# Identification de classes non-disjointes ayants des densités différentes

Hela Masmoudi*, Chiheb-Eddine Ben N'Cir**, Nadia Essoussi***

*LARODEC, ISG Tunis, Université de Tunis
hela.masmoudi@gmail.com
**LARODEC, ISG Tunis, Université de Tunis
chiheb.benncir@isg.rnu.tn
***LARODEC, ISG Tunis, Université de Tunis
nadia.essoussi@isg.rnu.tn

**Résumé.** La classification recouvrante correspond à un enjeu important en classification non-supervisée en permettant à une observation d'appartenir à plusieurs clusters. Plusieurs méthodes ont été proposées pour faire face à cette problématique en utilisant plusieurs approches usuelles de classification. Cependant, malgré l'efficacité de ces méthodes à déterminer des groupes non-disjoints, elles échouent lorsque les données comportent des groupes de densités différentes car elles ignorent la densité locale de chaque groupe et ne considèrent que la distance Euclidienne entres les observations. Afin de détecter des groupes non-disjoints de densités différentes, nous proposons deux méthodes de classification intégrant la variation de densité des différentes classes dans le processus de classification. Des expériences réalisées sur des ensembles de données artificielles montrent que les méthodes proposées permettent d'obtenir de meilleures performances lorsque les données contiennent des groupes de densités différentes.

## 1 Introduction

La classification non-supervisée a pour but de regrouper les observations proches dans un même groupe, tandis que les observations éloignées doivent être affectées à des groupes différents. Cette définition pourrait être insuffisante dans de nombreuses applications de regroupement dans lesquelles un objet peut appartenir à la fois à plusieurs classes. Ce type de problématique est appelée classification recouvrante ou encore classification non-exclusive. Plusieurs applications réelles nécessitent d'utiliser ce type de schéma de classification tels que le regroupement de documents où chaque document peut aborder plusieurs thèmes, la classification de vidéos où un film peut avoir différents genres (Snoek et al., 2006), la détection d'émotions où un morceau de musique peut évoquer plusieurs émotions distinctes (Wieczorkowska et al., 2006).

Afin de produire des classes non-exclusives, divers méthodes ont été proposées utilisants des approches hiérarchique (Diday, 1984), de partitionnement (Fu et Banerjee, 2008; Banerjee

et al., 2005; Heller et Ghahramani, 2007), de corrélation (Bonchi et al., 2011) et de la théorie des graphes (Fellows et al., 2009, 2011). Nos travaux se limitent à l'étude des méthodes recouvrantes basées sur le partitionnement. Les méthodes de partitionnement existantes utilisent des modèles de mélange de lois (Fu et Banerjee, 2008; Banerjee et al., 2005; Heller et Ghahramani, 2007) ou bien utilisent l'approche k-moyennes (Cleuziou, 2008). Des exemples de ces méthodes sont OKM (Cleuziou, 2008) et Parameterized R-OKM (Ben N'Cir et al., 2013). Ces dernière méthodes ne considèrent que la distance euclidienne entre chaque observation et ses représentants. En outre, la densité des observations dans un groupe pourrait être nettement différente des autres groupes dans l'ensemble de données. La métrique euclidienne utilisée évalue seulement la distance Euclidienne entre l'observation et le représentant, ou la combinaison de représentants, de clusters. Elle ne tient pas compte de la variation de la distance globale pour toutes les observations dans un groupe.

Récemment, une nouvelle mesure de distance a été proposée par Tsai et Lin (2011). Elle intègre la variation de distance au sein d'un même cluster et sert à régulariser la distance entre un objet et le représentant de cluster relativement à la densité interne du cluster en question. Nous proposons dans ce qui suit d'adapter cette nouvelle mesure pour détecter des groupes non-disjoints avec des densités différentes.

Cet article est organisé comme suit : la Section 2 décrit deux méthodes existantes de classification recouvrante à savoir OKM et Parameterized R-OKM. Ensuite, la Section 3 présente le problème d'identification des groupes avec une densité différente. La Section 4 décrit les méthodes proposées OKM-$\sigma$ et Parameterized R-OKM-$\sigma$ tandis que la Section 5 illustre les expériences réalisées sur des ensembles de données artificielles. La conclusion et les perspectives feront l'objet de la Section 6.

# 2 La classification recouvrante

Nous décrivons dans cette partie deux méthodes existantes de classification recouvrantes basées sur l'algorithme k-moyennes. Ces méthodes généralisent k-moyennes pour produire des recouvrements de classes.

## 2.1 Overlapping K-Means

La méthode OKM cherche des recouvrements optimaux plutôt que des partitions optimales, étant donné un ensemble d'objets à classifier $X = \{x_i\}_{i=1}^{N}$ avec $x_i \in \mathbb{R}^d$ et $N$ le nombre d'objets, OKM recherche un $k$ recouvrement de telle sorte que la fonction objective suivante soit optimisée :

$$J(\{\pi_k\}_{k=1}^{K}) = \sum_{i=1}^{N} ||x_i - (\overline{x_i})||^2 \tag{1}$$

où $\overline{x_i}$ désigne l'image de $x_i$ définie par la combinaison des centres des clusters auxquels $x_i$ appartient :

$$\overline{x_i} = \sum_{k \in \Pi_i} \frac{c_k}{|\Pi_i|} \tag{2}$$

avec $\Pi_i$ l'ensemble des affectations de l'objet $x_i$ aux différents clusters, c'est-à-dire les clusters auxquels $x_i$ appartient et $c_k$ correspond au représentant du cluster $k$. Le critère $J$ de la fonction objective généralise le critère des moindres carrés utilisé dans la méthode k-moyennes. Pour minimiser ce critère, deux étapes principales sont exécutées itérativement :

1. la mise à jour des représentants de classes $(C)$ puis.
2. l'affectation des objets à ces représentants $(\Pi)$.

les conditions d'arrêt de la méthode reposent sur plusieurs critères à savoir le nombre d'itérations maximales et le seuil minimal d'amélioration de la fonction objective entre deux itérations.

OKM produit des classes non-disjointes avec des zones de recouvrement trop larges. Cependant, les groupes ayants des zones de recouvrement larges ne sont pas appropriés à la plupart des applications réelles. Pour résoudre ce problème, une méthode récente, référencée Parameterized R-OKM (Ben N'Cir et al., 2013), propose un nouveau modèle qui produit des clusters non-disjoints avec possibilité de contrôle sur les zones de recouvrements.

### 2.2 Parameterized R-OKM

Afin de contrôler la taille des recouvrements entre les classes, Parameterized R-OKM restreint l'attribution d'une observation à plusieurs groupes en fonction du nombre d'affectations $|\Pi_i|$. Parameterized R-OKM est basée sur la minimisation du critère objetif suivant :

$$J(\{\pi_k\}_{k=1}^{K}) = \sum_{i=1}^{N} |\Pi_i|^{\alpha}.||x_i - (\overline{x_i})||^2 \tag{3}$$

où $\alpha \geq 0$ un paramètre fixé par l'utilisateur permettant de contrôler la taille des intersections. Lorsque $\alpha$ devient grand, Parameterized R-OKM construit des groupes avec des intersections plus réduites. Quand $\alpha = 0$, Parameterized R-OKM coïncide exactement avec OKM. Afin de minimiser le critère objectif, Parameterized R-OKM utilise les mêmes étapes que OKM.

## 3 Le problème des clusters avec des densités non-uniformes

Les méthodes OKM et Parameterized R-OKM ne considèrent que la distance Euclidienne entre chaque objet et le représentant du groupe. La métrique euclidienne utilisée ne tient pas compte de la distribution des objets autour du représentant du groupe. Par conséquent, OKM et Parameterized R-OKM échouent lorsque les classes ont des densités différentes conduisant à des groupes avec des intersections larges . La Figure 1 présente le problème des clusters avec des densités différentes où le groupe rouge a une faible densité par rapport au groupe bleu. Dans les applications réelles de classification recouvrante, la densité des objets dans un groupe peut être différente des densités des autres groupes contenus dans l'ensemble de données.

En fait, les méthodes basées sur k-moyennes ne permettent généralement pas de détecter des groupes de densités différentes. Cependant, certains travaux ont proposé des alternatives pour résoudre ce problème (Govaert, 1975). Récemment, une nouvelle alternative proposant

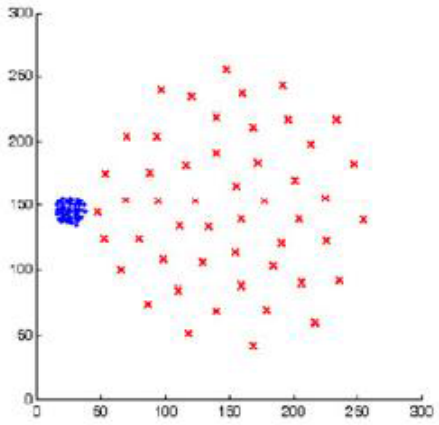


FIG. 1 – *Deux classes ayant le même nombre d'objets et des densités différentes : la classe rouge est caractérisée par une densité faible par rapport à la densité de la classe bleue.*

une mesure, intégrant la variation de densité entre les classes, à été introduite par Tsai et Lin (2011). Cette technique à été utilisée dans k-moyenne floues et a été référencée (FCM-$\sigma$ ).

FCM-$\sigma$ introduit la variation de densité pour chaque groupe de données afin de régulariser la distance entre un objet et le représentant. Cette mesure peut être mieux appliquée sur des données contenant des groupes de densités différentes. La nouvelle métrique de distance entre chaque objet $x_i$ et chaque représentant $c_k$ est définie par :

$$\hat{d}_{ik}^2 = \frac{||x_i - c_k||^2}{\sigma_k}, \tag{4}$$

où $\sigma_k$ représente la moyenne pondérée des distances dans un cluster $k$, définie par

$$\sigma_k = \left\{ \frac{\sum_{i=1}^{N} w_{ik}^p . ||x_i - c_k||^2}{\sum_{i=1}^{N} w_{ik}^p} \right\}^{1/2}, \tag{5}$$

avec $w_{ik} \in [0, 1]$ le degré d'appartenance de l'observation $x_i$ appartenant à la classe $k$ et $p$ un paramètre utilisé pour contrôler le flou. En utilisant la nouvelle distance $\hat{d}_{ik}^2$, FCM-$\sigma$ minimise la fonction objective suivante :

$$J_{FCM-\sigma} = \sum_{k=1}^{K} \sum_{i=1}^{N} w_{ik}^p . \frac{||x_i - c_k||^2}{\sigma_k}, \tag{6}$$

avec $\sum_{k=1}^{K} w_{ik} = 1 \; \forall i = 1, 2, ..., N$.

La mise à jour des centres et la mise à jour des degrés d'appartenance aux différents clusters sont déterminées par :

$$w_{ik} = \frac{1}{\sum_{c=1}^{K} \left( \frac{\hat{d}^2_{ik}}{\hat{d}^2_{ic}} \right)^{1/(p-1)}} \quad \forall i = 1, 2, ..., N \quad et \quad \forall k = 1, 2, ..., K. \tag{7}$$

et

$$c_k = \frac{\sum_{j=1}^{N} w_{jk}^p . x_j}{\sum_{j=1}^{N} w_{jk}^p} \tag{8}$$

L'évaluation de FCM-$\sigma$ sur des données de densités equilibrées montre que cette méthode donne les mêmes résultats que FCM (Tsai et Lin, 2011). Par contre, si les données contiennent des groupes de différentes densités, FCM-$\sigma$ est plus performant que FCM.

# 4 Méthodes Proposées

Dans les domaines d'applications de la classification recouvrante, l'algorithme d'apprentissage doit être capable de détecter des groupes non disjoints avec des densités uniformes et non uniformes. Par conséquent, nous proposons d'étendre OKM et Parameterized R-OKM pour détecter ces types de groupes en introduisant une régularisation de la densité dans la fonction objective de ces méthodes. Les nouvelles méthodes proposées sont désignées OKM-$\sigma$ et Parameterized R-OKM-$\sigma$.

## 4.1 Overlapping k-means-$\sigma$ (OKM-$\sigma$)

Pour tenir compte des différences de densité entre les classes, nous introduisons un facteur de régularisation $\sigma_i$ pour chaque observation $x_i$. Compte tenu de l'ensemble des $N$ observations, OKM-$\sigma$ minimise le critère objectif suivant :

$$\begin{aligned} J(\{\pi_k\}_{k=1}^{K}) &= \sum_{i=1}^{N} \hat{d}^2(x_i, (\overline{x_i})) \\ &= \sum_{i=1}^{N} \frac{\| x_i - (\overline{x_i}) \|^2}{\sigma_i}, \end{aligned} \tag{9}$$

où $\sigma_i$ est la valeur minimale des densités des groupes $\Pi_i$ auxquels l'observation $x_i$ appartient :

$$\sigma_i = Min_{k \in \Pi_i} \{\sigma_k\}, \tag{10}$$

avec $\sigma_k$ le poids local au cluster $k$ qui mesure le degré de déviation des observations contenues dans le cluster $k$ par rapport à leur image respective. Ce poids peut être décrit formellement par :

$$\sigma_k = \left\{ \frac{\sum_{i=1}^{N} P_{ik}.||x_i - (\overline{x_i})||^2}{\sum_{i=1}^{N} P_{ik}} \right\}^{1/2} , \quad (11)$$

où $P_{ik}$ une variable binaire indiquant l'appartenance de l'objet $x_i$ au cluster $k$.

## 4.2 Parameterized R-OKM-$\sigma$

En se basant sur le même principe de régularisation de densité, nous proposons la méthode Parameterized R-OKM-$\sigma$ permettant la régularisation des variances de densité entre les classes. Le nouveau critère objectif proposé est décrit par :

$$\begin{aligned} J(\{\pi_k\}_{k=1}^{K}) &= \sum_{i=1}^{N} |\Pi_i|^{\alpha} \hat{d}^2(x_i, (\overline{x_i})) \\ &= \sum_{i=1}^{N} |\Pi_i|^{\alpha} \frac{\| x_i - (\overline{x_i}) \|^2}{\sigma_i}, \end{aligned} \quad (12)$$

où $\sigma_i$ le facteur de régularisation locale de l'observation $x_i$ décrit de la même manière que dans l'Equation (10) de OKM-$\sigma$. Cependant, la nouvelle densité du groupe $\sigma_k$ est définie pour Parameterized R-OKM-$\sigma$ par :

$$\sigma_k = \left\{ \frac{\sum_{i=1}^{N} P_{ik}.|\Pi_i|^{\alpha}.||x_i - (\overline{x_i})||^2}{\sum_{i=1}^{N} P_{ik}.|\Pi_i|^{\alpha}} \right\}^{1/2} . \quad (13)$$

## 4.3 Résolution algorithmique

La minimisation de la fonction objective de chaque méthode proposée (OKM-$\sigma$ and Parameterized R-OKM-$\sigma$) est réalisée par itération de trois étapes indépendantes : (1) calcul de représentants de groupes $C$, (2) affectation multiple ($\Pi$) d'observations à un ou à plusieurs groupes et (3) le calcul de poids ($\sigma_k$) pour chaque classe.

Sachant que OKM-$\sigma$ est un cas particulier de Parameterized R-OKM-$\sigma$ (quand $\alpha = 0$), nous présentons dans l'algorithme 1 les différentes étapes de Parameterized R-OKM-$\sigma$. Cet algorithme utilise la fonction $ASSIGN-\sigma$ permettant l'affectation multiple de chaque observation à un ou plusieurs groupes. Cette fonction utilise une heuristique, utilisée aussi dans OKM et Parameterized R-OKM, permettant de réduire l'espace exponentiel (en terme de nombre de classes) des affectations possibles. Cette heuristique consiste à continuer l'affectation de l'observation au cluster le plus proche tant que le critère objectif est amélioré. Les différentes étapes de $ASSIGN-\sigma$ sont décrites dans l'algorithme 2.

**Algorithm 1** *Parameterized R-OKM-$\sigma$* $(X, t_{max}, \varepsilon, K) \rightarrow \Pi$

**ENTRÉE** $X$ : Ensemble de données décrites sur $\mathbb{R}^d$.
$t_{max}$ : nombre maximum d'itérations.
$\varepsilon$ : amélioration minimale de la fonction objective.
**SORTIE** $\Pi$ : affectation des observations aux $K$ clusters .

1: Initialiser les représentants des groupes $C^0$ aléatoirement dans $X$, initialiser le poids $\sigma_k^0$, initialiser la matrice d'appartenance $\Pi^0$ en utilisant $ASSIGN-\sigma$ et calculer la fonction objective $J(\Pi^0, C^0, \sigma^0)$ à l'itération 0.
2: $t = t + 1$.
3: Mettre à jour les représentants des clusters $C^t$.
4: Calculer les nouvelles affectations $\Pi^t$ en utilisant $ASSIGN - \sigma(x_i, C^t, \Pi_i^{t-1})\ \forall i$.
5: Mettre à jour les poids $\sigma^t$ (en utilisant l'équation 13).
6: Calculer la fonction objective $J(\Pi^t, C^t, \sigma^t)$.
7: Si ($t < t_{max}$ et $J(\Pi^{t-1}, C^{t-1}, \sigma^{t-1}) - J(\Pi^t, C^t, \sigma^t) > \varepsilon$) alors
8: Passer à l'étape 2.
9: Sinon
10: Retourner $\Pi^t$ la dernière matrice d'appartenance.
11: FinSi

**Algorithm 2** $ASSIGN - \sigma(x_i, \{c_1, ...c_K\}, \Pi_i^{old}) \rightarrow \Pi_i$

**ENTRÉE** $x_i$ : vecteur dans $\mathbb{R}^d$.
$\{c_1, ...c_K\}$ : $K$ représentants de classes.
$\Pi_i^{old}$ : Ancienne affectation de l'observation $x_i$.
**SORTIE** $\Pi_i$ : Nouvelle affectation pour $x_i$.

1: initialiser $\Pi_i = \{c^\star\}$ le centre le plus proche où $c^\star = \arg min_{c_k} \parallel x_i - c_k \parallel^2$.
2: rechercher le prochain groupe le plus proche $c^\star$ qui n'est pas inclus dans $\Pi_i$.
3: Calculer $(\overline{x_i}')$ et $\sigma_i'$ avec des affectations $\Pi_i' = \Pi_i \cup \{c^\star\}$.
4: Si $|\Pi_i'|^\alpha . \frac{\|x_i - \overline{x_i}'\|^2}{\sigma_i'} \leq |\Pi_i|^\alpha . \frac{\|x_i - \overline{x_i}\|^2}{\sigma_i}$ alors
5: $\Pi_i \leftarrow \Pi_i'$ et passer à l'étape 2.
6: Sinon si $|\Pi_i|^\alpha . \frac{\|x_i - \overline{x_i}\|^2}{\sigma_i} \leq |\Pi_i^{old}|^\alpha . \frac{\|x_i - \overline{x_i}^{old}\|^2}{\sigma_i^{old}}$ alors
7: Retourner $\Pi_i$.
8: Sinon
9: Retourner $\Pi_i^{old}$.
10: Fin Si
11: Fin Si

# 5 Expérimentations

Cette section évalue l'efficacité de OKM-$\sigma$ et Parameterized R-OKM-$\sigma$ sur des ensembles de données artificielles. Afin d'évaluer la capacité des méthodes proposées à produire des groupes non-disjoints sur des ensembles de données ayant des densités différentes, nous avons généré deux jeux de données artificielles référencés "Ensemble 1" et "Ensemble 2". Le premier

jeu de données contient deux classes où chaque classe est formée de $500$ observations décrites dans un espace à deux dimensions. Les deux classes ont des densités différentes : la classe “bleue” a une grande densité par rapport à la classe “rouge”. Pour le deuxième ensemble de données, nous avons modifié le rayon du groupe "rouge" qui devient plus grand.

Pour donner au lecteur une interprétation visuelle de la performance des méthodes proposées par rapport à OKM et Parameterized R-OKM, nous commençons par visualiser les classes obtenues dans un espace à deux dimensions.

TAB. 1 – *Comparaison entre OKM-$\sigma$ et OKM sur les ensembles de données artificielles*

| | **OKM** | | | **OKM-$\sigma$** | | |
|---|---|---|---|---|---|---|
| **Données** | Précision | Rappel | F-mesure | Précision | Rappel | F-mesure |
| Ensemble 1 | $0,810 \pm 0,00$ | $1,00 \pm 0,00$ | $0,895 \pm 0,00$ | $0,886 \pm 0,04$ | $1,00 \pm 0,00$ | $0,939 \pm 0,02$ |
| Ensemble 2 | $0,702 \pm 0,08$ | $0,998 \pm 0,00$ | $0,817 \pm 0,06$ | $0,854 \pm 0,01$ | $1,00 \pm 0,00$ | $0,921 \pm 0,01$ |

TAB. 2 – *Comparaison entre Parameterized R-OKM-$\sigma$ et Parameterized R-OKM avec $\alpha = 1$ sur les ensembles de données artificielles*

| | **Parameterized R-OKM** ($\alpha = 1$) | | | **Parameterized R-OKM-$\sigma$** ($\alpha = 1$) | | |
|---|---|---|---|---|---|---|
| **Données** | Précision | Rappel | F-mesure | Précision | Rappel | F-mesure |
| Ensemble 1 | $0,896 \pm 0,01$ | $0,998 \pm 0,00$ | $0,944 \pm 0,01$ | $0,935 \pm 0,01$ | $1,00 \pm 0,00$ | $0,966 \pm 0,01$ |
| Ensemble 2 | $0,874 \pm 0,01$ | $0,995 \pm 0,00$ | $0,930 \pm 0,01$ | $0,926 \pm 0,01$ | $1,00 \pm 0,00$ | $0,962 \pm 0,01$ |

TAB. 3 – *Comparaison entre Parameterized R-OKM-$\sigma$ et Parameterized R-OKM avec $\alpha = 2$ sur les ensembles de données artificielles*

| | **Parameterized R-OKM** ($\alpha = 2$) | | | **Parameterized R-OKM-$\sigma$** ($\alpha = 2$) | | |
|---|---|---|---|---|---|---|
| **Données** | Précision | Rappel | F-mesure | Précision | Rappel | F-mesure |
| Ensemble 1 | $0,947 \pm 0,00$ | $0,995 \pm 0,01$ | $0,970 \pm 0,02$ | $0,960 \pm 0,05$ | $0,999 \pm 0,05$ | $0,979 \pm 0,01$ |
| Ensemble 2 | $0,937 \pm 0,00$ | $0,994 \pm 0,00$ | $0,965 \pm 0,01$ | $0,954 \pm 0,05$ | $0,998 \pm 0,01$ | $0,975 \pm 0,01$ |

Les Figures 2 et 3 montrent les groupes produits par OKM-$\sigma$ et Parameterized R-OKM-$\sigma$ par rapport aux groupes produits par les méthodes existantes. En premier lieu, nous constatons que toutes les méthodes sont capables de produire des classes non-disjointes. Ensuite, pour OKM, nous remarquons les larges intersections (Points "verts") entres les deux classes formées. Ce problème a été partiellement résolu en utilisant OKM-$\sigma$ où les intersections entre les deux groupes sont réduites. Concernant Parameterized R-OKM, nous remarquons que les intersections peuvent être réduites à chaque fois que la valeur du paramètre $\alpha$ augmente. Cependant, Parameterized R-OKM produit des groupes plus pertinents à mesure que la densité augmente.

Pour confirmer les résultats schématisés sur les précédentes figures, nous donnons une évaluation quantitative en terme de Précision, Rappel et F-mesure. Ces mesures permettent

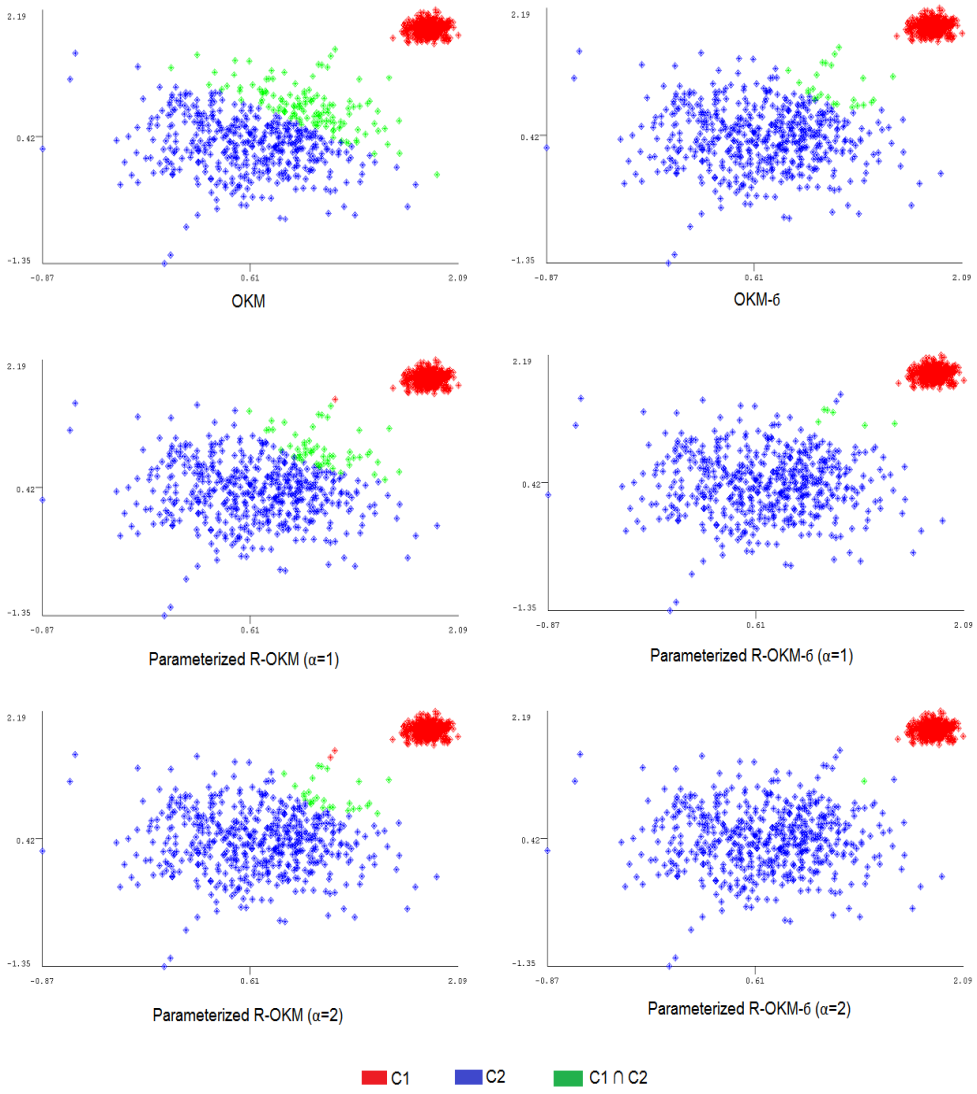


FIG. 2 – *Comparaison des classes obtenues avec OKM-$\sigma$ et Parameterized R-OKM-$\sigma$ par rapport aux classes obtenues avec OKM et Parameterized R-OKM sur l'ensemble* 1.

de vérifier le degré d'adéquation des classes obtenues avec les classes générées. Les tableaux 1 à 3 présentent les valeurs moyennes de Précision, Rappel et F-mesure obtenues avec dix exécutions de OKM-$\sigma$, Parameterized R-OKM-$\sigma$, OKM et Parameterized R-OKM sur les deux

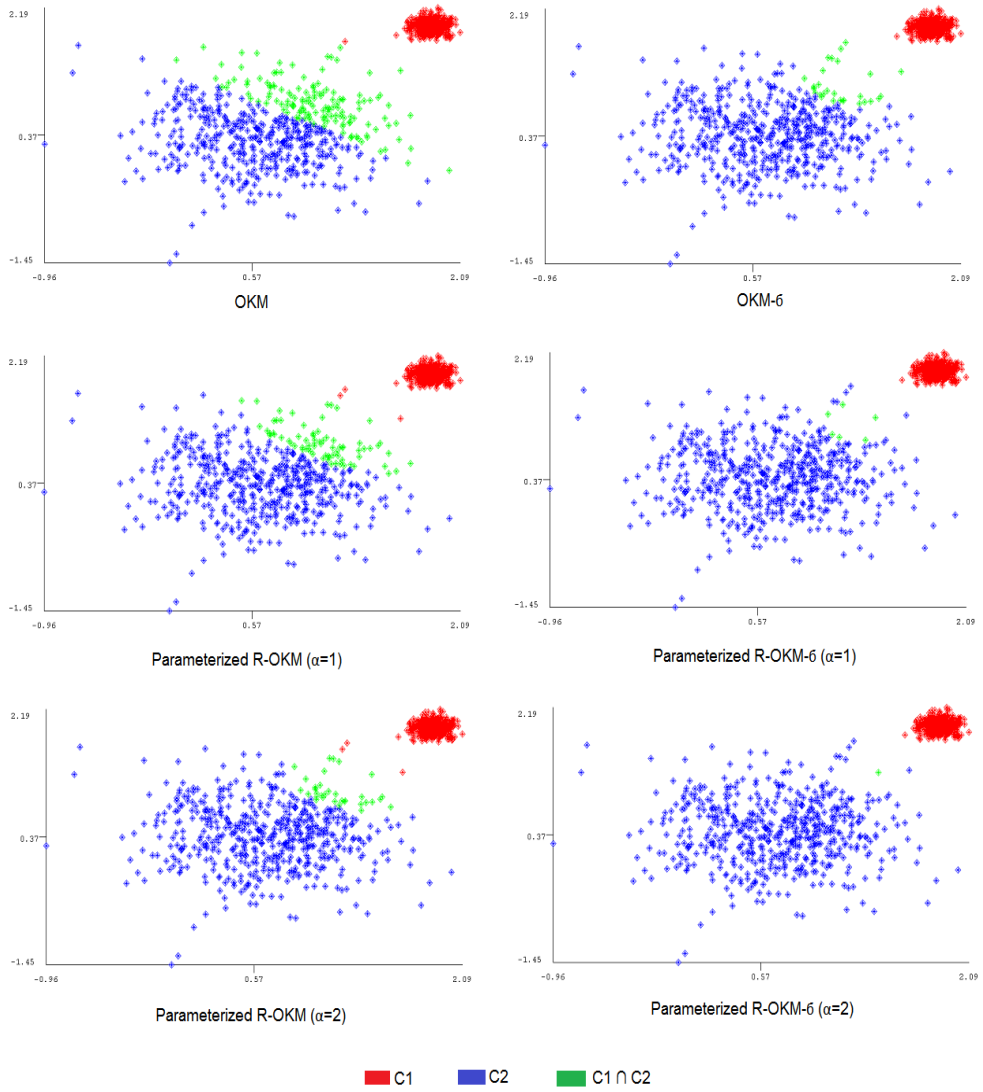


FIG. 3 – *Comparaison des classes obtenues avec OKM-$\sigma$ et Parameterized R-OKM-$\sigma$ par rapport aux classes obtenues avec OKM et Parameterized R-OKM sur l'ensemble* 2.

jeux de données artificielles. Ces résultats montrent que les méthodes proposées surpassent les méthodes originales en terme de F-mesure. Par exemple, en utilisant OKM-$\sigma$ sur "l'ensemble 2", la F-mesure obtenue augmente de $0.817$ à $0.921$ par rapport à OKM. L'amélioration de

la F-mesure sur les deux jeux de données est due à l'amélioration de la Précision. Ce résultat s'explique par le fait que les méthodes originales construisent des recouvrements de plus en plus larges à mesure que la densité entre les deux classes diffère. Cependant, en utilisant les méthodes proposées, le taux de recouvrements est réduit.

# 6 Conclusion

Nous avons proposé dans ce travail deux nouvelles méthodes capables de produire des classes non-disjointes lorsque les données s'organisent en groupes de densités différentes. Ces nouvelles méthodes s'appuient sur une nouvelle mesure de distance qui régularise la variation de densité entre les classes obtenues. Des expériences réalisées sur des ensembles de données artificielles montrent l'efficacité des méthodes proposées par rapport à celles existantes.

L'évaluation des méthodes proposées est effectuée sur des ensembles de données artificielles. Comme perspectives, nous envisageons de confirmer ces résultats sur des jeux de données réels.

# Références

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

Ben N'Cir, C.-E., G. Cleuziou, et N. Essoussi (2013). Identification of non-disjoint clusters with small and parameterizable overlaps. In *Computer Applications Technology (ICCAT), 2013 International Conference on*, pp. 1–6.

Bonchi, F., A. Gionis, et A. Ukkonen (2011). Overlapping correlation clustering. In *11th IEEE International Conference on Data Mining (ICDM)*, pp. 51–60.

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

Diday, E. (1984). Orders and overlapping clusters by pyramids. Technical Report 730, INRIA, France.

Fellows, M. R., J. Guo, C. Komusiewicz, R. Niedermeier, et J. Uhlmann (2009). Graph-based data clustering with overlaps. In *Proceedings of the 15th Annual International Conference on Computing and Combinatorics*, COCOON '09, Berlin, Heidelberg, pp. 516–526. Springer-Verlag.

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

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

Govaert, G. (1975). *Classification automatique et distances adaptatives*.

Heller, K. et Z. Ghahramani (2007). A nonparametric bayesian approach to modeling overlapping clusters. In *AISTATS*, Puerto Ric.

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

Tsai, D.-M. et C.-C. Lin (2011). Fuzzy c-means based clustering for linearly and nonlinearly separable data. *Pattern Recogn. 44*(8), 1750–1760.

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

## Summary

Overlapping clustering is an important issue in clustering which allows an observation to belong to more than one cluster. Several overlapping methods were proposed to solve this issue. Although the effectiveness of these methods to detect non disjoint clusters, they fail when clusters have different densities. In order to detect overlapping clusters with different densities, we propose two clustering methods based on a new distance metric that incorporates the distance variation in a cluster to regularize the distance between a data point and the cluster representative. Experiments performed on artificial data sets show that proposed methods with the new distance metric have better performances when clusters have different densities.