des lois du mélange ; enfin les algorithmes CBC et POBOC définissent les centres indépendamment de tout critère objectif de qualité du recouvrement induit par ces centres.

Ainsi définis, les centres sont déterminants pour la dernière étape d'affectation qui conduira au schéma final de classification. L'affectation est le plus souvent réalisée au moyen d'un seuil[3], difficile à déterminer, quitte à violer les fondements théoriques sur lesquels l'algorithme repose, par exemple la minimisation d'un critère objectif. Finalement, l'hypothèse sous-jacente formulée par ces approches vise à considérer qu' *"un recouvrement de qualité correspond nécessairement à l'extension d'une "bonne" partition"*.

2.2 Recouvrements et partitions étendues

L'hypothèse précédemment formulée ne semble pas incohérente de prime abord. En théorie tout recouvrement $\mathcal{R} = \{R_1, \ldots, R_k\}$ d'un ensemble d'objets X peut être obtenu par l'extension d'au moins une partition $\mathcal{P} = \{P_1, \ldots, P_k\}$ telle que $\forall i \in 1, \ldots, k\ ,\ P_i \subseteq R_i$. En revanche, partant d'une partition $\mathcal{P}$, l'ensemble des recouvrements possibles par extension

[2]Les approches de classification par mélange de lois consistent à rechercher les paramètres de ces lois ; l'un d'entre eux correspond à la moyenne, identifiable à un point de l'espace.

[3]Un objet est affecté à tous les centres pour lesquels sa distance au centre ou sa probabilité d'appartenance à cette classe est supérieure au seuil.

correspond à une classe de recouvrements (notée $\mathcal{C}_\mathcal{P}$), qui n'est qu'un sous ensemble de tous les recouvrements possibles. Ainsi, considérer qu'un "bon" recouvrement (selon un critère $W(.)$) correspond nécessairement à l'extension d'une "bonne" partition (selon un critère $V(.)$) supposerait que : *si $\mathcal{P}$ optimise le critère $V(.)$ et $\mathcal{R}$ optimise le critère $W(.)$ alors $\mathcal{R} \in \mathcal{C}_\mathcal{P}$*

Cette dernière propriété dépend bien sûr des critères $V(.)$ et $W(.)$ choisis. Nous montrons alors sur un exemple que l'on peut choisir des critères cohérents pour lesquels cette propriété n'est pas vérifiée.

Soit $X = \{x_1, \ldots, x_6\}$ un ensemble d'objets définis dans $\mathbb{R}^2$, présentés en figure 1 et que l'on souhaite organiser en deux classes (k=2). On pose $V(.)$ et $W(.)$ les critères objectifs pour l'évaluation respectivement d'une partition et d'un recouvrement et on les défini de la manière suivante :

$$V(\mathcal{P}) = \sum_{P_j \in \mathcal{P}} \sum_{x_i \in P_j} d(x_i, c_j) \quad \text{et} \quad W(\mathcal{R}) = \sum_{x_i \in X} d(x_i, \overline{x_i})$$

avec d la distance de Manhattan, c_j de centre de gravité des objets d'une classe P_j ou R_j selon le modèle, et $\overline{x_i}$ le centre de gravité de l'ensemble $\{c_j | x_i \in R_j\}$.

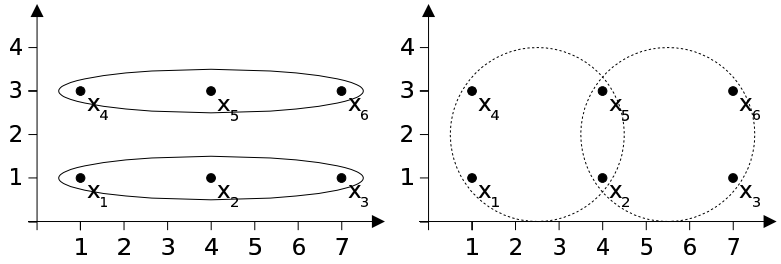

FIG. 1 – *Partition (à gauche) et recouvrement (à droite) optimaux selon les critères $V(.)$ et $W(.)$ respectivement.*

On peut calculer que la partition $\mathcal{P} = \{\{x_1, x_2, x_3\}, \{x_4, x_5, x_6\}\}$ minimise le critère $V(.)$ ($V(\mathcal{P}) = 12.0$) et que $\mathcal{R} = \{\{x_1, x_2, x_4, x_5\}, \{x_2, x_3, x_5, x_6\}\}$ minimise le critère $W(.)$ ($W(\mathcal{R}) = 12.0$ également). Pourtant $\mathcal{R}$ n'est pas une extension de la partition $\mathcal{P}$ et n'appartient donc pas à la classe des recouvrements $\mathcal{C}_\mathcal{P}$.

Nous venons donc de montrer par un exemple simple que la classification avec recouvrements ne se résume pas à étendre une partition par des affectations supplémentaires. Faire cette hypothèse consisterait à ne considérer qu'un sous-espace de l'espace de recherche d'une solution, ce dernier étant défini par l'ensemble des recouvrements possibles.

3 À la recherche d'un "bon" recouvrement

3.1 Définition du problème

Rechercher une partition $\mathcal{P}$ d'un ensemble $X = \{x_1, \ldots, x_n\}$ en k classes $P_1, \ldots, P_k$ selon un critère $V(.)$ défini sur l'ensemble des partitions possibles est un problème NP-difficile dans la mesure où l'espace de recherche est de taille exponentielle (k^n partitions possibles). L'algorithme bien connu des k-moyennes (MacQueen, 1967) propose une solution partielle

au problème d'optimisation du critère des moindres carrés (aussi appelé critère de variance intra-classe) : $V(\mathcal{P}) = \sum_{P_j \in \mathcal{P}} \sum_{x_i \in P_j} d^2(x_i, c_j)$.

Ce critère favorise les partitions dont les classes présentent une faible variance, autrement dit telles que les objets à l'intérieur d'une même classe sont faiblement dispersés. Cet algorithme procède par itérations de deux étapes (calcul des centres de classes puis affectation de chaque objet à son centre le plus proche) assurant la décroissance du critère et par la même, la convergence de la méthode vers une partition stable. La solution ainsi obtenue correspond à un minimum seulement local du critère et dépend de l'initialisation (tirage aléatoires de k centres) de l'algorithme.

Le problème de recherche d'un recouvrement minimisant un critère $W(.)$ ne peut pas être considéré comme plus facile que le précédent puisque l'espace de recherche est de taille beaucoup plus importante ($2^{k.n}$ recouvrements possibles). Par ailleurs le critère $V(.)$ permettant d'évaluer la qualité d'une partition n'est plus adapté dans le cas des recouvrements car si $\mathcal{R}$ est un recouvrement de X en k classes on peut montrer que $\forall \mathcal{P},\ \mathcal{R} \in \mathcal{C}_{\mathcal{P}} \Rightarrow V(\mathcal{P}) \leq V(\mathcal{R})$; un bon recouvrement selon $V(.)$ ne pouvant alors être qu'une partition.

Dans cette étude, notre proposition porte ainsi sur la définition d'un nouveau critère de qualité d'un recouvrement d'une part et d'une solution algorithmique permettant d'approcher un recouvrement optimal selon ce critère d'autre part. Pour y parvenir nous nous inspirons de l'algorithme simple et efficace des k-moyennes.

3.2 Critère objectif pour les recouvrements

Pour définir un critère de qualité d'un recouvrement il est indispensable de se reporter aux motivations premières qui nous conduisent à rechercher ce type d'organisation. Dans le cas d'un document par exemple, choisir une classe thématique et une seule pour ce document peut réduire considérablement la représentation que l'on conservera de ce document dans la classification. En revanche, autoriser ce document à s'afficher selon plusieurs thèmes rendra une image certainement plus juste de son contenu. La qualité d'un recouvrement pourra alors être mesurée relativement à l'écart entre le contenu réel des objets et l'"image" que la classification (ici le recouvrement) établie renvoie d'eux. Nous formalisons cette intuition dans le critère suivant :

$$W(\mathcal{R}) = \sum_{x_i \in X} d^2(x_i, \overline{x_i})$$

L'image d'un objet dans un recouvrement $\mathcal{R}$ est notée $\overline{x_i}$ dans ce critère et correspond à un compromis entre les différentes classes auxquelles cet objet appartient. Ainsi pour un recouvrement $\mathcal{R}$ en k classes $\{R_1, \ldots, R_k\}$ de centres respectifs $\{c_1, \ldots, c_k\}$, $\overline{x_i}$ est défini par le centre de gravité de l'ensemble $\{c_j | x_i \in R_j\}$.

3.3 L'algorithme OKM

L'algorithme OKM (*Overlapping k-means)* que nous détaillons dans cette section présente un squelette (figure 2) similaire à l'algorithme des k-moyennes. L'initialisation qui consiste à tirer aléatoirement k centres puis à dériver un premier recouvrement est suivie par l'itération de deux étapes : (1) la mise à jour des centres de classes puis (2) l'affectation des objets à ces centres.

Initialisation : t=0
choisir aléatoirement k centres $C^t = \{c_1^t, c_2^t, \ldots, c_k^t\}$ dans X,
Pour chaque $x_i \in X$: Affecter(x_i,C^t),
en déduire un recouvrement initial $\mathcal{R}^t = \{R_1^t, R_2^t, \ldots, R_k^t\}$.
Faire
t=t+1
• Mettre_à_jour($C^{(t-1)}$,$\mathcal{R}^{(t-1)}$) (*en déduire* C^t),
• Pour chaque $x_i \in X$: Affecter(x_i,C^t) (*en déduire* $\mathcal{R}^t$),
Tant que $\mathcal{R}^t \neq \mathcal{R}^{t-1}$

FIG. 2 – *Squelette de l'algorithme* OKM.

L'intérêt de l'algorithme OKM réside dans la méthode employée pour Mettre_à_jour les centres et pour Affecter chaque objet à un ou plusieurs centres. Ces deux opérations doivent d'une part assurer la cohérence des classes en regroupant ensemble des objets similaires et d'autre part permettre la convergence de la méthode par décroissance du critère $W(.)$.

Étant donné un ensemble $C = \{c_1, c_2, \ldots, c_k\}$ correspondant aux centres des k classes respectives $R_1, R_2, \ldots, R_k$ d'un recouvrement $\mathcal{R}$, la méthode d'affectation d'un objet x_i, présentée en figure 3 consiste à parcourir l'ensemble des centres de classes du plus proche au plus éloigné (suivant une métrique d) et à affecter x_i tant que son image est améliorée ($d(x_i, \overline{x_i})$ diminue). La nouvelle affectation de l'objet x_i ne sera finalement conservée que si l'image de x_i s'en trouve améliorée par rapport à l'ancienne affectation. Cette dernière précaution permet d'assurer la décroissance du critère $W(.)$ lors de l'étape d'affectation.

Affecter(x_i,C) :
Initialisation : Soit c^* le centre de C le plus proche de x_i ($\forall c_j \in C,\ d(x_i, c^*) \leq d(x_i, c_j)$),
$A = \{c^*\}$ (liste des affectations),
$C = C \setminus \{c^*\}$.
Faire
Soit c^* le centre de C le plus proche de x_i et $\overline{x_i}^A$ le centre de gravité des éléments de A,
Si $d(x_i, \overline{x_i}^{A \cup \{c^*\}}) < d(x_i, \overline{x_i}^A)$ alors $A \leftarrow \{c^*\}$ et $C = C \setminus \{c^*\}$,
Sinon **STOP**
Tant que $C \neq \emptyset$

Soit A' l'ancienne affectation de x_i,
Si $d(x_i, \overline{x_i}^A) < d(x_i, \overline{x_i}^{A'})$ alors affecter x_i aux centres de A,
Sinon conserver l'ancienne affectation A'.

FIG. 3 – *Méthode d'affectation utilisée dans l'algorithme* OKM.

Enfin, la mise à jour du centre c_j de la classe R_j est définie dans l'algorithme OKM par :

$$c_{j,v} = \frac{1}{\sum_{x_i \in R_j} \frac{1}{\delta_i^2}} \times \sum_{x_i \in R_j} \frac{1}{\delta_i^2} . \hat{x_i}_v^j \qquad (1)$$

Dans cette expression, $c_{j,v}$ désigne la $v^{\text{ième}}$ composante du vecteur c_j, δ_i correspond au nombre de classes de $\mathcal{R}$ auxquelles x_i appartient et $\hat{x_i}^j_v$ symbolise la $v^{\text{ième}}$ composante du centre c_j "idéal" pour l'objet x_i, c'est à dire le centre c_j tel que $d(x_i, \overline{x_i}) = 0$. De façon plus précise on a $\hat{x_i}^j_v = \delta_i.x_{i,v} - (\delta_i - 1).\overline{x_i}_v{}^{A\setminus\{c_j\}}$ où A désigne l'ensemble des centres des classes auxquelles x_i appartient. Il découle de ce qui précède une définition plus intuitive du nouveau centre c_j qui correspond finalement au centre de gravité du nuage de points $\{(\hat{x_i}^j, p_i) | x_i \in R_j\}$ où chaque $\hat{x_i}^j$ est pondéré par $p_i = \frac{1}{\delta_i^2}$.

On montre que chaque mise à jour d'un centre dans OKM permet d'assurer la décroissance du critère $W(.)$ mais également que le nouveau centre calculé est celui qui minimise ce critère.

Preuve

Soient X un ensemble d'objets définis dans $(\mathbb{R}^p, d)$ où d est la distance euclidienne, et $\mathcal{R}$ un recouvrement de X en k classes de centres $c_1, \ldots, c_k$. L'étape de mise à jour dans OKM consistant à recalculer chaque centre un par un, il suffit alors de montrer que le recalcul d'un nouveau centre c_j quelconque minimise le critère

$$W(\mathcal{R}) = \sum_{x_i \in X} d^2(x_i, \overline{x_i}) = \sum_{x_i \in X} \sum_{v=1}^{p} (x_{i,v} - \overline{x_i}_v)^2$$

Par réécriture du terme $\overline{x_i}$ et décomposition de la somme sur les objets de X on obtient :

$$W(\mathcal{R}) = \sum_{x_i \notin R_j} d^2(x_i, \overline{x_i}) + \sum_{x_i \in R_j} \sum_{v=1}^{p} \left[x_{i,v} - \frac{1}{\delta_i} (c_{j,v} + (\delta_i - 1).\overline{x_i}_v^{A\setminus\{c_j\}}) \right]^2$$

Pour les objets n'appartenant pas à la classe R_j, leur image $\overline{x_i}$ est indépendante de c_j, le premier terme est donc constant relativement à c_j. Le second terme constitue une fonction quadratique de c_j qui sera alors minimisée pour une dérivée égale à 0.

$$\frac{\partial W(\mathcal{R})}{\partial c_j} = 0 \Longleftrightarrow \sum_{x_i \in R_j} \frac{1}{\delta_i^2} \sum_{v=1}^{p} \left[c_{j,v} - \delta_i.x_{i,v} + (\delta_i - 1).\overline{x_i}_v^{A\setminus\{c_j\}} \right] = 0$$

d'où le résultat

$$c_{j,v} = \frac{1}{\sum_{x_i \in R_j} \frac{1}{\delta_i^2}} \times \sum_{x_i \in R_j} \frac{1}{\delta_i^2} \left[\delta_i.x_{i,v} - (\delta_i - 1).\overline{x_i}_v^{A\setminus\{c_j\}} \right]$$

□.

Notons pour conclure sur la présentation de l'algorithme, que la méthode des k-moyennes peut être considérée comme un cas particulier de OKM. En effet si on restreint dans OKM chaque objet à n'appartenir qu'à une seule classe (δ_i=1) on retrouve exactement le processus de classification utilisé dans l'algorithme k-moyennes. Il s'agit donc d'un algorithme non-déterministe puisque le résultat dépendra de l'initialisation ; de plus, chaque classe n'étant plus indépendante l'une de l'autre dans un recouvrement, l'algorithme OKM dépendra également de l'ordre de parcours des classes lors de l'étape de mise à jour des centres.

4 Évaluations et applications

L'évaluation des méthodes de classification non-supervisée reste un problème entier dans ce domaine de recherches. Une piste possible pour évaluer (au moins partiellement) une telle méthode est de mesurer sa capacité à retrouver un schéma de classification préétabli ; nous l'utiliserons pour évaluer l'algorithme OKM en insistant toutefois sur les précautions qu'il s'impose de prendre lors de l'interprétation des résultats quantitatifs, notamment du fait de l'hypothèse non vérifiée que l'ensemble de descripteurs est pertinent pour établir la classification attendue.

4.1 Observation du fonctionnement de l'algorithme

Nous proposons ici une première évaluation de l'algorithme OKM permettant au lecteur d'en appréhender le fonctionnement global en terme de : vitesse de convergence, importance et pertinence des recouvrements entre les classes et capacité à retrouver des classes attendues. La base de données Iris (D.J. Newman et Merz, 1998) est largement utilisée dans la communauté apprentissage et constitue un jeu d'évaluation simple constitué de 150 individus (Iris) décrits selon 4 descripteurs numériques (longueur et largeur des pétales et des sépales) et organisés en trois classes de 50 individus chacune (iris Setosa, Versicolour, Virginica).

Sans tenir compte de l'étiquette de classe assignée à chaque individu, nous effectuons une classification en trois classes (k=3) en utilisant la distance euclidienne dans l'espace de représentation des données ($\mathbb{R}^4$), préalablement centrées et réduites. Nous présentons sur les figures 4, 6 et 5. le résultat d'une exécution[4] des algorithmes k-moyennes et OKM dans des conditions initiales identiques.

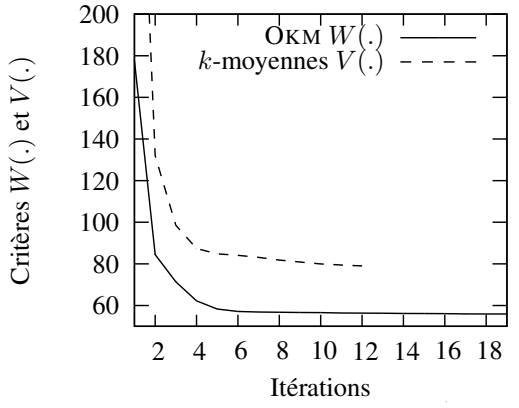

FIG. 4 – *Convergence des algorithmes.*

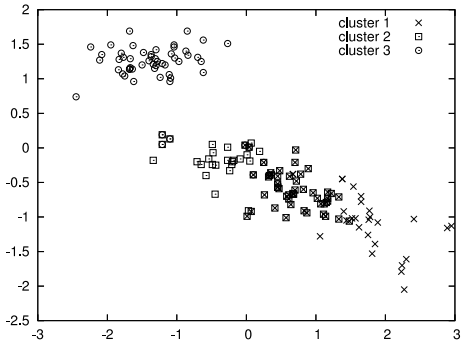

FIG. 5 – *Visualisation des classes par projection sur les deux premiers vecteurs propres.*

[4]L'exécution choisie correspond au meilleur résultat obtenu relativement au critère $W(.)$ final, sur un total de 50 exécutions.

Les courbes de la figure 4 mettent en évidence un phénomène prévisible : la convergence plus lente pour l'algorithme OKM vers un recouvrement stable des données (19 itérations), par rapport à k-moyennes qui obtient une partition stable en seulement 12 itérations. On retiendra cependant que dès la cinquième itération, les deux méthodes ont généré un résultat de qualité qui évoluera peu par la suite.

Étiquettes \ Classes	1	2	3
Iris Setosa			50 *(50)*
Iris Versicolour	26 *(3)*	50 *(47)*	9 *(0)*
Iris Virginica	49 *(36)*	27 *(14)*	

FIG. 6 – *Matrice de confusion pour* OKM *(et k-moyennes).*

La figure 6 présente la matrice de confusion et nous révèle que les deux méthodes identifient correctement les trois catégories d'iris puisque chacune des classes contient majoritairement l'une de ces trois catégories. Il est reconnu que, sur cette base de données Iris, la catégorie des "Iris Setosa" est plutôt facile à identifier tandis que les deux autres catégories sont réputées difficilement séparables, ce que l'on observe sur la figure 5. Ces phénomènes se vérifient également sur notre expérimentation :

- les 50 individus de la catégorie "Iris Setosa" se retrouvent exclusivement dans la classe n°3 avec les deux méthodes. Dans le recouvrement obtenu avec OKM, cette classe contient 9 individus de la catégorie "Iris Versicolour" en supplément, qu'elle partage avec les autres classes[5] (faible intersection) ;
- la séparation difficile des deux autres classes se manifeste par des erreurs de classification si l'on cherche à partitionner les données (k-moyennes) et par une intersection importante entre les deux classes lorsque ces données sont organisées en classes recouvrantes (OKM).

Sur cette première expérimentation, nous avons d'une part observé un comportement satisfaisant de l'algorithme OKM et d'autre part noté que la structuration en classes recouvrantes fournit un résultat informationnel plus riche qu'une simple partition, notamment en ce qui concerne l'organisation des classes entre elles.

4.2 Classification de documents multi-thématiques

Comme nous l'avons mentionné en introduction, les recherches menées autour de la classification avec recouvrements des classes sont motivées par des besoins apparaissant dans des domaines d'application où des données peuvent appartenir à plusieurs catégories prédéfinies. Dans cette seconde expérimentation, nous évaluons l'impact de notre contribution dans le domaine de la Recherche d'Information et plus précisément pour la classification de documents multi-thématiques.

L'expérimentation est conduite sur la collection de documents Reuters[6] initialement composée de 21578 articles journalistiques en langue anglaise. Chaque document peut être étiqueté par une ou plusieurs étiquettes parmi un ensemble de 114 catégories. Après filtrage, nous avons

[5] On observe visuellement (figure 5) que ces objets sont parmi les plus proches des objets propres à la classe n°3.

[6] http ://www.research.att.com/∼lewis/reuters21578.html

retenu 2739 documents, en ne choisissant que ceux pour lesquels au moins une catégorie est proposée, dont le corps de l'article n'est pas vide et appartenant au sous-ensemble "TEST" selon la répartition suggérée dans Apté et al. (1994).

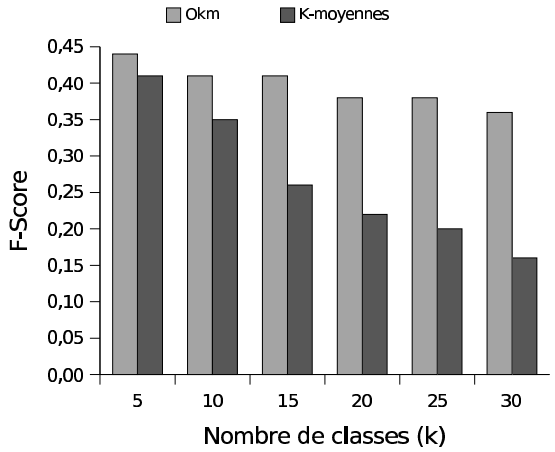

FIG. 7 – *Classification de documents issus de la collection Reuters.*

La représentation des documents est l'aboutissement d'une chaîne de traitements usuelle en Recherche d'Information : chaque document est représenté par un vecteur de dimension m dans lequel chaque composante $x_{i,v}$ correspond à l'information mutuelle du mot[7] w_v pour le document x_i. Une étape préalable de filtrage des descripteurs consiste à ne sélectionner que les m mots d'information mutuelle supérieure à un seuil τ fixé. Enfin, la similarité entre deux documents est évaluée au moyen du cosinus de Salton (Salton et McGill, 1983) :

$$sim(x_i, x_j) = \frac{\sum_v x_{i,v}.x_{j,v}}{\sqrt{\sum_v x_{i,v}^2 \times \sum_v x_{j,v}^2}}$$

L'évaluation que nous proposons consiste en des séries de 10 exécutions des algorithmes OKM et k-moyennes dans des conditions initiales identiques, sur un sous-ensemble de 300 documents pour un nombre de classes variant de 5 à 30. On appelle *association issue de* $\mathcal{R}$, une paire d'objets appartenant à une même classe de $\mathcal{R}$; on dira de plus que cette association est correcte si ces deux objets contiennent au moins une étiquette de catégorie en commun dans la classification préétablie. Chaque partition ou recouvrement $\mathcal{R}$ est alors évalué relativement au nombre d'associations de $\mathcal{R}$ (noté n_a), qui sont correctes (n_b) par rapport au nombre total d'associations correctes attendues (n_c). Nous recourons aux indicateurs traditionnels en Recherche d'Information : la précision, le rappel et l'indice de F_{score} (avec β=1).

$$\text{précision} = \frac{n_b}{n_a} \; ; \; \text{rappel} = \frac{n_b}{n_c} \; ; \; F_{\text{score}}(\beta) = \frac{(\beta^2+1) \times \text{précision} \times \text{rappel}}{\beta^2 \times \text{précision} + \text{rappel}}$$

[7]Un mot désigne en fait un radical : chaîne de caractères résultant d'un processus de segmentation puis de radicalisation (Porter, 1980).

Nous présentons en figure 7 les moyennes comparatives des mesures de F_{score} obtenues sur les recouvrements générés par OKM d'une part et les partitions obtenues avec k-moyennes d'autre part. La diminution observée du F_{score} s'explique dans les deux méthodes par la réduction logique du nombre d'associations (et donc du rappel) lorsque le nombre de classes augmente. Pourtant, si dans le cas des partitions, l'augmentation en précision ne permet pas de compenser la diminution importante du rappel, cette compensation est possible lorsqu'il s'agit de recouvrements du fait d'une perte de rappel atténuée sous l'effet des intersections.

5 Conclusion et Perspectives

Cette étude part du constat suivant : les méthodes de classification actuelles ne sont pas adaptées à la recherche d'une organisation des données en classes recouvrantes ; ce type de schéma de classification devient pourtant indispensable pour appréhender les domaines d'application actuels tels que les documents multimédia ou les données biologiques.

Nous avons alors proposé une première solution visant à rechercher dans l'ensemble des recouvrements possibles des données, un schéma correspondant au mieux à l'organisation de ces données. Cette proposition s'appuie d'une part sur la définition d'un critère objectif permettant d'évaluer les recouvrements, et d'autre part sur une méthode d'exploration de cet espace des possibilités (l'algorithme OKM).

Des expérimentations menées sur deux ensembles de données ont mis en évidence la cohérence globale de la méthode proposée (sur les données Iris) et justifié de l'intérêt d'organiser les données en classes recouvrantes afin d'en conserver une synthèse riche en informations (sur les données Reuters). Cependant cette première contribution suggère plusieurs améliorations et perspectives importantes à mener.

Tout d'abord on peut noter que, afin d'assurer la convergence du critère objectif, la méthode d'affectation proposée dans OKM favorise mais ne garantie pas que chaque objet soit affecté uniquement à ses centres les plus proches (cf. figure 3). Si cette situation est en pratique suffisamment rare pour ne pas remettre en cause la cohérence globale du schéma, il conviendra de proposer une solution théorique à ce problème.

Nous serons également amené à confirmer la justification de cette approche en montrant sur des études comparatives plus larges son intérêt par rapport à d'autres méthodes mentionnées dans ce papier, en particulier les algorithmes CBC, POBOC ou encore des algorithmes de classification floue complétés par une étape supplémentaire d'affectation.

Enfin, nous envisageons d'étudier l'intégration d'une pondération différente des descripteurs pour chaque classe en construction (Modha et Spangler, 2003). Cette perspective s'appuie sur l'hypothèse qu'un objet multi-classé doit l'être sur la base de critères différents.

Références

Apté, C., F. Damerau, et S. M. Weiss (1994). Automated learning of decision rules for text categorization. *ACM Trans. Inf. Syst. 12*(3), 233–251.

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.

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 ECAI'04*, Valencia, Spain, pp. 440–444.

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.

Diday, E. (1984). Une représentation visuelle des classes empiétantes : Les pyramides. Technical report, INRIA num.291, Rocquencourt 78150, France.

D.J. Newman, S. Hettich, C. B. et C. Merz (1998). UCI repository of machine learning databases. University of California, Irvine, Dept. of Information and Computer Sciences.

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

Lelu, A. (1994). Clusters and factors : neural algorithms for a novel representation of huge and highly multidimensional data sets. In E. D. Y. L. . al. (Ed.), *New Approaches in Classification and Data Analysis*, Berlin, pp. 241–248. Springer-Verlag.

MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. In *Proceedings of the Fifth Berkeley Symposium on Mathematical statistics and probability*, Volume 1, Berkeley, pp. 281–297. University of California Press.

Modha, D. S. et W. S. Spangler (2003). Feature weighting in k-means clustering. *Mach. Learn. 52*(3), 217–237.

Pantel, P. (2003). Clustering by Committee. Ph.d. dissertation, Department of Computing Science, University of Alberta.

Porter, M. F. (1980). An algorithm for suffix stripping. *Program 14*, 130–137.

Salton, G. et M. J. McGill (1983). *Introduction to Modern Information Retrieval*. McGraw-Hill, Inc.

Sneath, P. H. A. et R. R. Sokal (1973). Numerical Taxonomy - The Principles and Practice of Numerical Classification. San Francisco, W. H. Freeman and Compagny.

Summary

This paper deals with overlapping clustering, a trade off between crisp and fuzzy clustering. This kind of clustering is motivated by recent applications such as clustering multimedia documents. Even if several contributions have been proposed in this field of research, we observe that no theoretical solution exists.

Our study consists in reformulating the partitioning problem for clustering in order to discover a coverage of the data with overlapping clusters of similar objects. The proposed approach combines an objective criterion that evaluates the quality of a coverage with an algorithmic solution that optimizes this criterion.

A first experimentation on a simple well known dataset helps in understanding the global process of the algorithm ; a second evaluation gives a quantitative point of view of the approach on the task of text clustering.