Étude comparative de deux approches de classification recouvrante : MOC vs. OKM

Guillaume Cleuziou et Jacques-Henri Sublemontier

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

Résumé. La classification recouvrante désigne les techniques de regroupements de données en classes pouvant s'intersecter. Particulièrement adaptés à des domaines d'application actuels (e.g. Recherche d'Information, Bioinformatique) quelques modèles théoriques de classification recouvrante ont été proposés très récemment parmi lesquels le modèle MOC (Banerjee et al. (2005a)) utilisant les modèles de mélanges et l'approche OKM (Cleuziou (2007)) consistant à généraliser l'algorithme des k-moyennes. La présente étude vise d'une part à étudier les limites théoriques et pratiques de ces deux modèles, et d'autre part à proposer une formulation de l'approche OKM en terme de modèles de mélanges gaussiens, laissant ainsi entrevoir des perspectives intéressantes quant à la variabilité des schémas de recouvrements envisageables.

1 Introduction

La classification recouvrante (en anglais *overlapping clustering*) constitue un domaine de recherche étudié depuis les années 60 et relancé par des besoins applicatifs dans des domaines importants tels que la Recherche d'Information ou encore la Bioinformatique.

Le but recherché est alors d'extraire une collection de classes recouvrantes à partir d'une population d'individus de telle manière que : chaque individu appartienne à une ou plusieurs classes, les individus d'une même classe soient similaires, et deux individus n'appartenant pas au moins à une classe commune soient dissimilaires. Différentes directions ont été prospectées afin d'obtenir ce type de schéma de classification.

Des modèles hiérarchiques ont été proposés ; Jardine et Sibson (1971) ont permis, en introduisant les k-ultramétriques, d'envisager des structures hiérarchiques (ou pseudo-hiérarchiques) moins contraignantes que les arbres, par exemple des pyramides (Diday (1984)) ou encore des hiérarchies dites "faibles" étudiées par Bertrand et Janowitz (2003) notamment. L'un des avantages de ces modèles est de proposer une interprétation visuelle des classes et de leur organisation. En revanche, ces modèles ne permettent pas de prendre en compte la globalité des schémas de recouvrements possibles ; par exemple Bertrand et Janowitz (2003) montrent que dans une k-hiérarchie faible (le modèle hiérarchique le moins contraignant), "l'intersection de $(k+1)$ classes arbitraires peut être réduite à l'intersection de k de ces classes".

Les approches par partitionnement proposées ont consisté dans un premier temps à déterminer des centres, des axes ou des représentants de classes auxquels les individus sont affectés

relativement à un seuil d'appartenance. Il s'agit des travaux de Rocchio (1966) repris par Dattola (1968) et plus récemment de la méthode des k-moyennes axiales proposée par Lelu (1994). Ces approches, tout comme l'algorithme CBC (Pantel (2003)), sont motivés par le besoin de modèles spécifiques pour traiter les données textuelles mais souffrent d'un problème commun récurrent que constitue la détermination du seuil (similarité ou probabilité d'appartenance) qui décidera de l'affectation des individus aux classes extraites. Cleuziou et al. (2004) proposent alors l'algorithme POBOC pour se dégager de la contrainte du seuil ; l'affectation des individus est effectuée indépendamment d'un seuil fixé a priori, uniquement par l'étude de la distribution de leurs proximités avec l'ensemble des classes.

Les méthodologies de partitionnement mentionnées jusqu'ici s'appuient implicitement sur une hypothèse forte qui considère qu'un "bon" schéma de classification recouvrante (ou recouvrement) peut être obtenu par l'extension[1] d'un "bon" schéma de classification stricte (ou partition). D'autres pourront penser que de façon analogue, un "bon" recouvrement peut être obtenu par restriction[2] d'un "bon" schéma flou. La notion de "bon" schéma restant subjective à ce stade, Cleuziou (2007) propose un critère objectif générique de qualité d'un schéma de classification (recouvrant ou non) et montre que pour ce critère, il n'existe pas toujours une partition optimale qui, par extension, permettrait d'aboutir à un recouvrement optimal.

Les remarques précédentes nous amènent à considérer une nouvelle voie pour les méthodes de classification recouvrante : celle qui consiste à rechercher un "bon" schéma directement dans l'ensemble des recouvrements possibles. Cette démarche a été adoptée par Banerjee et al. (2005a) et par Cleuziou (2007) dans les algorithmes MOC et OKM respectivement. MOC (*Model-Based Overlapping Clustering*) peut être considéré de façon simplifiée comme une généralisation de la méthode EM (Dempster et al. (1977)) pour la classification recouvrante ; cette approche que nous détaillerons s'appuie en effet sur les modèles de mélanges qui s'avèrent être très performants pour les problématiques de classification stricte. OKM (*Overlapping-k-Means*) est une généralisation de l'algorithme bien connu des k-moyennes (MacQueen (1967)) qui allie simplicité et rapidité pour traiter des problèmes concrets de manière efficace.

En notant que la variante classificatoire (CEM) de EM se ramène, sous certaines conditions restrictives (lois normales, variances sphériques et égales, proportions égales) à l'algorithme des k-moyennes (Celleux et Govaert (1992)), il nous a semblé indispensable d'étudier les analogies entre les deux approches recouvrantes MOC et OKM. L'objectif de cet article est alors de proposer une formulation de OKM en terme de modèles de mélanges puis de la comparer théoriquement et expérimentalement à MOC.

L'article est organisé comme suit : les deux prochaines sections sont dédiées à la présentation des deux approches MOC et OKM respectivement ainsi qu'à la reformulation de OKM permettant une comparaison analytique des deux modélisations. La section 4 présente une discussion sur les modèles permettant d'identifier leurs principales différences et leurs conséquences prévisibles sur les schémas de classification. Cette section sera illustrée par quelques expérimentations sur des données textuelles et biologiques et sera suivie d'une synthèse de l'étude permettant de dégager les principales pistes de réflexions à mener.

[1] On entend par extension, le fait d'effectuer des affectations supplémentaires à un schéma initialement strict.

[2] On entend par restriction le fait de décider de l'affectation des individus, ce qui nous ramène au problème du seuil.

2 Le modèle MOC

Banerjee et al. (2005a) ont proposé récemment un modèle général pour le problème de classification recouvrante en utilisant les modèles de mélanges. Leur proposition s'appuie sur les modèles probabilistes relationnels ou PRM (Friedman et al. (1999)) d'une part, et sur des hypothèses de génération probabiliste des observations d'autre part. Nous décrivons l'essentiel de la méthode dans cette section en laissant le soin au lecteur de se reporter aux références citées pour en obtenir une présentation approfondie.

2.1 Un modèle inspiré de la BioInfo

Le modèle MOC (*Model-based Overlapping Clustering*) peut être vu comme une instanciation du modèle PRM permettant de modéliser les relations entre des gènes, des processus et des valeurs d'expressions de ces gènes mesurées sur des puces ADN. Cette instanciation repose sur l'hypothèse que le niveau d'expression d'un gène (observé dans une certaine condition expérimentale) dépend des processus auxquels le gène participe et de leur niveau d'influence (dans cette même condition expérimentale).

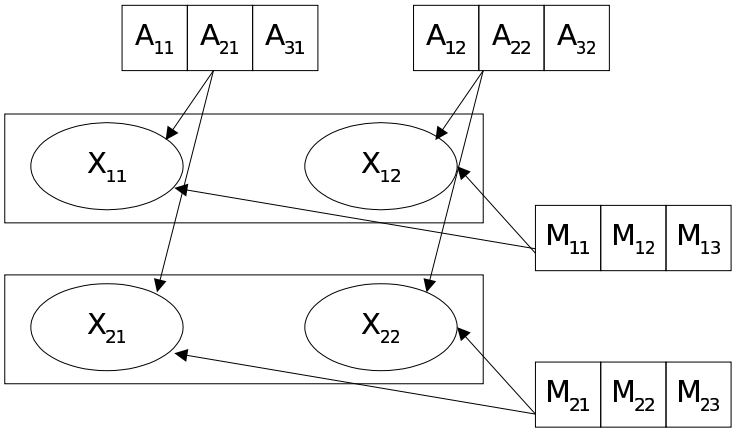

FIG. 1 – *Instanciation du modèle PRM.*

La figure 1 illustre l'instanciation du modèle PRM permettant de modéliser le fait que l'expression X_{ij} d'un gène X_i dans une condition j dépend :

- des niveaux d'influence $\{A_{hj}\}_h$ des processus $\{A_h\}$ (dans la condition j),
- de la participation M_{ih} (ou non) du gène i à un processus A_h.

Dans cet exemple on dispose de 2 gènes (données), 2 conditions (dimensions), 3 processus (classes).

Sous certaines hypothèses de distribution des observations $\{X_{ij}\}$, la détermination des paramètres $\{A_{hj}\}$ et $\{M_{ih}\}$ s'apparente à un problème de classification où les processus s'identifient aux classes ; plus précisément à un problème de classification recouvrante puisque un gène peut participer naturellement à plusieurs processus différents (donc appartenir à plusieurs classes).

2.2 Hypothèses de distribution et modèles associés

Un premier modèle classique consiste à faire l'hypothèse que les observations X_{ij} suivent des lois normales, de variances constantes σ. Selon le modèle général présenté précédemment, pour un nombre fixé de processus, les moyennes des gaussiennes associées sont déterminées par la somme des activités A_{hj} des processus auxquels X_i participe :

$$p(X_{ij}|M_i, A) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(X_{ij} - (MA)_{ij})^2}{2\sigma^2}\right) \quad (1)$$

Le problème de classification sera alors résolu par la recherche des paramètres M et A maximisant la vraisemblance du modèle. Segal et al. (2003) montrent que sous certaines conditions d'indépendance (entre M et A) et d'indépendance conditionnelle (de X_{ij} étant donnés M_i et $A_{.j}$), le problème revient alors à minimiser l'expression (2) ci-dessous par un algorithme du type EM (Dempster et al. (1977)).

$$\frac{1}{2\sigma^2}||X - MA||^2 - \log p(M) \quad (2)$$

Banerjee et al. (2005a) généralisent ce modèle aux familles de distributions exponentielles en s'appuyant sur le fait qu'il existe une bijection entre les distributions exponentielles et les divergences de Bregman (Banerjee et al. (2005b)). Ainsi, quelque soit la loi exponentielle considérée, la distribution des observations peut s'exprimer par

$$p(X_{ij}|M_i, A) \propto \exp\{-d_\phi(X_{ij}, (MA)_{ij})\} \quad (3)$$

avec d_ϕ la divergence de Bregman associée à la densité exponentielle choisie, en particulier :

- la distance euclidienne (élevée au carré) pour des densités Gaussiennes,
- la I-divergence[3] pour des densités de Poisson.

Maximiser la (log-)vraisemblance d'un tel modèle revient alors à minimiser l'expression générale (4).

$$\sum_{ij} d_\phi(X_{ij}, (MA)_{ij}) - \sum_{i,h} \log p(M_{ih}) \quad (4)$$

2.3 Algorithme de résolution

Banerjee et al. (2005a) complètent le modèle de classification MOC par une heuristique générale de minimisation du critère (4) qui consiste à itérer trois étapes de mises à jour : mise à jour des coefficients de mélange $p(M_{ih})$ (notés α_{ih} dans la suite), mise à jour de la matrice des appartenances M et mise à jour de la matrice A dite "matrice d'activité" en référence à l'inspiration bioinformatique de la méthode.

Si la mise à jour des α_{ih} peut être directement déduite des valeurs d'appartenance M et n'influence pas les autres paramètres, la mise à jour de M et de A nécessite un traitement plus complexe.

En particulier les auteurs proposent l'algorithme dynamicM pour résoudre partiellement le problème de la recherche des composantes binaires optimales d'un vecteur M_i représentant les appartenances d'un individu X_i aux processus (ou classes) A_1,...,A_h. Il s'agit d'une heuristique

[3] Également appelée divergence de Kullback-Leibler.

permettant d'explorer pour chaque M_i un sous-ensemble des 2^k-1 vecteurs M_i possibles pour retenir la solution minimisant le critère ou à défaut conserver les anciennes appartenances.

La mise à jour de A peut être réalisée dans un cas général (pour toute divergence de Bregman) au moyen d'un algorithme de descente de gradient de la forme :

$$A^{new} \leftarrow A - \eta M^T[(X - MA) \circ \phi''(MA)] \tag{5}$$

avec ϕ la fonction identifiant la divergence de Bregman utilisée et η un coefficient d'apprentissage fixé. Dans les cas particulier de divergences simples qui nous intéresseront plus particulièrement dans les expérimentations à venir, le problème de minimisation peut être résolu plus directement.

- pour la distance euclidienne il s'agit d'une minimisation de type moindres carrés résolue par (6) où $M^\dagger$ désigne la pseudo-inverse de M

$$A = M^\dagger X \tag{6}$$

- pour la I-divergence, les auteurs s'appuient sur des techniques de factorisation de matrices non-négatives pour aboutir à la règle de mise à jour suivante

$$A_{hj}^{new} = A_{hj} \frac{\sum_i M_{ih} X_{ij}/(MA)_{ij}}{\sum_i M_{ih}} \tag{7}$$

Dans cette dernière variante, on peut observer que la règle de mise à jour proposée correspond à une simplification de la règle plus générale (8) réétudiée récemment par Finesso et Spreij (2006) :

$$A_{hj}^{new} = A_{hj} \frac{\sum_i M_{ih} X_{ij}/(MA)_{ij}}{\sum_{ij} M_{ih} A_{hj} X_{ij}/(MA)_{ij}} \tag{8}$$

Dans le cas particulier où chaque individu X_i n'appartient qu'à une seule classe A_h alors $(MA)_{ij} = M_{ih}A_{hj}$; en ajoutant à cela le fait que la I-divergence mesure l'écart entre deux distributions p et q telles que $\sum_j p_j = \sum_j q_j = 1$ la simplification (7) devient en effet possible ($\sum_j X_{ij} = 1$). Cependant la méthode MOC s'intéressant justement aux cas où chaque individu peut appartenir à plusieurs classes, cette simplification devient fausse ; c'est la raison pour laquelle nous proposerons de conserver la règle de mise à jour originelle (8) afin d'assurer la décroissance du critère (4).

Les trois étapes de mises à jour que nous venons de présenter permettent d'assurer la décroissance du critère (4) et par conséquent d'améliorer à chaque itération (et après chaque étape de mise à jour) la vraisemblance du modèle probabiliste. L'initialisation du modèle (matrices M et A) est effectuée au moyen d'une étape de partitionnement (k-moyennes). Enfin, de façon assez classique l'algorithme MOC itère le processus de mises à jour un nombre maximum de fois ou jusqu'à observer une variation epsilonesque du critère objectif.

3 Le modèle OKM

L'algorithme des k-moyennes présente un modèle théorique simple et intuitif, facilement appréhendé par les praticiens de domaines d'application variés qui, de surcroît, apprécient généralement l'efficacité de cette méthode en terme de rapidité et de qualité des classes obtenues.

L'approche OKM, proposée par Cleuziou (2007), est le résultat d'une volonté de répondre de manière pragmatique aux besoins applicatifs actuels, en proposant d'étendre l'algorithme des k-moyennes à la recherche de recouvrements plutôt que de partitions.

3.1 Critère objectif et heuristique d'optimisation

Le critère des moindres carrés sur lequel repose l'algorithme des k-moyennes est une formalisation fidèle de l'objectif visé par les méthodes de partitionnement à savoir faire en sorte que deux individus d'une même classe soient similaires et deux individus de classes différentes soient dissimilaires. Comme nous l'avons mentionné en introduction, on peut assez naturellement considérer qu'un bon recouvrement sera caractérisé par :

- des individus similaires lorsqu'ils appartiennent plutôt aux mêmes classes,
- des individus dissimilaires lorsqu'ils appartiennent plutôt à des classes différentes.

Le critère utilisé dans OKM pour formaliser les caractéristiques précédentes introduit la notion d'image (que l'on notera $\psi(X_i)$) d'un individu X_i dans une classification $I_1, \ldots, I_k$. L'image de X_i correspond à un point, dans l'espace de représentation initial, et représentatif des classes auxquelles X_i appartient. Par exemple, en reprenant les notations utilisées précédemment, on pourra définir l'image de X_i dans un espace euclidien $(\mathbb{R}^m, d)$ par

$$\psi_j(X_i) = \frac{\sum_h M_{ih} A_{hj}}{\sum_h M_{ih}} \tag{9}$$

Dans (9), $\psi_j(X_i)$ désigne la $j^{\text{ième}}$ composante de l'image, $M_{ih} \in \{0, 1\}$ l'appartenance de X_i à la classe I_h et A_h correspond ici à la position du centre de la classe I_h. Le critère objectif que l'on cherchera à minimiser, évalue la qualité d'un recouvrement par la variance entre les individus et leur image dans la classification :

$$\sum_i d^2(X_i, \psi(X_i)) \tag{10}$$

Pour $\psi(.)$ bien choisie, on peut noter que ce critère se ramène exactement au critère des moindres carrés lorsque l'on oblige chaque individu à n'appartenir qu'à une seule classe ($\sum_h M_{i,h} = 1$).

L'heuristique de minimisation du critère objectif (10) dans OKM s'apparente à l'algorithme des k-moyennes. Après une étape d'initialisation aléatoire des centres de classes A, deux étapes de mises à jour (de M puis de A) sont itérées jusqu'à la vérification d'un critère d'arrêt portant sur le nombre d'itérations ou la variation du critère objectif.

Mise à jour de M. Pour chaque individu X_i, la mise à jour de M_i est réalisée en considérant qu'un individu ne doit appartenir qu'aux classes dont il est le plus proche au sens de la métrique choisie. Ce principe guide la construction du nouveau vecteur d'appartenance : initialisation avec affectation au plus proche centre de classe puis ajout de nouvelles affectations dans l'ordre de proximité des centres $\{A_h\}_h$ avec X_i tant que le critère objectif s'en trouve amélioré ; le nouveau vecteur ainsi obtenu ne remplaçant le vecteur M_i initial que s'il permet d'améliorer le critère objectif.

Mise à jour de A. Cleuziou (2007) montre que pour la distance euclidienne on peut définir localement le nouveau centre A_h de la classe I_h de façon optimale au sens du critère objectif.

Une heuristique d'optimisation globale de A consistera donc à itérer les optimisations locales, en parcourant l'ensemble des classes plusieurs fois[4], jusqu'à aboutir à un point fixe.

Pour compléter cette description, nous présentons dans le tableau 1 les différentes variantes de OKM en fonction des métriques considérées.

Métrique	Espace de représentation	Image ψ	Mise à jour de A_h
distance euclidienne	$\mathbb{R}^m$	Moyenne des centres (cf. (9))	centre de gravité pondéré (Cleuziou (2007))
I-divergence	$[0,1]^m$ avec $\sum_j X_{ij} = 1$	Moyenne des centres (cf. (9))	Règle multiplicative de Finesso&Spreij (cf. (8))
cosinus	$[0,1]^m$ avec $\sum_j X_{ij}^2 = 1$	Moyenne quadratique des centres	Non déterminé

TAB. 1 – *Variantes de* OKM *par métrique.*

3.2 Reformulation de OKM

En reprenant l'inspiration bioinformatique du modèle MOC, on considère qu'une observation X_{ij} est le résultat d'une certaine combinaison des processus auxquels l'individu X_i participe. Plutôt que de choisir comme combinaison l'addition des processus, nous en choisissons la moyenne. Ainsi sous l'hypothèse d'une distribution gaussienne (de variance constante σ) des valeurs X_{ij} nous obtenons

$$p(X_{ij}|M_i, A) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(X_{ij} - \mu_i.(MA)_{ij})^2}{2\sigma^2}\right) \tag{11}$$

avec μ un vecteur totalement défini par M tel que $\mu_i = 1/\sum_h M_{ih}$.

Dans un processus de classification, l'objectif consiste par exemple à rechercher les paramètres M (matrice binaire) et A (matrice réelle) maximisant la (log-)vraisemblance des paramètres étant données les observations : $\log \mathcal{L}(M, A|X) = \log p(X|M, A)$. Le modèle MOC fait l'hypothèse qu'il y a indépendance des observations X_{ij} conditionnellement aux M_i et $A_{.j}$. Sous ces mêmes hypothèses, la vraisemblance du modèle peut se décomposer ainsi

$$\mathcal{L}(M, A|X) = p(X|M, A) = \prod_{i,j} p(X_{i,j}|M_i, A_{.j})$$

En considérant à présent la log-vraisemblance et en introduisant l'hypothèse de distribution gaussienne (11), on note que

$\max_{M,A} \log \mathcal{L}(M, A|X) \equiv \max_{M,A} \left[\log \prod_{i,j} p(X_{i,j}|M_i, A_{.j})\right]$

$$\equiv \max_{M,A} \left[-\frac{1}{2\sigma^2} \sum_{i,j} (X_{ij} - \mu_i(MA)_{ij})^2\right] \equiv \min_{M,A} \left[\frac{1}{2\sigma^2} ||X - \mu^T I.MA||^2\right] \tag{12}$$

[4]En pratique un seul parcours de l'ensemble des classes suffit pour approcher le point fixe.

En observant que $\mu_i(MA)_{ij} = \psi(X_i)$ avec $\psi(X_i)$ l'image de X_i telle que définie en (9), on montre que minimiser le terme $||X - \mu^T I.MA||^2 = \sum_i \sum_j (X_{ij} - \mu_i(MA)_{ij})^2$ équivaut à minimiser $\sum_i d^2(X_i, \psi(X_i))$ (avec d la distance euclidienne) utilisé dans OKM (cf. (10)).

Nous avons donc démontré que l'approche OKM, dans sa version initiale utilisant la distance euclidienne, peut être réécrite comme un modèle de mélanges recouvrant faisant l'hypothèse que chacune des observations suit une lois normale dont la moyenne correspond à la moyenne des processus (ou classes) auxquelles l'individu participe.

4 Discussion et analyse comparative des modèles

Nous mènerons la discussion en deux temps : nous relèverons dans un premier temps les différences majeures et les limites théoriques des deux modèles MOC et OKM ; dans un second temps nous proposerons une étude comparative expérimentale des deux approches.

4.1 Discussion sur les modèles

Nous avons choisi de comparer analytiquement les deux modèles en les exprimant tous les deux en terme de modèles de mélanges recouvrants, plutôt que simplement comme des techniques de réallocation dynamique minimisant un critère d'inertie. On peut ainsi envisager plus facilement d'extraire ultérieurement des classes de formes, volumes et orientations variées. Cependant en l'état actuel des modèles, ces variations ne sont pas permises (hypothèse des variances toutes identiques) et les deux formalismes sont strictement équivalents.

Ce qui différencie les modèles MOC et OKM concerne la méthode de combinaison des "processus" déterminant les paramètres de la distribution d'une observation : pour une distribution exponentielle, c'est la moyenne de la distribution qui résulte de cette combinaison.

- Le modèle MOC propose une combinaison par addition, justifiée par le modèle bioinformatique sous-jacent qui suppose que l'expression observée d'un gène est le résultat de l'"addition" des processus dans lesquels ce gène intervient.
- Le modèle OKM utilise la notion d'image qui correspond effectivement à une combinaison des représentants des classes auxquelles l'individu appartient. La définition de l'image dépend de la métrique considérée et s'exprime comme une moyenne plutôt qu'une somme : moyenne arithmétique ou quadratique par exemple (cf. tableau 1)).

Ce choix de combinaison n'est pas anodin et peut avoir des conséquences notables sur la *validité* théorique du modèle d'une part, sur sa *sensibilité* aux données d'autre part.

La *validité* théorique du modèle peut être remise en cause lorsque la combinaison utilisée n'est pas un endomorphisme car le modèle implique de considérer la distance entre chaque individu et la combinaison associée. Par exemple, la I-divergence permet de comparer deux distributions p et q dans $[0, 1]^m$ telles que $\sum_j p_j = \sum_j q_j = 1$; si on peut montrer que l'image $\psi(X_i)$ utilisée dans OKM reste effectivement dans l'espace des distributions ($\psi(X_i) \in [0, 1]^m$ et $\sum_j \psi_j(X_i) = 1$), la combinaison $(MA)_i$ utilisée dans MOC ne vérifie pas les caractéristiques d'une distribution et l'expression $d_\phi(X_i, (MA)_i)$ utilisée dans (3) n'a pas de sens.

La combinaison peut également jouer un rôle important dans la *sensibilité* du modèle, notamment dans la répartition des données pour certaines hypothèses de lois de mélange. Prenons un exemple jouet composé de quatre individus décrits dans $\mathbb{R}$ $\{X_1 = (1.0), X_2 = (4.0), X_3 = (5.0), X_6 = (6.0)\}$ avec les hypothèses suivantes : distributions gaussiennes des observations résultant de deux processus/classes ($k = 2$). La figure 2 illustre la configuration des individus

à classer et, pour chacune des deux approches MOC et OKM, la projection des paramètres A dans le même espace de description après optimisation.

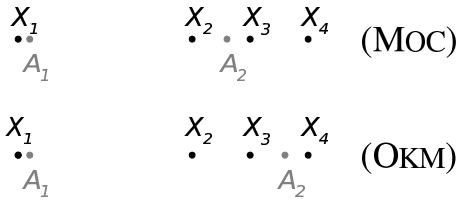

FIG. 2 – *Observation du paramètre A selon l'approche.*

La classification finale retournée par OKM s'obtient simplement en affectant chaque individu X_i à I_1, I_2 ou aux deux classes selon que l'individu est plus proche de A_1, A_2 ou $(\mathbf{A_1} + \mathbf{A_2})/\mathbf{2}$ respectivement ; sur l'exemple OKM retournera donc les classes $I_1 = \{X_1, \mathbf{X_2}\}$ et $I_2 = \{\mathbf{X_2}, X_3, X_4\}$. Dans l'approche MOC on obtient cette fois la classification finale en comparant les distances de X_i avec A_1, A_2 et $\mathbf{A_1} + \mathbf{A_2}$, aboutissant ainsi aux classes $I_1 = \{X_1, \mathbf{X}_4\}$ et $I_2 = \{X_2, X_3, \mathbf{X}_4\}$ dont l'intersection est totalement injustifiée. Il suffirait sur cet exemple de recentrer les individus en zéro tout en conservant les distances entre individus inchangées pour obtenir avec MOC les mêmes classes que pour OKM.

Cet exemple nous a donc permis de mettre en évidence que MOC peut être sensible aux translations de données, en particulier sous des hypothèses de distributions gaussiennes[5]. Cette limitation peut être observée sur un exemple réel de données d'expressions de gènes en bioinformatique[6] (figures 3 et 4). Même si, sur cet exemple, aucune organisation en classes n'est

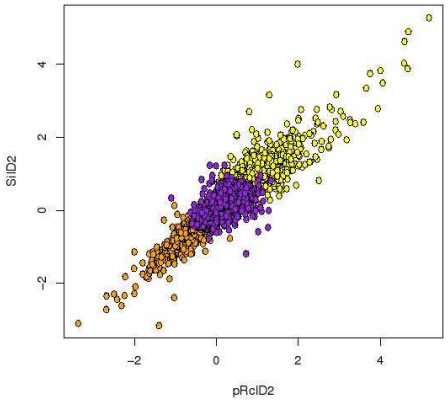

FIG. 3 – MOC *sur les données initiales.*

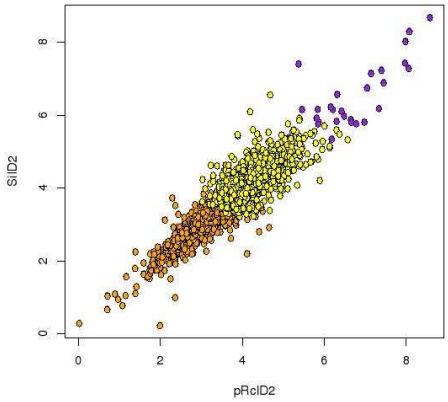

FIG. 4 – MOC *sur données translatées.*

observable, on remarque que l'intersection des deux classes (points violets) est cohérente sur les données initiales (figure 3) et de nouveau injustifiées sur les données translatées (figure 4). Il est donc important de préciser qu'en pratique, MOC produit des résultats intéressants sur des

[5]On peut montrer que cette sensibilité est annihilée pour d'autres types de distributions telles que la I-divergence grâce aux contraintes vérifiées par les individus (*e.g.* $\sum_j X_{ij} = 1$).

[6]Cette expérience a été réalisée dans le cadre du projet ANR/ARA Masse de données "Genomic data to Graph Structure" : http ://gd2gs.ibisc.univ-evry.fr/

données d'expressions de gènes, précisément parceque ces données sont par nature centrées globalement autour de zéro[7].

4.2 Comparaisons expérimentales

Pour achever la comparaison des deux approches de classification recouvrante étudiées, nous présentons les résultats obtenus par MOC et OKM sur une expérience de classification de documents textuels qui correspond à un domaine d'applications cible. L'étude est menée sur un sous-ensemble des documents du corpus Reuters, utilisé et présenté de façon détaillée par Cleuziou (2007). Ce corpus présente l'intérêt que chaque document possède une ou plusieurs étiquettes de classes. Nous n'utilisons pas cette information dans le processus de classification mais seulement pour évaluer la qualité des schémas de classification recouvrants générés par les méthodes. L'évaluation opérée consiste à mesurer l'écart entre les associations de documents connues (pré-étiquetage) et les associations effectivement retrouvées dans la classification, en utilisant les indices classiques de précision, rappel et F-mesure[8].

Nb.	F-mesure			Précision			Rappel		
classes	k-moy.	OKM	MOC	k-moy.	OKM	MOC	k-moy.	OKM	MOC
k=2	0.39	**0.42**	0.41	**0.27**	**0.27**	0.26	0.77	**0.95**	0.93
k=5	0.34	**0.41**	0.40	**0.27**	**0.27**	**0.27**	0.46	**0.89**	0.77
k =10	0.32	**0.41**	0.39	**0.30**	0.28	0.28	0.35	**0.83**	0.65
k =15	0.30	**0.40**	0.38	**0.33**	0.28	0.29	0.27	**0.76**	0.54
k =20	0.27	**0.40**	0.37	**0.34**	0.29	0.31	0.23	**0.66**	0.46

TAB. 2 – *Classification des données Reuters en utilisant la I-divergence.*

Nb.	F-mesure			Précision			Rappel		
classes	k-moy.	OKM	MOC	k-moy.	OKM	MOC	k-moy.	OKM	MOC
k=2	0.38	**0.39**	0.38	0.24	**0.25**	0.24	0.84	**0.97**	0.91
k=5	0.35	**0.40**	0.37	0.23	**0.26**	0.24	0.67	**0.90**	0.76
k =10	0.29	**0.40**	0.34	0.23	**0.26**	0.25	0.42	**0.84**	0.54
k =15	0.29	**0.39**	0.35	0.25	**0.26**	**0.26**	0.37	**0.80**	0.55
k =20	0.26	**0.38**	0.36	0.27	0.26	**0.29**	0.26	**0.69**	0.50

TAB. 3 – *Classification des données Reuters sous l'hypothèse de distributions gaussiennes.*

Nous présentons les résultats de l'évaluation des classifications en utilisant la I-divergence d'une part (tableau 2) et des distributions gaussiennes ou métrique euclidienne d'autre part (tableau 3). Les valeurs reportées dans les tableaux correspondent à des moyennes de 10 exécutions de chaque méthode dans des conditions initiales identiques.

Tout d'abord, nous retrouvons un phénomène connu en Recherche d'Information qui est que la I-divergence donne de meilleurs résultats que la métrique euclidienne pour la classification de documents. Ceci est notamment dû au fait que la I-divergence compare des distributions

[7] Certains gènes s'exprimant positivement, d'autres négativement.

[8] Cette technique d'évaluation des méthodes est utilisée par Cleuziou (2007) et par Banerjee et al. (2005a).

de mots plutôt que des vecteurs de fréquences, réduisant ainsi les effets liés aux variations de tailles entre documents. Le second résultat attendu est de constater que la précision augmente et que le rappel diminue quand on augmente le nombre de classes ; ceci s'explique par le fait que lorsque le nombre de classes augmente, le nombre de paires de documents associés diminue automatiquement.

Enfin, en comparant les résultats obtenus par les approches MOC et OKM on observe que pour les deux métriques utilisées, OKM génère d'avantage de recouvrements que MOC[9], ce qui se traduit par un taux de rappel plus élevé sans pour autant entraîner un fléchissement de la précision (qui reste d'ailleurs plutôt à l'avantage de OKM). Il semblerait donc, au regard de la globalité des résultats de cette expérience que : les deux approches de classification recouvrante étudiées permettent de générer des recouvrements pertinents (comparaison avec l'algorithme des k-moyennes) et que le gain obtenu soit plus net dans le cas de l'approche OKM notamment en utilisant la I-divergence. Cette dernière remarque corrobore les limites théoriques énoncées précédemment concernant le modèle MOC, et en particulier sous les hypothèses de distributions gaussiennes.

5 Conclusion et prespectives

Dans cet article nous avons étudié deux approches de classification recouvrante : l'approche MOC (Banerjee et al. (2005a)) et l'approche OKM (Cleuziou (2007)). En proposant une formalisation de OKM en terme de mélange de lois, nous avons pu constater les fortes analogies qui existent entre les deux approches. Nous avons détaillé leurs différences fondamentales et relevé quelques limites théoriques concernant le modèle MOC. Ces limites sont susceptibles de produire des résultats incohérents que nous avons observés expérimentalement.

Nous proposerons par la suite de confirmer les premières observations expérimentales sur d'autres corpus et d'autres domaines d'application. Nous poursuivrons l'extension des modèles traditionnels de classification dans le but d'extraire des recouvrements avec des classes de formes, volumes et orientations variées. Enfin nous travaillerons sur une variante sphérique du modèle OKM utilisant la mesure du cosinus, particulièrement adaptée au traitement des documents textuels.

Références

Banerjee, A., C. Krumpelman, J. Ghosh, S. Basu, et R. J. Mooney (2005a). 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.

Banerjee, A., S. Merugu, I. Dhillon, et J. Ghosh (2005b). Clustering with bregman divergences. *J. Mach. Learn. Res. 6*, 1705–1749.

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.

Celleux, G. et G. Govaert (1992). A classification EM algorithm for clustering and two stochastic versions. *Computational Statistics and Data Analysis 14*(3), 315–332.

[9] Sans détailler d'avantage ce point, il est vraisemblable que ce phénomène soit lié à l'initialisation *via* une exécution de l'algorithme k-moyenne dans l'approche MOC

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

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.

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.

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.

Finesso, L. et P. Spreij (2006). Nonnegative Matrix Factorization and I-Divergence Alternating Minimization. *Linear Algebra and its Applications 416*, 270–287.

Friedman, N., L. Getoor, D. Koller, et A. Pfeffer (1999). Learning probabilistic relational models. In *IJCAI*, pp. 1300–1309.

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

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.

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

Rocchio, J. (1966). Document retrieval systems - optimization and evaluation. Ph.d. thesis, harvard university, Report ISR-10 to National Science Foundation, Harvard Computation Laboratory.

Segal, E., A. Battle, et D. Koller (2003). Decomposing gene expression into cellular processes. *Pac Symp Biocomput*, 89–100.

Summary

This paper deals with overlapping clustering methodologies which consist in organizing data into classes with intersections. This kind of clustering scheem is suitable for important fields of application such as Information Retrieval or Bioinformatics. Two approaches have been recently proposed by Banerjee et al. (2005a) (MOC) and Cleuziou (2007) (OKM). These two approaches are compared on theoretical and experimental points of view in order to prospect for new general overlapping clustering models.