Passage aux noyaux en classification recouvrante

Guillaume Cleuziou*,**

*LIFO, Université d'Orléans, Rue Léonard de Vinci, 45067 Orléans Cedex 2
guillaume.cleuziou@univ-orleans.fr
**GREYC, UCBN, Bd. Maréchal Juin, 14032 Caen Cedex 5

Résumé. La classification recouvrante correspond à un domaine d'étude très actif ces dernières années et dont l'objectif est d'organiser un ensemble de données en groupes d'individus similaires avec la particularité d'autoriser des chevauchements entre les groupes. Parmi les approches étudiées nous nous intéressons aux extensions recouvrantes des modèles de type moindres carrés et constatons les difficultés théoriques et pratiques liées à leur adaptation aux noyaux. Nous formulons alors une nouvelle définition ensembliste pour caractériser un recouvrement de plusieurs classes, nous montrons que cette modélisation permet le recours aux noyaux et nous proposons une solution algorithmique efficace pour répondre au problème de la classification recouvrante à noyaux.

1 Introduction

La classification recouvrante consiste à organiser de manière non-supervisée un ensemble d'individus en classes chevauchantes ou recouvrantes composées d'individus similaires. L' étude des méthodologies associées vise avant tout à répondre aux besoins réels et pratiques communs à de nombreux domaines d'application : qu'il s'agisse de classer des documents (textes, images, vidéos), de constituer des communautés de personnes sur des critères sociaux ou marketing ou encore d'exhiber des groupes de gènes ou de molécules présentant des caractéristiques structurelles ou fonctionnelles communes, il est très fréquents d'être confronté à des données qui s'organisent naturellement en classes recouvrantes. Dans ces applications, le recours à des techniques usuelles de classification stricte ou disjointe apporterait un biais préjudiciable à l'usage final de la classification.

Depuis plus d'une trentaine d'années, la classification recouvrante est identifiée comme une problématique à part entière et donne lieu à des avancées constantes au fil de l'évolution des techniques de classification traditionnelles. Partant des premiers travaux de Shepard et Arabie (1979) portant sur le clustering (recouvrant) additif, la problématique s'est ensuite orientée vers l'acquisition et la théorisation des classifications hiérarchiques recouvrantes ou empiétantes (Diday, 1987; Diatta et Fichet, 1994; Bertrand et Janowitz, 2003) avant d'être reconsidérée plus récemment et de façon plus formelle du point de vue "partitionnement" (Banerjee et al., 2005; Cleuziou, 2008; Depril et al., 2012). Parmi ces dernières avancées on notera d'une part les modèles additifs ALS (Depril et al., 2012) et son équivalent en terme de modèle de mélanges recouvrants MOC (Banerjee et al., 2005) et d'autre part le modèle OKM (Cleuziou, 2008) que l'on pourrait qualifier de modèle géométrique tant il modélise les intersections de clusters par

un barycentre plutôt que par une somme des profils de clusters comme c'est le cas des modèles additifs précédents. Ces trois dernières méthodes se fondent sur un même cadre théorique qui consiste à explorer de manière efficace l'espace des solutions recouvrantes pour un nombre fixé de classes. Les algorithmes qui en découlent sont guidés par des critères objectifs quantifiant l'accumulation des imprécisions liées à l'affectation de chaque données à une classe ou à une combinaison de classes ; ce choix d'affectation posant lui même un problème combinatoire puisque le nombre de combinaisons possibles d'un ensemble fixé de classes est exponentiel.

Fort des avancées de cette dernière décennie dans le domaine de la classification recouvrante, nous nous intéressons à présent à la possibilité d'étendre ces modèles à l'utilisation de noyaux. La "kernélisation" des méthodes recouvrantes permettrait d'en étendre l'usage à des données de très grande dimensionalité (telles que les données textuelles), à des espaces non-euclidiens à l'origine, d'avoir recours aux mêmes procédés de projections que ceux qui ont fait leurs preuves en classification traditionnelle (non-recouvrante) et d'envisager de nouvelles avancées dans les domaines du clustering spectral ou semi-supervisé recouvrant (Dhillon, 2004; Filippone et al., 2008; Kulis et al., 2009) par exemple.

Dans cette contribution nous présentons tout d'abord les difficultés qui ont freiné l'évolution des modèles recouvrants vers le clustering à noyaux (Section 2). Nous proposons ensuite en Section 3 une nouvelle modélisation ensembliste des recouvrements. Cette modélisation se base sur le modèle OKM et permet à la fois de corriger certains biais du modèle originel et de rendre possible le passage aux noyaux via un algorithme adaptatif également présenté dans cette section. Nous confirmons enfin les attentes liées à ce nouveau modèle par le biais d'une étude empirique préliminaire sur des jeux de données réels et artificiels (Section 4).

2 Problématique des noyaux en classification recouvrante

En clustering, l'astuce du noyau, consiste à réaliser le processus de classification (classiquement les réallocations successives) dans un espace induit par le noyau sans jamais calculer explicitement les projections des données de départ dans ce nouvel espace. Considérons $X = \{x_1, \ldots, x_n\}$ dans $\mathbb{R}^p$ l'ensemble des données à traiter et K la matrice noyau induite par une projection implicite $\phi(.)$ telle que $K_{i,j} = \langle \phi(x_i), \phi(x_j) \rangle$. A priori, la méthode des k-moyennes ne peut pas être "kernélisée" directement puisqu'elle considèrerait la minimisation du critère d'inertie suivant

$$J_{kmeans}(\{\pi_c\}_{c=1}^k, \{m_c\}_{c=1}^k) = \sum_{c=1}^{k} \sum_{x_i \in \pi_c} \|\phi(x_i) - \phi(m_c)\|^2$$

et que ce dernier s'appuie sur k centre mobiles (modélisant les profils des clusters) $\{m_c\}_{c=1}^k$ dont on ne doit pas calculer explicitement les projections $\phi(m_c)$. Cependant en observant que les centres mobiles sont redéfinis à chaque itération de façon optimale par les moyennes (centres de gravité) des individus de chaque classe, (Dhillon, 2004) ont proposé de se passer des variables associées à ces centres et d'intégrer leur définition dans le critère initial :

$$J_{kmeans}(\{\pi_c\}_{c=1}^k) = \sum_{c=1}^{k} \sum_{x_i \in \pi_c} K_{i,i} - \frac{2 \sum\limits_{x_j \in \pi_c} K_{i,j}}{|\pi_c|} + \frac{\sum\limits_{x_j, x_l \in \pi_c} K_{j,l}}{|\pi_c|^2}.$$

Cette astuce permet d'envisager un clustering totalement identique à k-moyennes dans n'importe quel espace induit par un noyau K, cependant il est important d'observer que cette transformation a un coût (raisonnement sur les paires d'individus).

Lorsque l'on tente de kernéliser les méthodes de classification recouvrante de type "réallocation dynamique" (e.g. ALS, MOC ou OKM), il semblerait naturel de procéder de façon similaire. Si l'on choisit - sans perte de généralité - le modèle OKM [1], on serait amené à considérer le critère objectif suivant

$$J_{OKM}(\{\pi_c\}_{c=1}^k, \{m_c\}_{c=1}^k) = \sum_{x_i \in X} \left\| \phi(x_i) - \frac{\sum_{c=1}^k \mathbf{1}_{x_i \in \pi_c} \phi(m_c)}{\sum_{c=1}^k \mathbf{1}_{x_i \in \pi_c}} \right\|^2 \quad (1)$$

où $\mathbf{1}_{x_i \in \pi_c} = 1$ si $x_i \in \pi_c$ et 0 sinon. Ce critère quantifie une somme de distances entre chaque individu x_i et une combinaison (la moyenne dans OKM) des profils des clusters auxquels il appartient. Malheureusement, la définition d'un profil de cluster ne correspond plus à un simple centre de gravité : les profils dépendent les uns des autres ce qui rend impossible la réécriture à partir des seules données relationnelles du noyau (produits scalaires entre individus).

Ben N'Cir et Essoussi (2012) se sont intéressés à la kernélisation de OKM ; afin de contourner cette difficulté, ils ont proposé de conserver les variables de profils mais de limiter leur domaine de définition à l'ensemble des individus X, à la manière de médoïdes. L'individu choisi comme profil du cluster π_c sera celui qui minimise un critère d'inertie sur π_c :

$$m_c = \arg\min_{x_i \in \pi_c} \frac{\sum_{x_j \in \pi_c} w_j \|\phi(x_i) - \phi(x_j)\|^2}{\sum_{x_j \in \pi_c} w_j} \quad \text{où} \quad w_j = \sum_{c=1}^k \mathbf{1}_{x_i \in \pi_c}. \quad (2)$$

On voit effectivement que les profils de clusters pourront ainsi être choisis à partir des informations issues du noyau uniquement et une fois ces profils déterminés, les auteurs montrent que l'heuristique utilisée dans OKM pour l'affectation peut être réalisée à l'identique dans l'espace de projection. Néanmoins le recours aux médoïdes est une solution coûteuse nécessitant de considérer toutes les paires d'individus de chaque cluster pour leur mise à jour. Enfin et surtout, la définition du médoïde (2) reste un choix arbitraire qui n'est pas induit par le critère initial (1) et ne garantit pas la convergence de l'algorithme qui en découle.

Dans la suite de l'article nous proposons une approche différente pour le problème du clustering recouvrant à noyau. Il s'agit d'une approche ensembliste qui diffère légèrement du modèle OKM originel en corrigeant certains de ses inconvénients mais surtout permettant l'utilisation des noyaux dans un processus recouvrant et convergeant.

1. La problématique reste identique pour les autres méthodes ALS ou MOC.

3 OKSETS : modèle et algorithme

3.1 Le modèle de combinaison

Nous commencerons par l'observation de Depril et al. (2012) concernant la méthode ALS, qui est également valable pour toute méthode de classification recouvrante fondée sur une combinaison de profils de clusters : dans ces approches, la notion de "profils" de clusters ne correspond plus à l'idée intuitive de "centres" que l'on peut avoir dans les méthodes non-recouvrantes de type k-moyennes ; en effet, l'optimisation du critère objectif conduit à des profils pouvant être éloignés des données des clusters qu'ils représentent. Plus précisément, Cleuziou (2008) montre par dérivation de (1) que dans OKM les profils peuvent être mis à jour l'un après l'autre de façon optimale avec la règle suivante :

$$m_c^* = \frac{\sum_{x_i \in \pi_c} \frac{m_c^i}{w_i^2}}{\sum_{x_i \in \pi_c} \frac{1}{w_i^2}} \text{ où } m_c^i = x_i.w_i - \sum_{q \neq c} \mathbf{1}_{x_i \in \pi_q} m_q$$

le profil optimal m_c^* est alors obtenu par une moyenne pondérée des m_c^i, c'est-à-dire des profils de π_c idéaux vis-à-vis de chaque individu du cluster. Ce phénomène, induit par le critère objectif initial, peut conduire progressivement à augmenter les recouvrements entre clusters de façon artificielle : plus deux clusters se recouvriront, plus leurs profils seront éloignés de leur centre de gravité et plus ils risquent d'engendrer des recouvrements supplémentaires.

Nous proposons de profiter de la nécessité de faire abstraction des variables de profils dans le passage aux noyaux, pour corriger le phénomène de "sur-recouvrements" que nous venons d'évoquer. Pour cela nous commençons par introduire la notion de *nuage* dans une classification.

Définition 1 *Étant donné un ensemble de clusters* $\Pi = \{\pi_1, \ldots, \pi_q\}$*, on appellera "nuage" de* Π *l'application* $N(.)$ *qui lui associe l'union de ses extensions*

$$N(\Pi) = \bigcup_{\pi_c \in \Pi} \pi_c.$$

On parlera également de "nuage" associé à un individu $N(x_i)$ *pour indiquer de façon analogue l'union des clusters auxquels il appartient*

$$N(x_i) = \bigcup_{\pi_c | x_i \in \pi_c} \pi_c.$$

Nous définissons ensuite un nouveau critère objectif pour la classification recouvrante. Ce nouveau critère est défini sur la base d'une somme d'erreurs locales modélisées par les distances des individus au centre de gravité de leur nuage associé :

$$J_{OKSets}(\{\pi_c\}_{c=1}^k) = \sum_{x_i \in X} \left\| x_i - \sum_{x_j \in N(x_i)} \frac{x_j}{|N(x_i)|} \right\|^2 \quad (3)$$

De cette manière on est assuré que l'affectation d'un individu à un cluster se réalise sur la base de sa distance au centre de gravité du cluster. De plus, un individu sera affecté à plusieurs clusters si le centre de gravité du nuage de cette combinaison est plus proche de cet individu. Enfin, on montre que ce critère est adapté à l'utilisation de noyaux :

$$J_{K-OKSets}(\{\pi_c\}_{c=1}^k) = \sum_{x_i \in X} \left\| \phi(x_i) - \sum_{x_j \in N(x_i)} \frac{\phi(x_j)}{|N(x_i)|} \right\|^2$$
$$= \sum_{x_i \in X} K_{i,i} - \frac{2\sum_{x_j \in N(x_i)} K_{i,j}}{|N(x_i)|} + \frac{\sum_{x_j, x_l \in N(x_i)} K_{j,l}}{|N(x_i)|^2}$$

et on observe que dans le cas d'un clustering non-recouvrant, chaque individu appartient à un seul cluster $x_i \in \pi_c \Leftrightarrow N(x_i) = \pi_c$, ce qui nous ramène exactement au critère objectif de l'algorithme des k-moyennes à noyaux (cf. Section 2). Le modèle K-OKSETS défini précédemment est donc une généralisation recouvrante du modèle des k-moyennes à noyaux.

3.2 L'algorithme adaptatif

Contrairement au contexte du k-moyennes à noyaux où l'on sait que le centre de gravité d'un cluster correspond à sa représentation optimale, dans le contexte recouvrant nous avons évoqué le fait que cela n'est plus vérifié. Ainsi un algorithme (batch) classique qui consisterait à réaffecter itérativement tous les individus avant de remettre à jour globalement tous les nuages n'assurerait pas la décroissance du critère (3) et n'aurait donc aucune raison de converger vers un recouvrement et des profils stables. Nous proposons donc un algorithme adaptatif guidé par le critère $J_{K-OKSets}$ (Figure 1). L'étape cruciale de l'algorithme réside dans la pro-

K-OKSETS

Entrées : X un ensemble de n individus, K une matrice noyau sur X, k le nombre de clusters attendu, $\{x^1, \ldots, x^k\}$ k prototypes d'initialisation et max_{iter} un nombre maximum d'itérations

1. Initialiser la classification avec $\pi_c = \{x^c\}$ $\forall c \in 1..k$
2. $\forall x_i \in X \setminus \{x^1, \ldots, x^k\}$, AFFECTER($x_i$) et en déduire $J_{K-OKSets}$ ($t \leftarrow 0$)
3. Tant que $J_{K-OKSets}$ décroît et $t < max_{iter}$
   4. Pour $i = 1..n$: AFFECTER(x_i) et en déduire $J_{K-OKSets}$
   5. $t \leftarrow t + 1$

Sortie : la classification recouvrante $\{\pi_c\}_{c=1}^k$

FIG. 1 – *Principe de l'algorithme adaptatif* K-OKSETS.

cédure d'affectation d'un individu à un ou plusieurs clusters. Cette procédure doit être réalisée efficacement et assurer la décroissance du critère $J_{K-OKSets}$. En théorie, la recherche de l'affectation optimale nécessiterait de considérer toutes les combinaisons possibles de clusters (au

nombre de $2^k - 1$), déterminer le nuage associé à chacune de ces combinaisons, puis calculer la distance de l'individu au centre de ce nuage afin de choisir la combinaison minimisant cette distance. En pratique nous choisissons d'une part de suivre l'heuristique d'affectation proposée pour OKM (de complexité quasi-linéaire sur k) et d'autre part de stocker dans une structure appropriée les nuages associés à chacune des combinaisons existantes pour éviter de devoir les recalculer à chaque nouvelle affectation.

L'heuristique d'affectation que nous utilisons consiste, pour un individu x_i à :

1. Ordonner les clusters du plus proche au plus éloigné de x_i selon la distance

$$\left\| \phi(x_i) - \sum_{x_j \in \pi_c} \frac{\phi(x_j)}{|\pi_c|} \right\|^2 = K_{i,i} - \frac{2 \sum_{x_j \in \pi_c} K_{i,j}}{|\pi_c|} + \frac{\sum_{x_j, x_l \in \pi_c} K_{j,l}}{|\pi_c|^2}$$

2. Affecter x_i au premier cluster puis aux suivants tant que l'erreur locale (rappelée ci-dessous) diminue

$$K_{i,i} - \frac{2 \sum_{x_j \in N(x_i)} K_{i,j}}{|N(x_i)|} + \frac{\sum_{x_j, x_l \in N(x_i)} K_{j,l}}{|N(x_i)|^2}$$

3. Revenir à l'ancienne affectation si l'erreur locale n'est pas améliorée

La structure de données sur laquelle se base l'algorithme conserve les informations utiles à l'ajout ou la suppression d'un élément de manière à simplifier les calculs d'erreurs locales et donc de l'erreur globale. Elle est organisée de manière arborescente où chaque nœud de l'arbre correspond à une combinaison de clusters et contient : l'ensemble des individus affectés à cette combinaison, le nuage associé ainsi qu'un score correspondant à la moyenne des produits scalaires sur les couples contenus dans le nuage : $score(N) = \sum_{x_i, x_j \in N} K_{i,j}/|N|^2$. La Figure 2, illustre une structure ainsi que la classification en 3 classes associée : par exemple l'individu x_6 apparaît à l'intersection des 3 clusters, on le retrouve donc dans le nœud d'étiquette $\pi_1 \wedge \pi_2 \wedge \pi_3$ dans lequel on trouve également l'information sur le nuage et le score associé à ce nuage. Ainsi la réaffectation d'un individu x_i nécessitera les étapes suivantes :

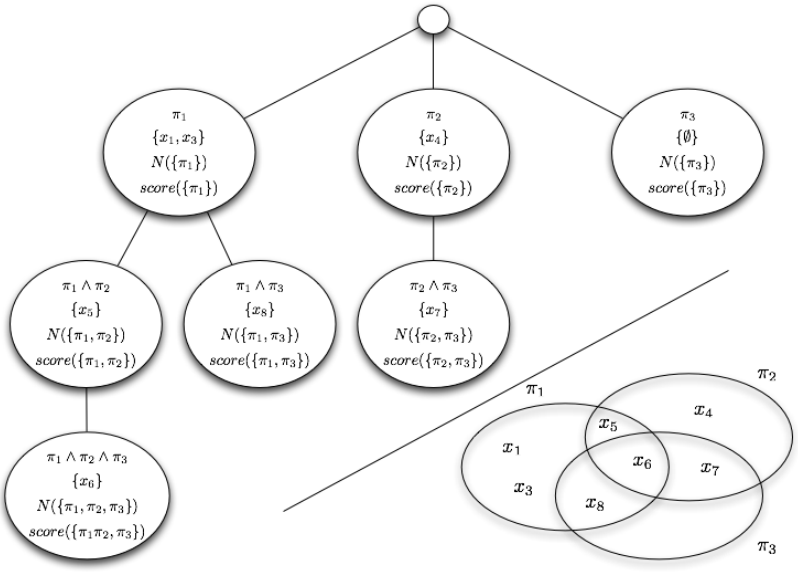
1. Un premier parcours partiel de l'arbre pour supprimer x_i du nœud auquel il est affecté, puis une mise à jour de tous les nuages et scores des nœuds prédécesseurs :

$$Score(N \setminus x_i) = \frac{Score(N).|N|^2 - 2. \sum_{x_j \in N} K_{i,j} - K_{i,i}}{(|N| - 1)^2}$$

2. Un second parcours partiel pour l'heuristique d'ajout nécessitant les mises à jour inverses selon le même procédé.

Compte-tenu de l'augmentation exponentielle du nombre théorique de combinaisons de clusters avec le nombre de clusters k, la structure proposée n'est envisageable que sous l'hypothèse que seul un petit nombre de combinaisons est effectivement exploré par l'heuristique d'affectation et la condition que seule les combinaisons explorées soient stockées. Il n'est en effet pas déraisonnable de penser que des combinaisons de clusters éloignés ont peu de chances d'accueillir des individus qui sont affectés à leurs plus proches clusters d'après l'heuristique choisie. Nous observerons dans l'étude empirique qui suit les premiers gages de confirmation de cette hypothèse.

FIG. 2 – *Structure classificatoire pour l'algorithme* K-OKSETS.



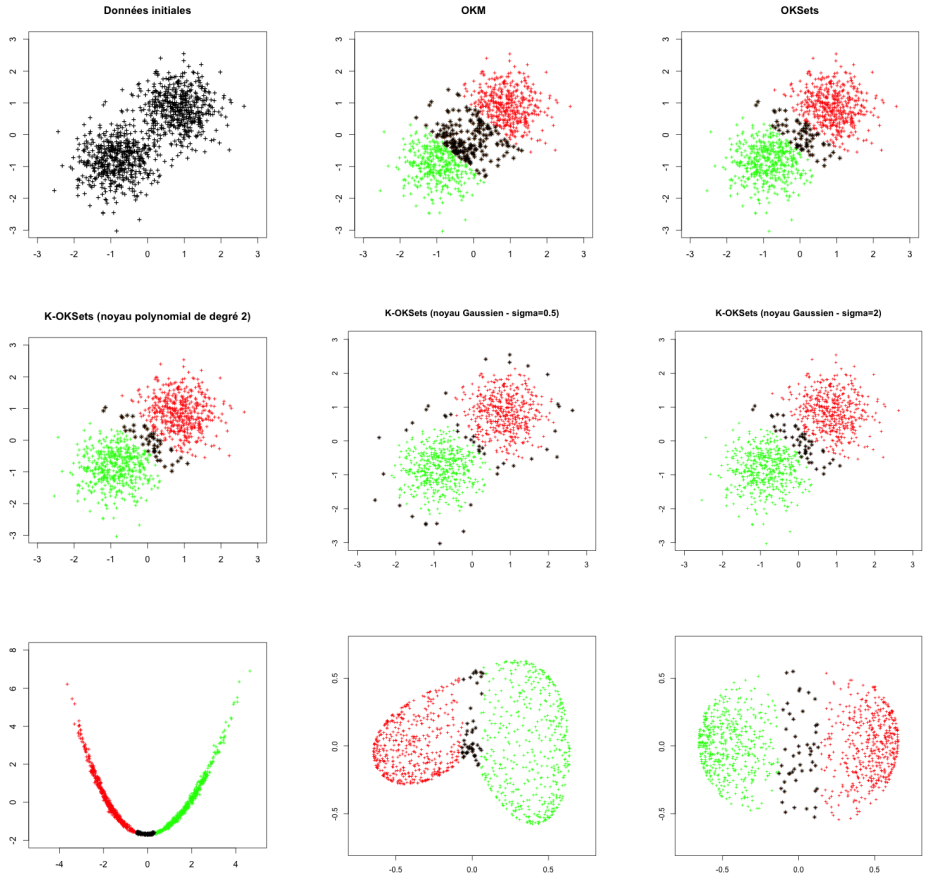
4 Étude empirique

Nous proposons dans cette section une première évaluation empirique du modèle de classification ensembliste OKSETS et de son utilisation dans le cadre du clustering à noyau. Nous proposons en Figure 3 différentes classifications d'un même jeu de données artificielles généré par deux gaussiennes légèrement chevauchantes de 500 individus chacune en deux dimensions. On observe en premier lieu (ligne supérieure de la Figure 3) que OKM et sa variante ensembliste génèrent des résultats comparables en terme de nature du recouvrement entre les deux classes (individus en noir) ; cependant OKSETS présente une "bande" de recouvrement plus étroite du fait de l'absence de profils de clusters mobiles. Les autres visualisations rendent compte des classifications obtenues par K-OKSETS avec différents noyaux : polynomial de degré 2 et gaussiens avec variances de 0.5 et 2 ; nous visualisons les classifications (et en particulier les recouvrements) à la fois dans l'espace initial et dans l'espace de projection approximé par un positionnement multidimensionnel (MDS) à partir des distances euclidiennes induites par le noyau. On notera de façon générale la cohérence des classes et recouvrement générés et en particulier qu'il semblerait possible de détecter par les recouvrements le contour des classes en utilisant un noyau approprié.

Dans un second temps, nous procédons à une évaluation dite "externe" des classifications obtenues. Il s'agit alors de mesurer l'adéquation entre la classification générée par l'algorithme de façon totalement non-supervisée et une classification de référence attendue par des experts du domaine. Nous utilisons trois jeux de données réelles : Iris, EachMovie et Scene.

– Iris (D.J. Newman et Merz, 1998) est bien connu des spécialistes de classification (super-

FIG. 3 – *Visualisations des recouvrements sur données artificielles : sans noyau (ligne sup.), avec noyau (ligne du milieu) et dans l'espace de projection reconstruit (ligne inf.).*



visée ou non), il est constitué de 150 fleurs décrites selon 4 caractéristiques numériques, chacune des fleurs étant étiquetée par une seule des trois catégories Setosa, Versicolor ou Virginica, dont la première est réputée facilement identifiable, contrairement aux deux autres catégories plus mélangées.

– EachMovies [2] est plutôt utilisé en classification "multi-étiquettes", il s'agit d'extraits de films décrits par les préférences d'utilisateurs (3 notes par film) et organisés en classes de genres où chaque film peut être associé à plusieurs genres. Nous utilisons ici un sous-ensemble de 3 genres correspondant à 75 extraits de films associés en moyenne à 1.14 genres chacun (taux de recouvrement).
– Scene [3] correspond à un ensemble de 2407 photos décrites selon 294 attributs numé-

2. http ://grouplens.org/datasets/eachmovie/
3. http ://mlkd.csd.auth.gr/multilabel.html

riques (modèle LUV) et étiquetées manuellement selon 6 catégories (*beach*, *sunset*, *fall foliage*, *field*, *mountain* et *urban*) avec un taux de recouvrement moyen de 1.07.

Nous suggérons de quantifier l'adéquation (entre classifications) par la mesure F-BCubed étendue pour les classifications recouvrantes et proposée récemment par Amigó et al. (2009). Cette mesure évalue en terme de précision/rappel, les paires d'individus correctement associées et dissociées par la classification en tenant compte des associations multiples induites par les recouvrements. La Tableau 1 contient les moyennes (et écart-types) de précision, rappel, F-BCubed et Taux de recouvrement sur 5 (pour Scene) ou dix (pour Iris et EachMovies) initialisations différentes des algorithmes [4]. Nous comparons ainsi les performances des méthodes k-means et KkMEANS [5] (k-moyennes à noyau) qui ne produisent pas de recouvrements avec OKM, OKSETS et KOKSETS qui génèrent des classifications recouvrantes. Toutes les méthodes ont été paramétrées avec un nombre de clusters attendus (k) égal au nombre d'étiquettes associées au jeu de données. Les résultats les plus remarquables pour chaque type de classification sont identifiés en gras dans le tableau. On observe sur Iris que, bien que la classification attendue soit non recouvrante, il est possible d'obtenir une meilleure adéquation en autorisant des recouvrements limités : en effet sans recours aux noyaux OKSETS (linéaire) obtient un score de 0.82 (FBCubed) légèrement supérieur au meilleur score obtenu sans recouvrements (0.81 pour KkMeans avec noyau Gaussien). Ceci s'explique en particulier par le chevauchement naturel entre les catégories Versicolor et Virginica qui conduit, lorsqu'on l'autorise, à un recouvrement ; en terme d'adéquation, mieux vaut attribuer un individu à deux classes plutôt qu'à une seule avec le risque de choisir la mauvaise. On notera que la projection opérée par un noyau polynomial n'est pas pertinente sur Iris, tandis que le noyau Gaussien permet d'améliorer la qualité des classifications ; en particulier le score obtenu par KOKSETS avec le noyau gaussien ($\sigma = 1$) est intéressant dans la mesure où il égale le meilleur score tout en réduisant les recouvrements (1.07 contre 1.14).

Les tests réalisés sur les données EachMovie sont encore plus encourageants. La possibilité d'introduire les noyaux dans le clustering recouvrant conduit à un score inégalé de 0.69 obtenu en alliant la qualité à la fois en précision et en rappel, et avec un taux de recouvrement identique à la référence (1.14). Enfin, les expérimentations comparatives menées sur Scene valident sur un jeu de données plus important, l'intérêt de la modélisation ensembliste proposée dans OKSETS sans d'avantage de gain à attendre de l'utilisation de noyaux, que ce soit en partitionement ou en recouvrement.

Pour terminer cette première étude expérimentale, nous avons cherché à confirmer empiriquement l'hypothèse utilisée par l'algorithme KOKSETS selon laquelle le nombre de combinaisons (de clusters) explorées durant le processus de classification reste raisonnable par rapport à l'ensemble théorique des combinaisons possibles. La Figure 4 rend compte d'une première simulation qui tend à confirmer l'hypothèse : par exemple pour 15 clusters sur Iris, au maximum 52 combinaisons seront considérées et stockées dans la structure, contre plus de 32 000 combinaisons théoriques envisageables.

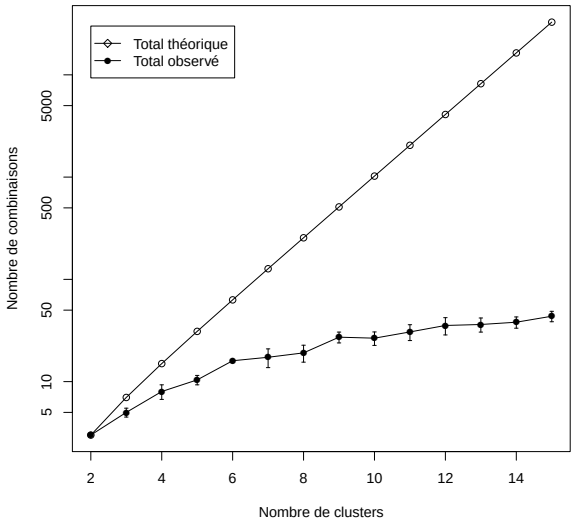
4. Le même ensemble d'initialisations est utilisé pour comparer toutes les méthodes.

5. Nous utilisons l'implémentation de KKmeans proposée dans la librairie R "kernlab" ; celle-ci propose un ajustement automatique des paramètres du noyau gaussien.

TAB. 1 – *Évaluation externe comparative d'algorithmes de classification recouvrante et non-recouvrante en utilisant différents noyaux.*

Données	Méthode	Précision	Rappel	FBcubed	Tx Overlap
Iris	kMeans	0.77 ± 0.09	0.83 ± 0.02	0.80 ± 0.06	1.00 ± 0.00
	KkMEANS (Linéaire)	0.77 ± 0.09	0.83 ± 0.02	0.80 ± 0.06	1.00 ± 0.00
	KkMEANS $K_{i,j} = (1 + \langle x_i, x_j \rangle)^2$	0.74 ± 0.07	0.79 ± 0.02	0.76 ± 0.05	1.00 ± 0.00
	KkMEANS $K_{i,j} = (1 + \langle x_i, x_j \rangle)^3$	0.73 ± 0.07	0.76 ± 0.06	0.74 ± 0.06	1.00 ± 0.00
	KkMEANS (Gaussien σ=1.57$\pm$0.42)	**0.77**$\pm$0.08	**0.85**$\pm$0.03	**0.81**$\pm$0.05	**1.00**$\pm$0.00
	OKM	0.55 ± 0.12	0.98 ± 0.01	0.70 ± 0.10	1.43 ± 0.11
	OKSETS (linéaire)	**0.73**$\pm$0.07	**0.93**$\pm$0.02	**0.82**$\pm$0.05	**1.14**$\pm$0.05
	KOKSETS $K_{i,j} = (1 + \langle x_i, x_j \rangle)^2$	0.72 ± 0.02	0.89 ± 0.01	0.80 ± 0.01	1.13 ± 0.02
	KOKSETS $K_{i,j} = (1 + \langle x_i, x_j \rangle)^3$	0.64 ± 0.03	0.87 ± 0.02	0.74 ± 0.02	1.17 ± 0.02
	KOKSETS (Gaussien $\sigma = 0.5$)	0.64 ± 0.12	0.87 ± 0.08	0.73 ± 0.09	1.07 ± 0.05
	KOKSETS (Gaussien $\sigma = 1$)	**0.75**$\pm$0.09	**0.92**$\pm$0.04	**0.82**$\pm$0.05	**1.07**$\pm$0.06
	KOKSETS (Gaussien $\sigma = 2$)	0.66 ± 0.15	0.94 ± 0.05	0.76 ± 0.12	1.16 ± 0.18
Each Movie	kMeans	**0.66**$\pm$0.10	**0.61**$\pm$0.05	**0.63**$\pm$0.06	**1.00**$\pm$0.00
	KkMEANS (Linéaire)	**0.66**$\pm$0.10	**0.61**$\pm$0.05	**0.63**$\pm$0.06	**1.00**$\pm$0.00
	KkMEANS $K_{i,j} = (1 + \langle x_i, x_j \rangle)^2$	0.65 ± 0.08	0.61 ± 0.05	0.62 ± 0.05	1.00 ± 0.00
	KkMEANS $K_{i,j} = (1 + \langle x_i, x_j \rangle)^3$	0.63 ± 0.08	0.61 ± 0.04	0.62 ± 0.04	1.00 ± 0.00
	KkMEANS (Gaussien σ=0.87$\pm$0.91)	0.61 ± 0.09	0.52 ± 0.06	0.56 ± 0.07	1.00 ± 0.00
	OKM	0.43 ± 0.03	0.90 ± 0.03	0.58 ± 0.02	1.69 ± 0.12
	OKSETS	0.66 ± 0.07	0.69 ± 0.03	0.67 ± 0.03	1.12 ± 0.04
	KOKSETS $K_{i,j} = (1 + \langle x_i, x_j \rangle)^2$	**0.66**$\pm$0.06	**0.73**$\pm$0.02	**0.69**$\pm$0.03	**1.14**$\pm$0.02
	KOKSETS $K_{i,j} = (1 + \langle x_i, x_j \rangle)^3$	0.65 ± 0.03	0.72 ± 0.02	0.68 ± 0.01	1.16 ± 0.03
	KOKSETS (Gaussien $\sigma = 0.5$)	0.46 ± 0.03	0.81 ± 0.10	0.58 ± 0.03	1.04 ± 0.04
	KOKSETS (Gaussien $\sigma = 1$)	0.58 ± 0.09	0.68 ± 0.06	0.62 ± 0.06	1.14 ± 0.09
	KOKSETS (Gaussien $\sigma = 2$)	0.49 ± 0.06	0.75 ± 0.07	0.59 ± 0.03	1.34 ± 0.13
Scene	kMeans	**0.49**$\pm$0.01	**0.43**$\pm$0.01	**0.45**$\pm$0.01	**1.00**$\pm$0.00
	KkMEANS (Linéaire)	**0.49**$\pm$0.01	**0.43**$\pm$0.01	**0.45**$\pm$0.01	**1.00**$\pm$0.00
	KkMEANS $K_{i,j} = (1 + \langle x_i, x_j \rangle)^2$	0.40 ± 0.02	0.42 ± 0.04	0.41 ± 0.03	1.00 ± 0.00
	KkMEANS $K_{i,j} = (1 + \langle x_i, x_j \rangle)^3$	0.38 ± 0.02	0.41 ± 0.03	0.40 ± 0.02	1.00 ± 0.00
	KkMEANS (Gauss σ=0.066$\pm$0.001)	0.40 ± 0.03	0.40 ± 0.04	0.40 ± 0.03	1.00 ± 0.00
	OKM	0.17 ± 0.02	0.95 ± 0.01	0.28 ± 0.03	2.52 ± 0.19
	OKSETS (linéaire)	**0.41**$\pm$0.00	**0.60**$\pm$0.02	**0.49**$\pm$0.01	**1.30**$\pm$0.01
	KOKSETS $K_{i,j} = (1 + \langle x_i, x_j \rangle)^2$	0.40 ± 0.03	0.58 ± 0.02	0.48 ± 0.02	1.25 ± 0.05
	KOKSETS $K_{i,j} = (1 + \langle x_i, x_j \rangle)^3$	0.37 ± 0.02	0.60 ± 0.05	0.45 ± 0.01	1.22 ± 0.04
	KOKSETS (Gaussien $\sigma = 0.067$)	0.19 ± 0.04	0.90 ± 0.12	0.31 ± 0.04	1.54 ± 0.42

FIG. 4 – *Évolution du nombre de combinaisons considérées en fonction du nombre de clusters (k), sur Iris.*



5 Conclusion et perspectives

Nous avons abordé dans cette contribution la problématique de la classification recouvrante à travers la possibilité d'étendre de façon théoriquement bien fondée les modèles actuels vers le clustering à noyau. Nous nous sommes concentré pour le moment sur l'extension du modèle OKM en proposant une version ensembliste légèrement corrigée nommée OKSETS pour *Overlapping k-Sets* qui se prête aisément au passage aux noyaux via l'algorithme K-OKSETS (*Kernelized*-OKSETS). Les retours d'expériences ont permis de confirmer la faisabilité et l'intérêt de recourir aux noyaux en classification recouvrante.

Cependant, même si une solution théorique au problème est à présent effective et confirmée de manière pratique, une seconde phase d'étude sera indispensable afin de proposer une expérimentation plus étendue faisant intervenir d'avantage de jeux de tests, plus variés en terme de domaines et plus complexes en terme de recouvrements (nature et taille). Nous avons également observé que l'utilisation de noyaux offre les possibilités inattendues de réguler la taille des recouvrements et dans une certaine mesure d'utiliser les recouvrements pour détecter les contours des clusters ; il s'agira alors d'exploiter le potentiel de ces observations. Enfin, nous envisageons de la même manière d'adapter le modèle additif ALS aux noyaux puis, à moyen terme, de s'appuyer sur ces nouvelles avancées pour appréhender des problématiques nouvelles telles que le clustering spectral et semi-supervisé recouvrant.

Références

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

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

Ben N'Cir, C.-E. et N. Essoussi (2012). Overlapping patterns recognition with linear and non-linear separations using positive definite kernels. *International Journal of Computer Applications 56*(9), 1–8. Published by Foundation of Computer Science, New York, USA.

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

Cleuziou, G. (2008). An extended version of the k-means method for overlapping clustering. In *19th ICPR Conference*, Tampa, Florida, USA, pp. 1–4.

Depril, D., I. V. Mechelen, et T. F. Wilderjans (2012). Lowdimensional additive overlapping clustering. *Journal of Classification 29*(3), 297–320.

Dhillon, I. S. (2004). Kernel k-means, spectral clustering and normalized cuts. pp. 551–556. ACM Press.

Diatta, J. et B. Fichet (1994). From Apresjan hierarchies and Badelt-Dress weak hierarchies to quasi-hierarchies. In E. Diday et al. (Ed.), *New Approaches in Classification and Data Analysis*, pp. 111–118. Sringer-Verlag.

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

Filippone, M., F. Camastra, F. Masulli, et S. Rovetta (2008). A survey of kernel and spectral methods for clustering. *Pattern Recogn. 41*(1), 176–190.

Kulis, B., S. Basu, I. Dhillon, et R. Mooney (2009). Semi-supervised graph clustering : a kernel approach. *Machine Learning Journal 74*(1), 1–22.

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

Summary

Overlapping Clustering is currently a very attractive research domain that consists in structuring a dataset into clusters of similar data but allowing overlaps between clusters. Among the existing methods we focus our study on the ones based on least-square criteria and we notice that their kernelization is a tedious task. To solve the observed limitations, a new model is proposed using a set theoretical modeling of the cluster combinations. The new model and its associated algorithm are assessed on a preliminary empirical study that confirms the expectations.