Structuration de données par apprentissage non-supervisé : applications aux données textuelles

Guillaume Cleuziou

Université d'Orléans

Laboratoire d'Informatique Fondamentale d'Orléans

Discipline : Informatique

Habilitation à Diriger les Recherches

Soutenue le 16 décembre 2015 par

Guillaume Cleuziou

Structuration de données par apprentissage non-supervisé : applications aux données textuelles

Composition du jury :

Président du jury :	Pr. Christine Largeron	Université Jean Monnet (Saint-Etienne)
Rapporteurs :	Pr. Massih-Reza Amini	Université Joseph Fourrier (Grenoble)
	Pr. Vladimir Makarenkov	Université du Québec à Montréal
	Pr. Djamel A. Zighed	Université Lumière (Lyon 2)
Examinateurs :	Pr. Christine Largeron	Université Jean Monnet (Saint-Etienne)
	Pr. Marcilio de Souto	Université d'Orléans
	Pr. Rosanna Verde	Seconda Università di Napoli
	Pr. Christel Vrain	Université d'Orléans

Recherches réalisées au
Laboratoire d'Informatique Fondamentale d'Orléans
Université d'Orléans
Bât. 3IA – 6, rue Léonard de Vinci
45 067 ORLÉANS Cedex 2

Table des matières

Introduction

"Le monde n'a peut-être pas de sens, mais il a des structures, et tout est là."

Jean-Claude Clari, Romancier québécois.

Ce qui différencie l'homme de l'animal est qu'il a conscience de son existence. Son évolution s'est manifestée par le passage d'une organisation guidée par l'instinct et subissant le monde vers une structuration active de son environnement accompagnée d'un besoin de comprendre ce monde qui l'entoure.

L'homme a été acteur dans la structuration de son environnement social puis économique, plus ou moins consciemment parfois. La société, originellement gouvernée par de simples relations d'appartenances (communautés, familles) et de dominations (tribues), s'est dotée de structures de gouvernance à l'échelle de la planète (nations). La communication entre les hommes a progressivement évolué avec le développement de languages structurés. Enfin, l'homme a organisé un système économique structurant les relations d'échanges à l'échelle des individus ou des peuples.

L'homme est aussi spectateur d'un monde dont il cherche à comprendre les structures gouvernantes. Dans une quête perpétuelle de compréhension il a développé les mathématiques afin de proposer des modèles susceptibles d'expliquer ses observations physiques, chimiques ou biologiques.

L'informatique est, à la base, un outil que l'homme a créé pour le suppléer dans la réalisation de tâches nombreuses et répétitives (robotisation, calculs) puis il s'en est servi pour simuler et prédire à partir des modèles qu'il proposait et pour l'assister dans sa propre structuration du monde (informatique de gestion, communication). Aujourd'hui c'est l'informatique qui structure le monde à travers les réseaux sociaux, par des liens virtuels que l'homme a parfois du mal à maîtriser. Mais l'informatique permet aussi d'aller plus loin dans la compréhension du monde ; de la simple simulation ou prédiction, nous avons été en mesure de proposer des algorithmes capables d'apprendre des modèles de structuration à partir de l'observation du monde.

Notre rôle en tant qu'informaticiens, spécialisés en fouille de données et en apprentissage automatique, est à la fois de (1) proposer de nouveaux modèles de structuration, (2) de les mettre en œuvre informatiquement afin qu'ils rendent comptent de la structure des choses et (3) de tenter de les apprendre automatiquement à partir d'observations (méta-modèles).

Les contributions que vous trouverez dans ce mémoire, synthétisent dix années de recherches guidées par cette vision de notre place dans la société et dans sa quête. À travers une *structuration* en deux parties et cinq chapitres nous tenterons de contribuer modestement aux trois attentes exprimées.

La première partie se concentre sur la proposition de nouveaux modèles théoriques de structuration complexe en classes d'individus ; il s'agit alors de proposer des structures de classification plus proches de l'organisation réelle des données telles qu'observées (classification recouvrante, formes symétriques), de rendre ces structures à la fois ro-

bustes (tolérance au bruit) et manipulables par l'homme (visualisation, paramétrage) et enfin d'être en mesure de les expliquer (sémantique des classes).

La seconde partie s'intéresse à une donnée particulière qui structure la communication entre les hommes : la langage et plus précisément la donnée textuelle. Dans un premier temps nous nous concentrons sur la mise en œuvre de modèles rendant compte de la structure thématique d'une collection de textes courts dans un contexte de recherche d'information. Le dernier chapitre, enfin, présente un méta-modèle permettant d'apprendre automatiquement un modèle de structuration sémantique d'un ensemble de termes.

Première partie

Classification par apprentissage non-supervisé

Sommaire

En fouille de données, le succès d'une méthode tient au fait qu'elle permet de répondre par un algorithme intuitif à un besoin pratique bien théorisé. Dans leur article [Wu et al., 2008], les auteurs identifient dix algorithmes présentés comme les plus influents dans le domaine du data mining ; tous vérifient les conditions du succès évoquées ci-dessus. Les premiers d'entre-eux sont les algorithmes C4.5, k-moyennes, SVM, Apriori et EM qui répondent tous à un besoin pratique attesté (prédire une classe, découvrir une typologie, extraire des associations fréquentes), par un algorithme généralement assez intuitif (découpage arborescent, réallocation dynamique, recherche par sélection/agrégation, etc.) et conçu de manière à résoudre un problème bien posé (optimisation numérique, statistique, combinatoire, etc.). C'est avec cet éclairage que nous débutons la première partie de ce mémoire qui présente un ensemble de contributions pour la tâche de clustering.

Le clustering est généralement présenté comme un problème de structuration d'un ensemble d'objets en classes, de telle sorte que les objets dans une même classe doivent être similaires entre eux, et les objets appartenant à des classes différentes doivent être dissimilaires. Cette tâche correspond à un besoin naturel d'étudier la typologie afin de simplifier/résumer un ensemble de données qu'il serait impossible d'appréhender dans sa globalité (e.g. par visualisation). Depuis les années 1950, de nombreux algorithmes ont été proposés dans le but d'automatiser cette tâche de classification. Progressivement les besoins ont évolués et les méthodologies se sont adaptées à la forme des données à traiter (numériques, symboliques, relationnelles, graphes, textes, séquences, etc.), à leur quantité, leur qualité ou leur disponibilité (big data, bruit, données manquantes, flux de données, etc.), à la forme des résultats attendus (classes convexes ou non, partitions strictes ou floues, structures arborescentes, cartes visuelles, explication des classes, etc.).

Durant ces dix dernières années nous avons travaillé à la proposition de nouvelles approches ou nouvelles extensions permettant de répondre à des situations concrètes qui nécessitent de remettre en cause la problématique du clustering telle qu'énoncée initialement : l'organisation des objets en classes recouvrantes (ou chevauchantes) remet en cause le principe de classes bien séparées puisque certains individus peuvent appartenir à des classes différentes ; l'autre principe de forte similarités intra-classe est quant à lui caduc lorsque l'on recherche par clustering des formes spécifiques telles que des clusters symétriques ; enfin les deux principes à la fois peuvent être partiellement transgressés pour une tâche de clustering visant à la recherche d'une solution consensuelle (clustering multiple).

Nous présentons dans cette première partie, une synthèse des contributions réalisées pour répondre à la problématique du clustering recouvrant et du clustering symétrique. Nous évoquons également dans l'Annexe A, un travail réalisé dans un contexte d'encadrement doctoral portant sur la classification non-supervisée multi-vues.

Le premier chapitre est dédié à la présentation de l'algorithme Okm, formalisé comme une généralisation de k-moyennes et qui permet d'extraire directement une classification recouvrante d'un ensemble d'objets sans recourir à une étape préliminaire de partitionnement strict ou flou.

La généricité du modèle sous-jacent à Okm est ensuite montrée dans le second chapitre où tout une série d'extensions est proposée. Chacune de ces extensions correspond à un réel besoin en terme d'usage (e.g. robustesse, contrôle, sémantique ou visualisation des classifications) et nécessite le recours à des outils et formalismes complémentaires

issus de l'analyse de données, de la fouille de données ou de l'intelligence artificielle (e.g. l'analyse formelle de concepts, l'analyse de données symboliques, les méthodes d'optimisation convexe, les approches neuronales, les méthodes à noyau).

Enfin, le troisième chapitre présente une contribution récente dans un domaine lui aussi plutôt récent que constitue la reconnaissance de formes symétriques par clustering. Sans s'éloigner du principe de clustering par réallocation dynamique, qui constitue un principe de base sur lequel repose une grande partie des contributions présentées dans ce mémoire, nous montrons qu'il est possible de généraliser les approches existantes en leur permettant de rechercher des clusters symétriques dans tout espace de projection induit par un noyau.

1 Modèle et algorithme de classification recouvrante

Il est fréquent, dans les applications du monde réel, que les données s'organisent naturellement en groupes non-disjoints. Il est alors nécessaire de rechercher une couverture de l'ensemble des données, plutôt qu'une partition. Il s'agit là d'un domaine de recherches appelé "classification recouvrante" (*overlapping clustering*) qui a été d'abord étudié dans les années 80 [Shepard and Arabie, 1979, Diday, 1987], puis réinvesti depuis une dizaine d'années à travers des applications aux données biologiques et textuelles entre autres [Banerjee et al., 2005, Depril et al., 2008]. Le champ d'application s'est alors considérablement étendu de sorte que la classification recouvrante est aujourd'hui utilisée pour de nombreuses tâches pratiques :

- analyse des **réseaux sociaux** : afin de détecter des communautés sociales chevauchantes [Gregory, 2008, Tang and Liu, 2009, Wang et al., 2010, Zhao and Zhang, 2011, Fellows et al., 2011],
- **recherche d'information** : par exemple en classification automatique d'extraits vidéos selon leur(s) genre(s) [Snoek et al., 2006], ou d'extraits musicaux selon leur(s) émotion(s) [Wieczorkowska et al., 2006] ou enfin de documents textuels selon leur(s) thématique(s) [Gil-García and Pons-Porrata, 2010, Pérez-Suárez et al., 2013],
- **bioinformatique** : pour la découverte de classes de gènes ou protéines selon la/les fonction(s) métabolique(s) à laquelle/auxquelles ils participent [Segal et al., 2003, Banerjee et al., 2005, Becker et al., 2012]
- **réseaux de capteurs** : pour le routage, la localisation ou encore la synchronisation de protocoles [Youssef et al., 2009].

Observons enfin que, contrairement au clustering flou, la classification recouvrante suppose que chaque observation peut totalement appartenir à plusieurs classes, sans recourir à un degré d'appartenance.

1.1 Aperçu du domaine

De manière simplifiée, deux types d'approches ont été proposées pour réaliser une classification recouvrante : les approches heuristiques et les solutions théoriques. Dans le premier cas il s'agit soit de redéfinir un nouveau processus intuitif de clustering indépendamment d'un quelconque modèle mathématique sous-jacent [Pantel and Lin, 2002, Cleuziou et al., 2004], soit de modifier la sortie d'un algorithme standard de clustering (une partition floue par exemple) de manière à forcer la construction de classes recouvrantes [Lingras and West, 2004, Zhang et al., 2007]. Ces approches heuristiques conduisent souvent à de bons résultats en pratiques mais leur amélioration ou leur extension restent limités du fait de l'absence de modèle théorique sous-jacent.

Les solutions théoriques, quant à elles, correspondent souvent à des extensions de modèles de classification usuels tels que les approches hiérarchiques, de décomposition de graphes, ou encore les méthodes par réallocation dynamique. Par exemple, les pyramides [Diday, 1987] ou plus généralement les hiérarchies faibles [Bertrand and Janowitz, 2003] constituent des structures classificatoires chevauchantes qui étendent les structures hiérarchiques usuelles ; de manière résumée, les méthodes classificatoires sous-jacentes permettent de construire des structures dont les distances (e.g. robinsoniennes) induites sont

plus proches du tableau de distances initial que les distances ultramétriques induites par des hierarchies traditionnelles. Cependant, on peut montrer que ces pseudo-hiérarchies chevauchantes sont à la fois restrictives en terme de recouvrements autorisés et complexes à générer et à visualiser.

Les modèles de mélanges recouvrants [Banerjee et al., 2005, Heller and Ghahramani, 2007, Fu and Banerjee, 2008] constituent des extensions de l'algorithme originel EM (*Expectation Maximization*) [Dempster et al., 1977]. Ces modèles ont été initialement motivés par des applications en bioinformatique pour la classification de gènes participant à de multiples processus métaboliques. Ils se fondent sur l'hypothèse que chaque observation est le résultat d'une combinaison de distributions [1] : cette combinaison pouvant être additive [Banerjee et al., 2005] ou multiplicative [Heller and Ghahramani, 2007, Fu and Banerjee, 2008]. Ces choix de combinaison ne sont pas sans lien avec l'application envisagée, en effet les auteurs justifient ce choix par le fait qu'un gène cumule son expression en fonction des processus auxquels il participe. Nous verrons plus tard que cette modélisation (cumulative) des recouvrements, même si elle semble convenir au traitement des données génétiques, n'est pas universelle.

Nous nous sommes intéressés à un autre type d'approches de classification, les méthodes fondées sur la minimisation d'un critère d'inertie. Dans le cas non-recouvrant, on sait qu'un algorithme de type "nuées dynamiques" peut être vu comme un cas particulier d'un modèle de mélanges [Celleux and Govaert, 1992]. La même analogie peut être faite entre les modèles de mélanges recouvrants additifs [Banerjee et al., 2005] et les algorithmes fondés sur la minimisation d'un moindre carré additif [Shepard and Arabie, 1979, Mirkin, 1987, Depril et al., 2008]. Cependant, comme nous allons le voir dans ce chapitre, ce dernier type d'approches permet d'explorer plus facilement de nouvelles formes de combinaison de classes pour modéliser des recouvrements (e.g. combinaisons géométriques) et ouvre la voie à de nombreuses extensions (distances, données symboliques, cartes auto-organisatrices, etc.) tant la littérature est abondante et les connexions variées dans ce domaine.

1.2 OKM : modèle et algorithme

Nous avons proposé la méthode de classification recouvrante OKM (*Overlapping k-means*) [Cleuziou, 2006, Cleuziou, 2007, Cleuziou, 2008]. Celle-ci se fonde sur un modèle d'inertie favorisant la structuration d'un ensemble de données en classes recouvrantes, composées d'individus similaires.

Étant donnée une matrice $X = (x_1, \ldots, x_N)^T$ décrivant N observations dans $\mathbb{R}^M$ à structurer en K classes, OKM se fonde sur le critère des moindres carrés suivant

$$J_{okm}(\Pi, C) = \sum_{x_i \in X} \|x_i - \frac{\sum_k \pi_{i,k} c_k}{\sum_k \pi_{i,k}}\|^2. \quad (1)$$

Dans cette expression, $\|.\|$ désigne la distance euclidienne et $\Pi = (\pi_1, \ldots, \pi_K)$ est une matrice binaire $(N \times K)$ formalisant une classification recouvrante, c'est-à-dire une

1. Classiquement des lois gaussiennes ou plus généralement des lois exponentielles.

répartition des N individus parmi les K classes telle que

$$\pi_{i,k} = \begin{cases} 1 \text{ si } x_i \in \pi_k \\ 0 \text{ sinon.} \end{cases} \quad \text{avec } \sum_k \pi_{i,k} \geq 1 \ \ \forall i = 1, \ldots, N. \tag{2}$$

Notons que dans la suite nous utiliserons la notation π_k pour désigner soit l'ensemble des individus de la $k^{\text{ième}}$ classe soit la $k^{\text{ième}}$ colonne de la matrice Π selon le contexte. Enfin, la variable $C = (c_1, \ldots, c_K)^T$ correspond à une matrice $(K \times M)$ de profils de classes dans $\mathbb{R}^M$. Nous choisissons volontairement de ne pas utiliser le terme de "représentant" ou de "prototype" de classe pour c_k car, du fait des recouvrements et contrairement aux modèles habituels (type nuées dynamiques), les profils de classes pourront se positionner de façon excentrée voire à l'extérieur de l'enveloppe convexe définie par les individus de la classe concernée.

Le critère objectif J_{okm} (1), qui constitue le point de départ de cette première partie, doit être interprété (cf. Figure 1) comme une somme d'erreurs : chacune de ces erreurs étant quantifiée par la distance euclidienne entre un individu x_i et l'isobarycentre des profils des classes auxquelles il appartient ($\frac{\sum_k \pi_{i,k} c_k}{\sum_k \pi_{i,k}}$).

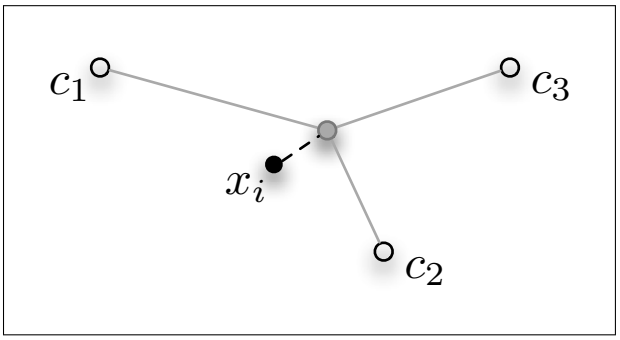


Figure 1 – Illustration de l'erreur induite par un individu x_i affecté à trois classes de profils respectifs c_1, c_2 et c_3. La distance euclidienne entre x_i et l'isobarycentre des trois profils est matérialisée par le segment en pointillés.

La minimisation du critère (1) tend à structurer les données en classes dont les recouvrements correspondent à une interprétation visuelle ou géométrique des intersections entres les classes. Dans la suite, ce type de modélisation des recouvrements est appelé "modélisation géométrique", par opposition aux modélisations additives (ou cumulatives) déjà mentionnées, qui se justifient davantage par des considérations fonctionnelles et/ou logiques liées aux types des données considérés (e.g. données biologiques). Le modèle de clustering additif proposé par [Mirkin, 1987] est analytiquement très proche du modèle Okm :

$$J_{mirkin}(\Pi, C) = \sum_{x_i \in X} \|x_i - \sum_k \pi_{i,k} c_k\|^2 \tag{3}$$

Ce critère peut de plus être réécrit aisément en $\|X - \Pi C\|_F^2$ (où $\|.\|_F$ désigne la norme de Froebenius). La réécriture matricielle du critère Okm nécessiterait quant à elle une normalisation des lignes de la matrice Π :

$$J_{okm}(\Pi, C) = \|X - \tilde{\Pi} C\|_F^2 \ \text{ où } \ \tilde{\pi}_{i,k} = \frac{\pi_{i,k}}{\sum_l \pi_{i,l}}. \tag{4}$$

Ce rapprochement analytique d'une part facilitera la comparaison expérimentale de la modélisation géométrique que nous proposons avec la modélisation additive proposée jusqu'alors et d'autre part offre un cadre théorique commun à la résolution du problème d'optimisation qui s'en suit.

Observons que les critères de moindres carrés (1) et (3) constituent tous les deux des généralisations du critère d'inertie (5) utilisé dans k-moyennes [Forgy, 1965] et rappelé ci-après :

$$J_{kmeans}(\Pi, C) = \sum_{k=1}^{K} \sum_{x_i \in \pi_k} \|x_i - c_k\|^2 \tag{5}$$

auquel les deux modèles recouvrants se ramènent lorsqu'on contraint chaque individu à n'appartenir qu'à une seule classe ($\sum_k \pi_{i,k} = 1$).

Enfin, la forme de la matrice Π permet d'appréhender facilement la taille de l'espace de solutions selon que l'on considère une classification stricte : K^N classifications possibles ou une classification recouvrante : $(2^K - 1)^N$ recouvrements possibles.

La recherche d'une solution minimisant un critère de moindres carrés (tel que les critères (1), (3) et (5) précédents) passe par un processus itératif en deux étapes, chacune visant la minimisation du critère selon l'une des deux variables Π et C et considérant la seconde variable fixe. Ce processus assure la convergence vers un optimum (ici un minimum) local du critère objectif, dépendant de l'initialisation des variables.

Si les sous-problèmes issus des deux étapes de minimisation peuvent être résolus assez aisément dans le cas d'une classification stricte (k-moyennes), la généralisation aux classifications recouvrantes n'est pas triviale est nécessite de définir une stratégie qui soit efficace, étant donnée la taille de l'espace de solutions, et qui préserve la convergence du critère objectif vers un minimum local.

Affectation des individus. Cette étape est résolue simplement dans k-moyennes par l'affectation de chaque individu à la classe dont le profil est le plus proche de l'individu, au sens de la distance euclidienne : $\pi_{i,k} = 1$ *ssi* $c_k = \arg\min_{c_l} \|x_i - c_l\|^2$.

Dans le contexte de la classification recouvrante, la recherche de l'affectation optimale nécessiterait de considérer, pour chaque individu x_i, l'ensemble des $(2^K - 1)$ combinaisons d'affectations possibles, soit toutes les possibilités d'instancier π_i, en excluant la non-affectation. Cette solution coûteuse est celle retenue dans l'algorithme ALS [Depril et al., 2008] pour la minimisation du critère J_{mirkin} (3). Nous observerons dans le tableau d'expérimentations (Table 1) qu'ALS n'est plus envisageable dès que le nombre de classes attendu atteint 6 ou plus.

Nous avons proposé dans OKM une heuristique d'exploration d'un sous-ensemble de combinaisons d'affectations pour un individu x_i donné, qui consiste à affecter en premier lieu x_i à la plus proche classe puis à considérer itérativement les affectations aux autres classes (par ordre de proximité toujours) tant que l'erreur associée à x_i diminue (cf. Algorithme 1).

Mise à jour des profils de classes. Contrairement au critère (5) utilisé dans k-moyennes, les combinaisons (additives ou géométriques) des profils, telles que définies

Algorithme 1 Stratégie d'affectation dans Okm

procedure Affectation(x_i, π_i, C)
  Initialisation : $\pi'_i = (0, \ldots, 0)$
  Recherche de la plus proche classe : $l = \arg\min_{k=1,\ldots,K} \|x_i - c_k\|^2$
  repeat
    Affectation : $\pi'_{i,l} = 1$
    Recherche de la plus proche classe suivante : $l = \arg\min_{k|\pi'_{i,k}=0} \|x_i - c_k\|^2$
  until Augmentation de l'erreur : $\|x_i - \frac{c_l + \sum_k \pi'_{i,k} c_k}{1+\sum_k \pi'_{i,k}}\|^2 \geq \|x_i - \frac{\sum_k \pi'_{i,k} c_k}{\sum_k \pi'_{i,k}}\|^2$
  if $\|x_i - \frac{\sum_k \pi'_{i,k} c_k}{\sum_k \pi'_{i,k}}\|^2 \leq \|x_i - \frac{\sum_k \pi_{i,k} c_k}{\sum_k \pi_{i,k}}\|^2$ **then**
    Acceptation de la nouvelle (meilleure) affectation : $\pi_i \leftarrow \pi'_i$
  else
    Conservation de l'ancienne (meilleure) affectation
  end if
  Return π_i
end procedure

dans les modèles recouvrants (1) et (3), ne permettent pas une décomposition du critère objectif en une somme de termes indépendants pour chaque profil de classe. Une solution optimale du problème global peut être obtenue par $C = \tilde{\Pi}^+ X$ pour Okm et $C = \Pi^+ X$ pour le modèle recouvrant additif de Mirkin, où Π^+ désigne la pseudo-inverse $\Pi^+ = (\Pi^T \Pi)^{-1} \Pi^T$.

Comme alternative à une décomposition en valeurs singulières (qui peut s'avérer coûteuse notamment lorsque le nombre de classes K est grand), nous avons proposé une mise à jour successive des profils de classes dans le cas du modèle Okm. Il s'agit alors de calculer successivement pour chaque classe π_l, son profil c_l optimal en considérant les autres profils fixés. Nous obtenons alors le résultat suivant :

$$\nabla J_{okm}(c_l) = 0 \Leftrightarrow c_l = \frac{\sum_{x_i \in \pi_l} \frac{1}{|\pi_i|^2} \hat{x}_i^l}{\sum_{x_i \in \pi_l} \frac{1}{|\pi_i|^2}} \text{ où } |\pi_i| = \sum_k \pi_{i,k} \text{ et } \hat{x}_i^l = |\pi_i| x_i - \sum_{k \neq l} \pi_{i,k} c_k. \quad (6)$$

De manière plus intuitive, le centre c_l optimal correspond au barycentre du nuage $\{(\hat{x}_i^l, \frac{1}{|\pi_i|^2})\}_{x_i \in \pi_l}$, où chaque point $\hat{x}_i^l$ correspond à la déviation [2] de l'un des individus x_i de π_l pondéré inversement selon le (carré du) nombre de classes auxquelles il appartient.

Précisons qu'à l'itération t de l'algorithme Okm, le calcul d'un nouveau profil de classe $c_l^{(t)}$ repose d'une part sur la connaissance des profils déjà mis à jour $\{c_k^{(t)}\}_{k<l}$ et d'autre part sur les profils qu'il reste à calculer $\{c_k^{(t-1)}\}_{k>l}$. Autrement dit, les points de $\mathbb{R}^M$ utilisés sont calculés ainsi :

$$\hat{x}_i^l = |\pi_i| x_i - \sum_{k<l} \pi_{i,k} c_k^{(t)} - \sum_{k>l} \pi_{i,k} c_k^{(t-1)}. \quad (7)$$

Finalement, l'algorithme Okm (Algorithme 2) procède par itérations successives des

2. Position du profil de classe c_l permettant à un individu d'annuler l'erreur qu'il induit sur le système.

procédures d'affectation (multiple) des individus puis de mise à jour des profils de classes telles que proposées précédemment. Cet algorithme ne permet pas de minimiser le critère

Algorithme 2 OKM

procedure OKM(X, K, T)
  $t \leftarrow 0$
  Initialisation aléatoire des profils : $c_k^{(t)} \leftarrow Random(x_1, \ldots, x_N), \quad k = 1 \ldots K$
  Initialisation des classes : $\forall i,\ \pi_i^{(t)} \leftarrow$ AFFECTATION$(x_i,\ \pi_i^{(t)},\ C^{(t)})$
  repeat
    $t \leftarrow t + 1$
    for $l = 1 \ldots K$ **do** Mise à jour des profils de classes par (6) et (7)
      $$c_l^{(t)} \leftarrow \frac{1}{\sum_{x_i \in \pi_l^{(t-1)}} |\pi_i^{(t-1)}|^2} \sum_{x_i \in \pi_l^{(t-1)}} \frac{1}{|\pi_i^{(t-1)}|^2} \left(|\pi_i^{(t-1)}| x_i - \sum_{k<l} \pi_{i,k}^{(t-1)} c_k^{(t)} - \sum_{k>l} \pi_{i,k}^{(t-1)} c_k^{(t-1)} \right)$$
    end for
    for $i = 1 \ldots N$ **do** Affectation de chaque individu
      $\pi_i^{(t)} \leftarrow$ AFFECTATION$(x_i,\ \pi_i^{(t-1)},\ C^{(t)})$
    end for
  until $\Pi^{(t)} = \Pi^{(t-1)}$ ou $t = T$ Conditions d'arrêt
  Return $\Pi^{(t)}$
end procedure

objectif J_{okm} (1) localement sur aucun des deux sous-problèmes associés à chacune des variables Π et C. Cependant il en assure la décroissance au sens large, sur un ensemble fini de solutions. L'algorithme OKM converge ainsi vers une solution proche d'un optimum local du critère des moindres carrés initial.

1.3 Discussion et positionnement

On remarquera que l'algorithme OKM (tout comme les algorithmes ALS et MOC) constitue une généralisation de l'algorithme des k-moyennes, dans le sens où si on autorise chaque individu à n'appartenir qu'à une seule classe :

1. la procédure proposée (Algorithme 1) réalisera l'affectation à la classe de plus proche profil,
2. la définition de chaque nouveau profils (6) revient exactement au centre de gravité de chaque classe ($|\pi_i| = 1$ et $\hat{x}_i^l = x_i$ pour tout i et l).

À la différence des méthodes existantes, OKM propose une résolution algorithmique efficace (en $O(NTK\log(K))$ associée à une modélisation géométrique, plutôt qu'additive, des recouvrements. En ce sens cette première contribution représente une réponse pragmatique à des besoins réels en fouille de données, comme en témoignent dores et déjà les usages qui sont faits de OKM, par exemple, dans les domaines suivants :

— biologie [Régis et al., 2012, Mishra, 2012],
— réseaux de capteurs [Fouchal et al., 2012],

— recherche d'information [Rizoiu et al., 2011, Dupuch et al., 2013],
— réseaux sociaux [Forestier et al., 2010].

En tant que généralisation des k-moyennes, OKM hérite des contraintes bien connues des approches par réallocation dynamique, parmi lesquelles : le problème de la sélection de modèle (choix du nombre K de classes) et le caractère non-déterministe de l'algorithme de résolution.

OKM n'a pas été étudié sous l'angle de la sélection de modèle jusqu'à présent [3], néanmoins une étude approfondie sur l'adaptation de critères usuels (e.g. Bayesian Information Criterion [Schwarz et al., 1978]) ou sur l'extension de OKM vers des modèles bayésiens non-paramétriques qui ne requièrent pas la connaissance de K [Kulis and Jordan, 2011] est à envisager. En revanche, nous avons étudié dans [Ben N'cir et al., 2014] le choix du modèle de recouvrement (additif ou géométrique) et montré d'une part que si la modélisation géométrique est insensible au pré-traitement des données, un simple centrage des données peut, en revanche, rendre un recouvrement additif proche d'une modélisation géométrique ; nous avons observé empiriquement qu'un modèle de recouvrement géométrique est, au pire, équivalent à un modèle additif et, généralement, plus adapté à la sémantique des recouvrements quelque soit le type d'application (vidéo, musique, image, biologie).

Concernant l'étape d'initialisation, qui rend l'algorithme non-déterministe, même si des heuristiques ont été étudiées pour k-moyennes qui pourraient certainement être reprises ou adaptées au clustering recouvrant (e.g. [Redmond and Heneghan, 2007]), nous proposons en pratique d'effectuer plusieurs réalisations de l'algorithme avec différentes initialisations puis de conserver la classification correspondant à la plus faible valeur du critère objectif. Cette solution classique permet d'augmenter les chances d'approcher d'un optimum global du critère objectif et reste envisageable pour OKM du fait de sa faible complexité.

Enfin, il est intéressant de noter que, outre les approches MOC et ALS, d'autres modèles existent qui présentent de fortes ressemblances analytiques avec OKM mais sont issus de théories différentes. En théorie des fonctions de croyance, [Masson and Denœux, 2008] ont proposé l'algorithme ECM (*Evidential c-Means*) qui pourrait être considéré comme une version floue de OKM dans laquelle chacune des 2^K combinaisons de classes est explicitement considérée dans la fonction objective. Le résultat de l'algorithme est une *partition crédale* modélisée par une matrice d'appartenance floue $N \times 2^K$ où chaque élément quantifie la crédibilité de l'appartenance d'un individu x_i à un ensemble de classes (e.g. $\{\pi_2 \cup \pi_3\}$). En dépit de l'aspect combinatoire, qui limite la pratique de ce modèle à un petit nombre de classes (au plus 6 en général), l'utilisation des fonctions de croyance est intéressante dans le contexte du clustering en ceci qu'elles génèrent une classification recouvrante modélisant une incertitude sur l'appartenance d'un individu x_i à l'une des classes π_2 *ou* π_3, plutôt qu'une information d'appartenance à l'ensemble des classes du recouvrement π_2 *et* π_3. Cette même notion d'incertitude est également reprise, de manière topologique plutôt que probabiliste, via la théorie des ensembles approximatifs qui a conduit à l'algorithme RCM (*Rough c-Means*) [Lingras and West, 2004] ; chaque cluster π_k est alors décomposé en une zone de certitude $\underline{\pi_k}$ (borne inférieure) et une zone du possible $\overline{\pi_k}$ (borne supérieure) de telle manière

3. Dans la suite nous considèrerons toujours connu le nombre de classes souhaitées.

que la différence des deux ensembles modélise une aire d'incertitude $bn(\pi_k)$ (*boundary area*). Ces zones sont apprises par un processus classique de réallocations itératives visant à minimiser conjointement les inerties des zones de certitude et d'incertitude de chaque classe, tout en respectant certaines contraintes telles que le non recouvrement entre bornes inférieures et l'obligation pour tout individu x_i situé dans une intersection de plusieurs classes d'appartenir à leur borne supérieure.

Nous sommes convaincus que chacune des formalisations précédentes apporte son lot de bénéfice pour répondre à la tâche de classification recouvrante. Il serait tout à fait intéressant de réaliser une étude plus large sur ces travaux émanant de sous-communautés scientifiques différentes, afin de faire émerger d'une part les conditions d'un cadre théorique unifié et d'autre part les contextes d'application plutôt favorables à chacune des approches.

1.4 Évaluation préliminaire

Comme toute méthode non-supervisée, l'évaluation des classifications produites par des approches recouvrantes telles que OKM, ALS ou MOC, reste un problème ouvert qui, de plus, est peu armé en terme d'indices d'évaluation par rapport aux approches de classification strictes, notamment en ce qui concerne les mesures d'évaluation externe (i.e. par rapport à une classification pré-établie) [Halkidi et al., 2002]. Nous avons proposé dans [Cleuziou et al., 2012] une adaptation du critère de Rand corrigé [Hubert and Arabie, 1985] pour la comparaison de deux classifications recouvrantes. Ce nouveau critère considère la possibilité que chaque paire d'individus puisse être observée plusieurs fois dans chacune des classifications du fait de leur caractère recouvrant. Ce critère est cependant coûteux à calculer et ne représente pas actuellement une mesure de référence. Afin de positionner empiriquement la méthode OKM par rapport à d'autres approches recouvrantes, strictes ou floues, nous utiliserons le plus souvent la mesure $F - Bcubed$ proposée par [Amigó et al., 2009], dont une adaptation permet de comparer deux classifications recouvrantes par combinaison des termes de précision et de rappel (multiples) tels que définis dans Définition 1.

Définition 1 *Soient X un ensemble d'individus à structurer en classes, $\Pi = (\pi_1, \ldots, \pi_K)$ le résultat d'une classification recouvrante et $\Delta = (\delta_1, \ldots \delta_{K'})$ un recouvrement dit de référence. Les termes de précision et de rappel comparant Π relativement à Δ sont définis respectivement par :*

$$Précision\ BCubed(\Pi, \Delta) = Avg_{x_i}\left[Avg_{x_j|\pi_i\cap\pi_j\neq\emptyset}\left[\frac{Min(|\pi_i \cap \pi_j|, |\delta_i \cap \delta_j|)}{|\pi_i \cap \pi_j|}\right]\right]$$

$$Rappel\ BCubed(\Pi, \Delta) = Avg_{x_i}\left[Avg_{x_j|\delta_i\cap\delta_j\neq\emptyset}\left[\frac{Min(|\pi_i \cap \pi_j|, |\delta_i \cap \delta_j|)}{|\delta_i \cap \delta_j|}\right]\right]$$

La *précision multiple* évalue la qualité des associations d'individus réalisées par Π, en considérant que deux individus ne doivent pas être associés plus de fois avec Π que dans la classifications de référence Δ. De manière duale, le *rappel multiple* évalue la proportion d'associations de Δ effectivement retrouvées par Π, en considérant que deux individus ne doivent pas être associés moins de fois dans Π que dans Δ.

La mesure $F - Bcubed$ finale, combine les termes de précision (P) et de rappel (R) par la F-mesure standard :

$$F_\beta(R, P) = \frac{(1 + \beta^2).P.R}{(\beta^2.P + R)} \tag{8}$$

où le choix β=1 (que nous ferons par la suite) conduit à une moyenne harmonique des deux critères P et R. Bien que cette mesure d'évaluation soit réputée sensible aux déséquilibres entre les classes [de Souto et al., 2012], nous l'utiliserons de manière complémentaire avec des visualisations, lorsque les données le permettent, non pas dans une optique d'expérimentations et de comparaisons à grande échelle mais plutôt avec l'objectif d'attester de la validité du modèle et de l'algorithme proposés.

Nous illustrons en Figure 2 le résultat produit par OKM (K=2) sur des données générées artificiellement selon deux gaussiennes (équilibrées) de 500 individus chacune, dans $\mathbb{R}^2$.

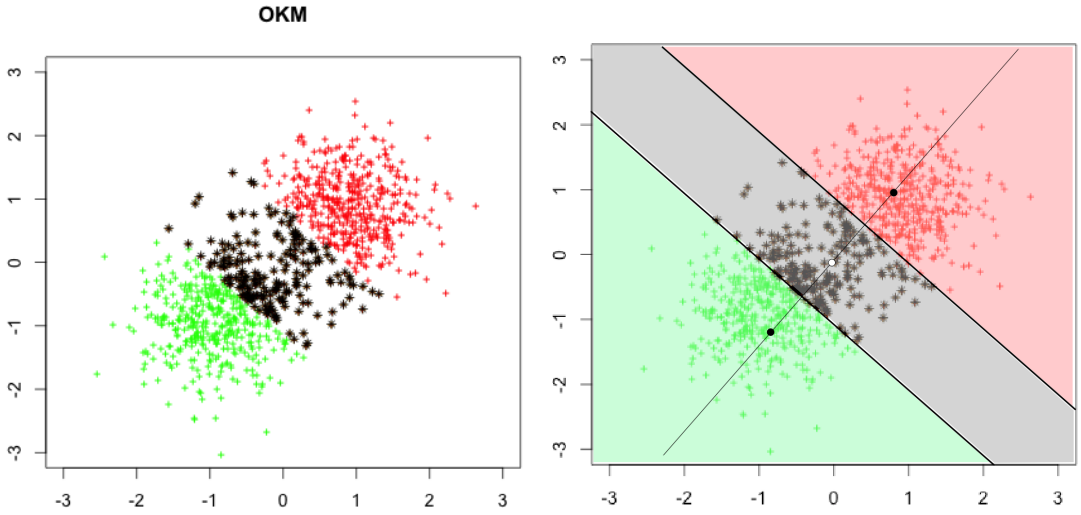


FIGURE 2 – Résultat de OKM sur données artificielles (gauche) et interprétation par cellules de Voronoï (droite).

On observera en particulier la large bande correspondant au recouvrement des deux classes, retrouvées par OKM. La figure de droite offre une interprétation d'un partitionnement de l'espace au moyen de cellules de Voronoï : les deux profils de classe sont légèrement excentrés des moyennes attendues, tandis que l'isobarycentre des deux profils joue comme un nouveau profil (non-libre), à l'origine d'une troisième cellule de Voronoï, totalement dépendante des deux premières et matérialisant la zone de recouvrement.

Nous présentons enfin dans le Tableau 1 une évaluation comparative et quantitative sur trois jeux de données de l'UCI [D.J. Newman and Merz, 1998] et de Mulan [Tsoumakas et al., 2011] dont nous donnons une description synthétique ci-après :

— Iris : 150 individus (fleurs), décrits par 4 attributs numériques et répartis en 3 classes mono-label,

— Scene : 2404 individus (images), décrits par 294 attributs numériques et répartis en 6 classes multi-labels (faible taux de recouvrement : 1.07),
— Yeast : 2417 individus (protéines), décrits par 103 attributs numériques et répartis en 14 classes multi-labels (fort taux de recouvrement : 4.23).

Le taux de recouvrement correspond au nombre moyen de classes (ou labels) par individu :

$$TxOverlap(\Pi) = \frac{\sum_{i=1}^{N}\sum_{k=1}^{K}\pi_{i,k}}{N}$$

TABLE 1 – Évaluation comparative de OKM sur trois jeux de données de l'UCI et de Mulan (centrés et réduits).

Données	Méthode	Précision	Rappel	FBcubed	Tx Overlap
Iris	k-Moyennes	**0.72**±0.05	0.77 ± 0.04	0.74 ± 0.02	**1.00**±0.00
	OKM	0.49 ± 0.08	**0.98**±0.01	0.65 ± 0.07	1.51 ± 0.09
	MOC	0.43 ± 0.03	0.95 ± 0.02	0.59 ± 0.02	1.63 ± 0.04
	ALS	0.41 ± 0.06	0.94 ± 0.04	0.57 ± 0.06	1.64 ± 0.11
	k-moyennes flou	0.71 ± 0.00	**0.98**±0.00	**0.82**±0.00	1.26 ± 0.00
Scene	k-Moyennes	**0.47**±0.02	0.46 ± 0.03	**0.46**±0.03	1.00 ± 0.00
	OKM	0.16 ± 0.02	0.94 ± 0.02	0.28 ± 0.03	2.48 ± 0.24
	MOC	0.19 ± 0.01	0.88 ± 0.01	0.32 ± 0.01	2.49 ± 0.04
	ALS	-	-	-	-
	k-moyennes flou	0.29 ± 0.00	0.78 ± 0.00	0.42 ± 0.00	**1.12**±0.08
Yeast	k-Moyennes	**0.81**±0.00	0.04 ± 0.00	0.07 ± 0.00	1.00 ± 0.00
	OKM	0.71 ± 0.01	0.61 ± 0.04	0.66 ± 0.02	**4.45**±0.21
	MOC	0.63 ± 0.01	**0.78**±0.02	**0.69**±0.00	5.94 ± 0.13
	ALS	-	-	-	-
	k-moyennes flou	0.57 ± 0.06	0.42 ± 0.03	0.48 ± 0.01	4.67 ± 0.67

Avec toutes les précautions d'interprétation qu'une telle étude requiert [4], on notera sans surprises que les approches recouvrantes (MOC et OKM) sont adaptées pour l'analyse typologique de données très recouvrantes telles que Yeast où elles permettent d'obtenir un très bon rappel sans dégradation significative de la précision. En revanche, lorsque les données sont peu recouvrantes (Iris et Scene), k-moyennes ou k-moyennes flou (en recherchant de manière précise un bon seuil d'affectation a posteriori) sont davantage pertinentes.

Publications associées au chapitre 1

[Ben N'cir et al., 2014] Ben N'cir, C., Cleuziou, G., and Essoussi, N., (2014). Overview of Partitional Clustering Methods. In *Partitional Clustering Algorithms*, E.

4. du fait de l'utilisation d'une seule mesure d'évaluation et surtout du principe même d'évaluation externe.

Celebi Editor, Springer International Publishing, pages 245-275.

[Cleuziou et al., 2004] Cleuziou, G., Martin, L., and Vrain, C., (2004). PoBOC : an Overlapping Clustering Algorithm. Application to Rule-Based Classification and Textual Data. In R. López de Mántaras and L. Saitta, IOS Press, editor, *Proceedings of the 16th ECAI Conference*, pages 440–444, Valencia, Spain.

[Cleuziou, 2006] Cleuziou, G., (2006). Classification avec recouvrement des classes : une extension des k-moyennes. In *13èmes rencontres de la Société Francophone de Classification*, pages 68–72, Paris.

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

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

[Cleuziou and Sublemontier, 2008] Cleuziou, G. and Sublemontier J.H., (2008). Étude comparative de deux approches de classification recouvrante : Moc vs. Okm. In *8èmes journées d'Extraction et de Gestion des Connaissances (EGC'2008)*, pages 667-678.

2 Variantes en classification recouvrante

Ce second chapitre présente un ensemble d'extensions pour l'approche Okm à la base du chapitre précédent. Les variantes et améliorations proposées visent à répondre encore et toujours aux problématiques pratiques rencontrées par les utilisateurs ; il s'agit entre autres de tenir compte de leur besoin de contrôler l'importance des recouvrements, de visualiser les classifications générées, d'augmenter la robustesse de la méthode, de la rendre compatible avec d'autres types de données (symboliques, conceptuelles) et avec d'autres mesures de proximités. Pour chacune de ces requêtes nous présentons de manière synthétique : l'approche envisagée (en terme de modèle et/ou d'algorithme), les problématiques liées au domaine de la classification recouvrante ainsi qu'une illustration des principaux résultats obtenus.

2.1 Régulation des recouvrements (R-Okm)

Depuis sa publication en 2008, plusieurs études indépendantes ont été consacrées à l'amélioration de l'approche Okm. Nous mentionnerons par exemple les travaux de Ben N'Cir qui s'est intéressé successivement à la classification à base de noyaux [Ben N'Cir and Essoussi, 2012] [5], à la prise en compte de clusters de densités variées [Ben N'Cir and Essoussi, 2013] et à la gestion des outliers [Ben N'Cir et al., 2014], toujours dans le cadre théorique du modèle Okm. Une autre problématique, propre à la classification recouvrante, a été abordée dans [Lu et al., 2012] et concerne la nécessité (bien que discutable dans un contexte non-supervisé) de contrôler, limiter ou du moins gérer, l'importance des recouvrements ; l'algorithme Sparse-OC (*Sparse Overlapping Clustering*) reprend alors le critère objectif de Okm en ajoutant, selon le contexte applicatif :

- une contrainte explicite sur le nombre de classes auxquelles un individu est autorisé à appartenir ($|\pi_i| \leq \delta_i$)
- une contrainte implicite prenant la forme d'un terme pénalisant globalement les solutions à fort recouvrement ($\lambda\|\Pi\|_1$).

Dans le même temps nous avons proposé deux nouvelles modélisations pour permettre à l'utilisateur de réguler l'importance des recouvrements [Ben N'Cir et al., 2014]. Le premier modèle, R_1-Okm, conditionne l'affectation multiple d'un individu x_i au bénéfice que celle-ci induit sur l'erreur locale :

$$J_{r_1okm}(\Pi, C) = \sum_{x_i \in X} |\pi_i|^\alpha \|x_i - \frac{\sum_k \pi_{i,k} c_k}{\sum_k \pi_{i,k}}\|^2. \tag{9}$$

Par exemple, avec ce modèle et pour $\alpha = 1$, un individu x_i appartiendra à deux classes plutôt qu'à une seule seulement si l'erreur locale induite est au moins deux fois plus petite.

Le second modèle de régulation, R_2-Okm, introduit un terme de pénalité visant à limiter l'affectation d'un individu à des classes dispersées :

$$J_{r_2okm}(\Pi, C) = \sum_{x_i \in X} \|x_i - \frac{\sum_k \pi_{i,k} c_k}{\sum_k \pi_{i,k}}\|^2 + \lambda \sum_{x_i \in X} \frac{\sum_k \pi_{i,k} \|x_i - c_k\|^2}{|\pi_i|}. \tag{10}$$

5. sujet sur lequel nous reviendrons en fin de chapitre.

Dans ce modèle, l'appartenance d'un individu à plusieurs classes n'est envisageable qu'en observant le rapport entre la diminution de l'erreur locale et l'augmentation de la dispersion des classes concernées (proportionnellement à λ).

Les deux modèles R_1-OKM et R_2-OKM présentent la caractéristique de couvrir à la fois le modèle originel OKM (pour $\alpha = 0$ et $\lambda = 0$ respectivement) mais aussi le modèle des k-moyennes (pour α et $\lambda \to +\infty$) avec la possibilité de rechercher une solution moins recouvrante (α et $\lambda > 0$) ou au contraire davantage recouvrante (α et $\lambda < 0$) par rapport à OKM. La Figure 3 illustre cette variation sur la taille des recouvrements ainsi que les caractères non-linéaire et linéaire des modèles R_1-OKM et R_2-OKM respectivement.

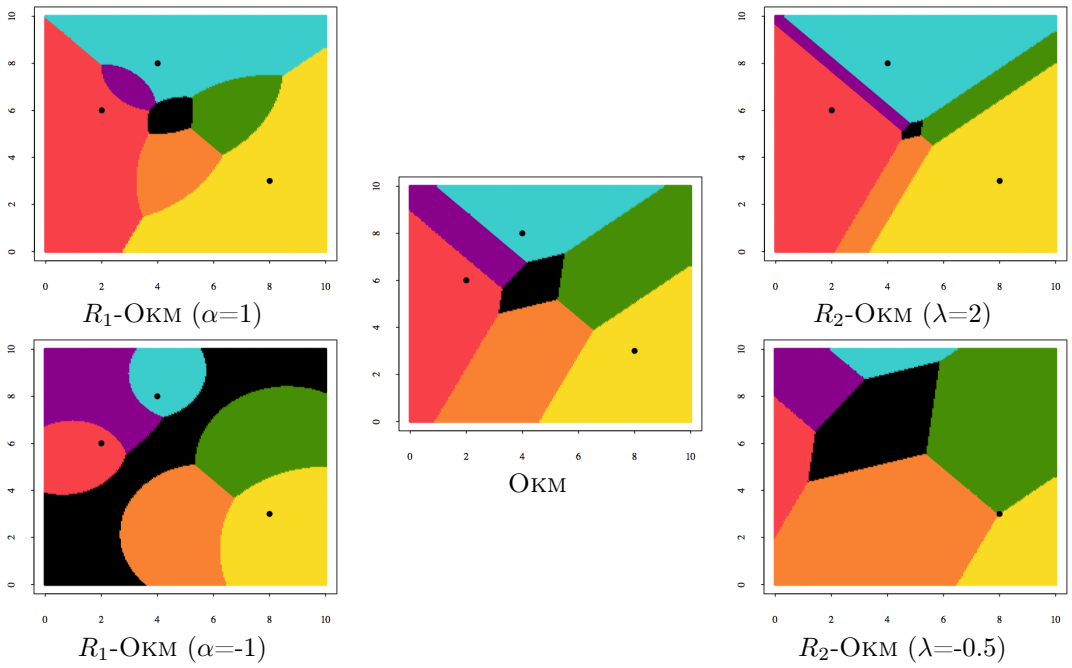


FIGURE 3 – Exemple en 2 dimensions des cellules de Voronoï obtenues pour un recouvrement en 3 classes à partir de trois profils (points noirs). Les cellules de couleurs primaires (rouge, jaune et bleu) identifient les zones où les objets ne seront affectés qu'à une seule classe, les mélanges de couleurs (orange, vert, violet) identifient les zones de recouvrement entre deux classes et la cellule en noir identifie un recouvrement sur les trois classes.

D'un point de vue algorithmique, ces deux nouveaux modèles trouvent une résolution semblable à celle utilisée pour OKM :

— Le principe d'affectation multiple reste inchangé (cf. heuristique définie dans l'Algorithme 1) mais considère le nouveau critère objectif dans la prise de décision.

— la mise à jour des profils de classes nécessite de recalculer les gradients des critères (9) et (10) relativement au profil à corriger ; les profils optimaux de chacun des modèles sont présentés dans la Table 2.

Nous donnons en Figure 4 un aperçu du comportement des modèles R_1-OKM et R_2-OKM, comparativement aux approches suivantes : k-moyennes, OKM, MOC, ALS et k-moyennes flou (avec recherche fine d'un seuil d'affectation a posteriori). En complément des trois jeux de données déjà utilisés précédemment (Iris, Scene et Yeast), nous

TABLE 2 – Synthèse des mises à jour des profils de classe pour les modèles OKM, R_1-OKM et R_2-OKM.

OKM	R_1-OKM	R_2-OKM
$c_k^* = \frac{\sum_{x_i \in \pi_k} \frac{1}{\lvert\pi_i\rvert^2} \hat{x}_i^k}{\sum_{x_i \in \pi_k} \frac{1}{\lvert\pi_i\rvert^2}}$	$c_k^* = \frac{\sum_{x_i \in \pi_k} \frac{1}{\lvert\pi_i\rvert^{2-\alpha}} \hat{x}_i^k}{\sum_{x_i \in \pi_k} \frac{1}{\lvert\pi_i\rvert^{2-\alpha}}}$	$c_k^* = \frac{\sum_{x_i \in \pi_k} \left(\frac{1}{\lvert\pi_i\rvert^2} \hat{x}_i^k + \frac{\lambda}{\lvert\pi_i\rvert} x_i \right)}{\sum_{x_i \in \pi_k} \left(\frac{1}{\lvert\pi_i\rvert^2} + \frac{\lambda}{\lvert\pi_i\rvert} \right)}$

introduisons un quatrième jeu de données *Emotion* issu du projet Mulan [6] et renfermant 593 extraits musicaux encodés selon 72 descripteurs numériques et pré-étiquetés suivant 6 classes d'émotions (en moyenne 1.86 étiquettes par extrait musical).

Ces expérimentations confirment d'abord le caractère générique des deux nouveaux modèles qui produisent effectivement la même partition que k-moyennes lorsque leur paramètre est choisi suffisamment grand, et atteignent le même recouvrement que par OKM lorsque leur paramètre s'annule. L'observation des recouvrements produits par les deux modèles R_1-OKM et R_2-OKM pour un paramètrage (et donc un taux de recouvrement) intermédiaire, révèle l'intérêt du second modèle qui seul permet d'atteindre des solutions faiblement recouvrantes meilleures qu'une partition lorsque les références sont peu ou pas recouvrantes (e.g. pour Iris et Scene).

Tout récemment, en 2015, Whang, Dhillon et Gleich se sont également intéressés à cette problématique en introduisant dans leur approche NEO-K-Means (*Non-Exhaustive Overlapping k-Means*) [Whang et al., 2015] la possibilité de produire une classification non-exhaustive, c'est-à-dire dans laquelle certains individus ne sont affectés à aucune classe. Leur proposition, en terme de contrôle sur les recouvrements, consiste à imposer une première contrainte sur le nombre d'affectations ($trace(\Pi^T\Pi) = \|\Pi\|_1 = (1+\alpha)N$) et une seconde sur le nombre maximal d'individus non-affectés ($\sum_{x_i \in X} \mathbb{1}_{|\pi_i|=0} \leq \beta N$). Ces contraintes sont notamment utilisées à bon escient pour guider le processus de classification sans qu'il soit nécessaire de recourir à une quelconque hypothèse (e.g. additive ou géométrique) sur la forme des recouvrements.

Sur les mêmes jeux de données (en particulier Scene, Emotion et Yeast) ainsi que d'autres données générées artificiellement, les auteurs de NEO-K-Means comparent, pour la première fois, et plutôt avantageusement NEO-K-Means, OKM et R-OKM (probablement R_2-OKM) à l'approche SPARSE-OC citée précédemment, ainsi qu'à MOC et k-moyennes flou. Une étude expérimentale plus étendue, à la manière de celle reportée en Figure 4, reste à réaliser afin de positionner les méthodes SPARSE-OC et NEO-K-Means relativement au taux de recouvrement produit.

6. http ://mulan.sourceforge.net/datasets-mlc.html

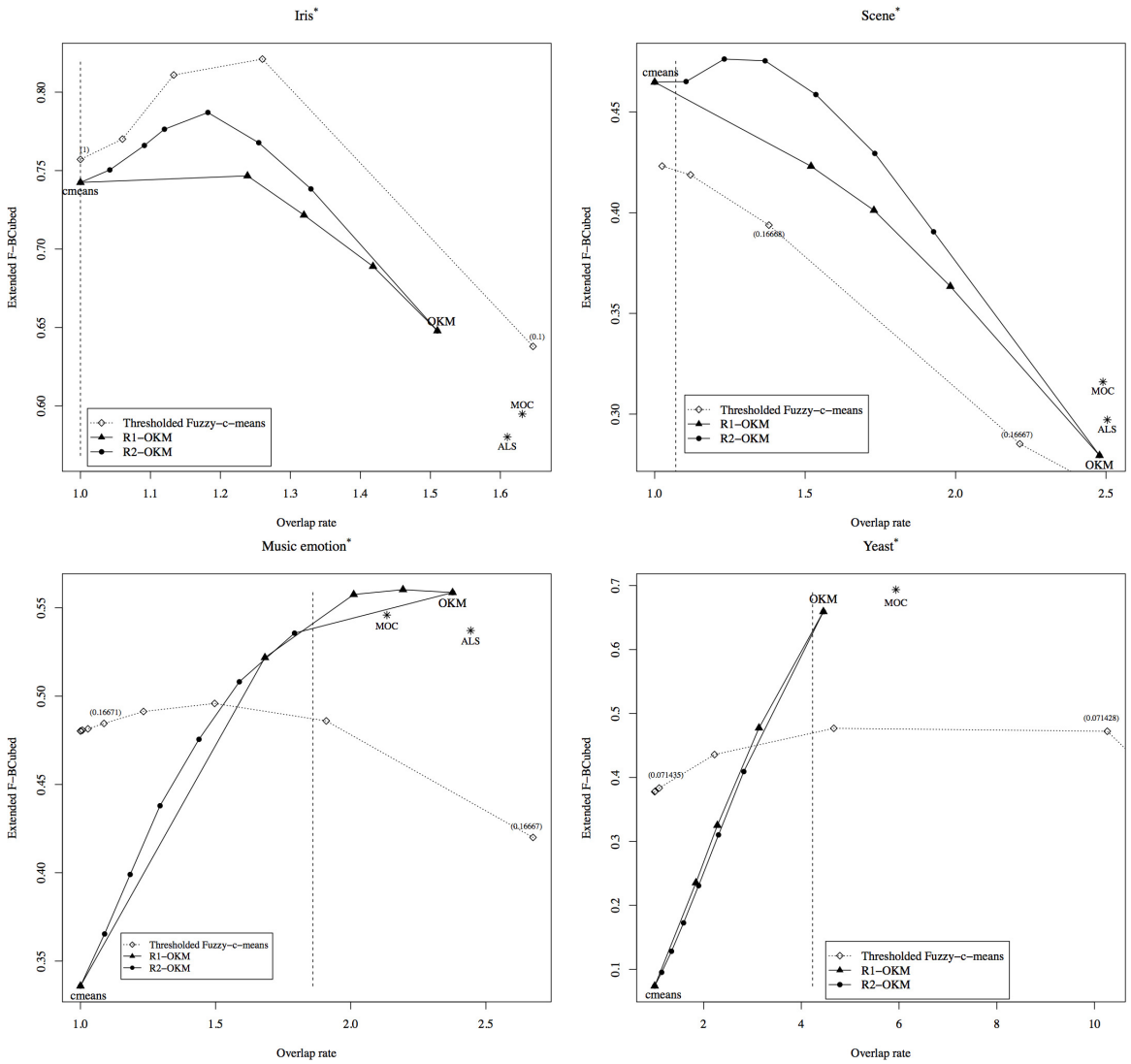


FIGURE 4 – Comportement des modèles R_1-OKM et R_2-OKM sur données réelles : Iris, Scene, Emotion et Yeast. La position de chaque méthode correspond au score *F-Bcubed* (étendu) obtenu en fonction du taux de recouvrement produit. La ligne verticale en pointillés indique le taux de recouvrement de la référence.

2.2 Cartes de Kohonen recouvrantes (OSOM)

Nous avons étudié le modèle OKM sous l'angle des cartes auto-organisatrices, couramment appelées cartes topologiques ou cartes de Kohonen, en référence à l'auteur de l'algorithme SOM (*Self-Organizing Map*) [Kohonen, 1984]. Cette étude a été réalisée dans le but de répondre à quatre objectifs majeurs en classification recouvrante :

1. proposer une **représentation visuelle** des classes recouvrantes générées,
2. simplifier le choix et réduire la **sensibilité au nombre de classes** (paramètre K),
3. améliorer la **cohérence topologique des recouvrements**,
4. générer une première segmentation adaptée à un **post-traitement pseudo-hiérarchique**.

L'un des avantages majeurs des approches hiérarchiques ou pseudo-hiérarchiques [7] est de proposer, en sortie de l'algorithme, une représentation visuelle fournissant à l'utilisateur une clé de lecture sur l'analyse réalisée. Le recours aux cartes auto-organisatrices permet de produire une sortie visuelle via une carte (ou grille), le plus souvent en deux dimensions, correspondant à une projection non-linéaire des individus. Par ailleurs, les grilles considérées possèdent généralement un grand nombre de cellules qu'il conviendra a posteriori de segmenter pour en dégager les classes finales. Cela permet donc de s'affranchir du choix de K et de reporter la décision au moment de la segmentation. Enfin, rendre visuellement interprétable une classification impose de satisfaire certaines contraintes, telles que l'absence de croisements ou d'inversions dans une pyramide ; dans le cas d'une carte topologique recouvrante, il s'agira d'interdire les recouvrements entre des classes correspondant à des cellules non connectées sur la grille. Cette contrainte de "cohérence topologique" permet naturellement d'éviter l'affectation d'un individu à des classes éloignées, rappelant l'intuition du modèle R_2-Okm d'une part, et confirmant l'heuristique d'affectation (aux plus proches classes) proposée dans Okm d'autre part.

Nous sommes partis du modèle théorique proposé par Heskes pour les cartes topologiques non-recouvrantes [Heskes, 1999] qui introduit une fonction d'énergie (ou critère objectif) permettant de guider le processus de construction de la carte topologique. Nous présentons cette fonction en (11) en l'adaptant aux notations introduites jusqu'ici :

$$J_{Heskes}(\Pi, C) = \sum_{x_i \in X} \sum_{k=1}^{K} h_{\pi_k, \pi(x_i)} \|x_i - c_k\|^2. \tag{11}$$

Dans cette expression, K correspond au nombre de cellules dans la grille considérée, $\pi(.)$ désigne une fonction renvoyant la classe (ici une cellule) d'appartenance de l'individu x_i et h_{π_k,π_l} formalise la proximité [8], sur la grille, entre deux cellules π_k et π_l. Une carte de bonne qualité sera caractérisée par une faible valeur du critère objectif, atteinte lorsque chaque individu x_i est :

- proche du profil de sa classe ($h_{\pi(x_i),\pi(x_i)} = 1$),
- plutôt proche des profils des classes voisines sur la carte ($h_{\pi_k,\pi(x_i)}$ proches de 1),
- éventuellement éloigné des profils des classes distantes sur la carte ($h_{\pi_k,\pi(x_i)}$ proches de 0).

Afin de générer des cartes recouvrantes, nous avons considéré la possibilité pour chaque individu x_i d'être affecté, non plus à une seule cellule, mais à un sous-ensemble de cellules G_i satisfaisant la contrainte topologique de former une clique sur la grille. De façon plus formelle, on notera $\mathcal{P}(\Pi)$ l'ensemble des 2^K combinaisons possibles de cellules et $\mathcal{C}(\Pi) \subset \mathcal{P}(\Pi)$ les sous-ensembles de cellules formant une clique. Ainsi, $G_i \in \mathcal{C}(\Pi)$ forme un ensemble (de cellules) convexe sur la grille. Nous avons défini le critère objectif suivant pour guider le processus de génération de cartes topologiques recouvrantes :

$$J_{OSOM}(\Pi, C) = \sum_{x_i \in X} \sum_{G \in \mathcal{C}(\Pi)} h_{G,G_i} \|x_i - \frac{\sum_{\pi_k \in G} c_k}{|G|}\|^2. \tag{12}$$

Le fait de raisonner sur le powerset plutôt que sur l'ensemble de cellules nécessite de redéfinir la fonction de proximité $h : \mathcal{P}(\Pi) \times \mathcal{P}(\Pi) \to [0, 1]$. Nous avons alors proposé

7. Par exemple les pyramides [Diday, 1987].
8. Généralement h est choisie dans $[0, 1]$ telle que $h_{\pi_k,\pi_k} = 1$.

d'utiliser une fonction de voisinage de type Gaussien, relative à la distance de Hausdorff [Hausdorff, 1962] mesurant une distance entre deux ensembles de cellules. Plus formellement,

$$h^t_{G,G'} = \exp\left(-\frac{d_H(G,G')}{2\sigma_t^2}\right) \text{ où } d_H(G,G') = max\left\{\begin{array}{l}\max\limits_{\pi_k\in G}\min\limits_{\pi_l\in G'} d(\pi_k,\pi_l)\\ \\ \max\limits_{\pi_l\in G'}\min\limits_{\pi_k\in G} d(\pi_k,\pi_l)\end{array}\right\} \quad (13)$$

où $d(.)$ renvoie la distance, sur la grille, entre deux cellules et t paramètre la zone d'influence topologique de façon à permettre de la concentrer sur les cellules (ou sous-ensembles de cellules) les plus proches à mesure des itérations de l'algorithme.

L'adaptation du modèle de Heskes aux cartes topologiques recouvrantes permet de dériver l'algorithme OSOM de construction adaptatif (*on line*) dont le principe général est le suivant : après une initialisation aléatoire des cellules de la grille, chaque itération (t) nécessite de

1. tirer aléatoirement un individu $x_i \in X$ et rechercher le meilleur sous-ensemble de cellules $G_i \in \mathcal{C}(\Pi)$ au sens du critère objectif (12), soit

$$G_i = \arg\min_{G'\in\mathcal{C}(\Pi)} \sum_{G\in\mathcal{C}(\Pi)} h^t_{G,G'} \|x_i - \frac{\sum_{\pi_k\in G} c_k}{|G|}\|^2 \quad (14)$$

2. mettre à jour tous les profils de classes $c_1 \dots c_K$:

$$c_k \leftarrow c_k + \epsilon_t \sum_{G|\pi_k\in G} \frac{h^t_{G,G_i}}{|G|}(x_i - \frac{\sum_{\pi_l\in G} c_l}{|G|}) \quad (15)$$

Cet algorithme souffre d'une complexité directement liée à la taille, a priori exponentielle, du sous-ensemble $\mathcal{C}(\Pi) \subset \mathcal{P}(\Pi)$ considéré. On observera alors qu'en pratique, si on considère la classe des cliques pour $\mathcal{C}(\Pi)$ et une structure de grille 2D avec un voisinage carré (8 voisins par cellules), la taille de $\mathcal{C}(\Pi)$ reste linéaire relativement à K ($|\mathcal{C}(\Pi)|$<10.K). D'autre part, nous avons proposé dans [Cleuziou, 2013] une version heuristique (*fast*-OSOM) dans laquelle la recherche du meilleur sous-ensemble de cellules G_i passe par l'identification d'une cellule c_i, pour être ensuite élargie à un ensemble G_i.

En revanche, d'un point de vue théorique, l'adaptation du modèle de Heskes nous permet de décomposer le critère objectif (12) de la manière suivante :

$$J_{OSOM}(\Pi, C) = \sum_{x_i\in X} \|x_i - \frac{\sum_{\pi_k\in G_i} c_k}{|G_i|}\|^2 + \sum_{x_i\in X}\sum_{G\neq G_i} h_{G,G_i}\|x_i - \frac{\sum_{\pi_k\in G} c_k}{|G|}\|^2 \quad (16)$$

avec un premier terme qui reprend le critère objectif de OKM, et un second terme permettant de contrôler la cohérence topologique de la carte. De même que le modèle de Heskes peut-être vu comme une extension de k-moyennes aux cartes topologiques, le modèle que nous avons proposé doit ainsi être interprété comme une extension de OKM aux cartes topologiques recouvrantes.

Nous illustrons en Figure 5 une carte topologique recouvrante obtenue par l'algorithme *fast*-OSOM sur le jeu de données *Emotion* précédemment introduit. Cette représentation visuelle particulière permet de compléter l'interprétation d'une carte topologique classique par une information sur la connexité des cellules entre elles, offrant par la même occasion un post-traitement de la carte facilité (e.g. segmentation automatique).

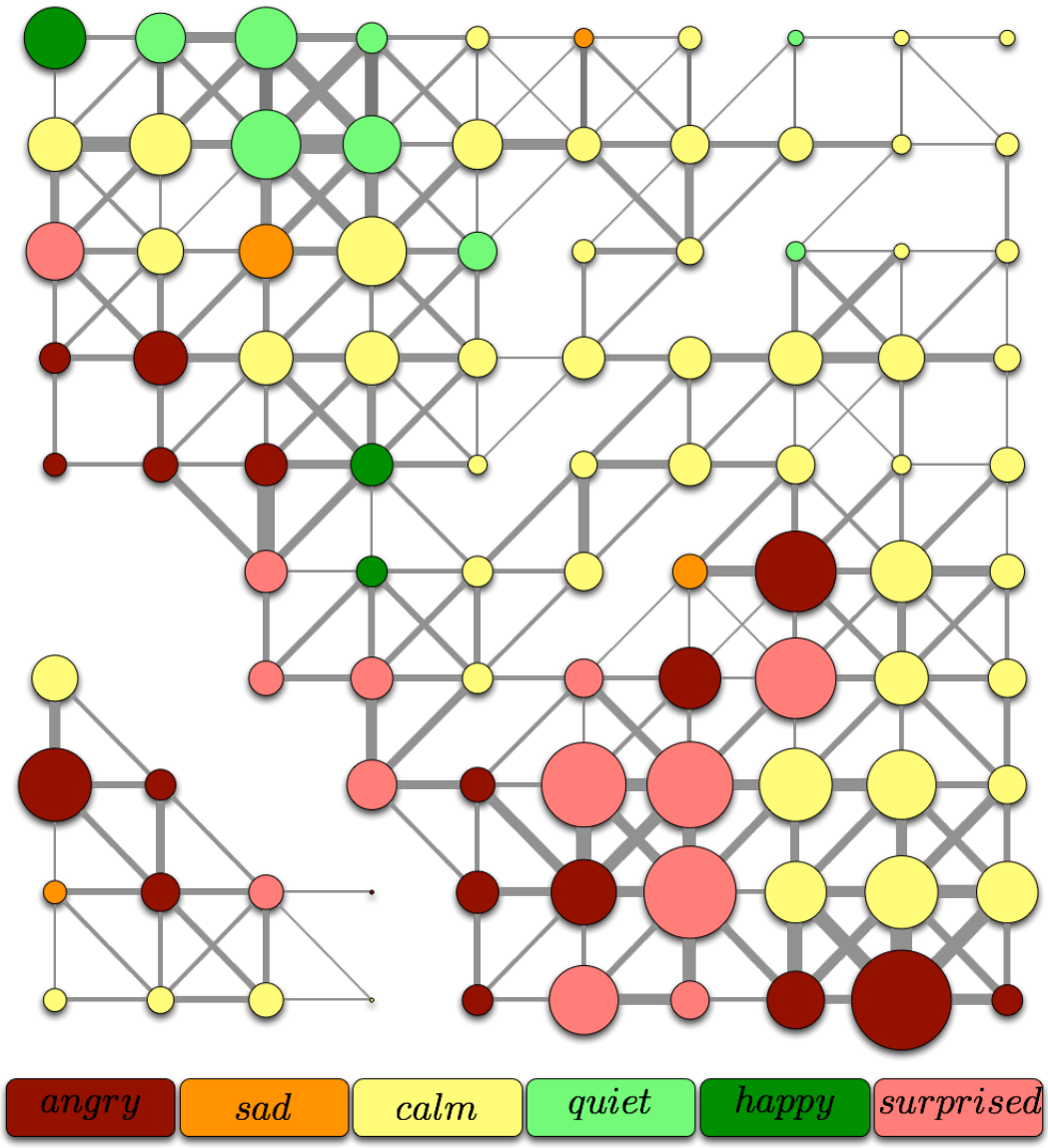


FIGURE 5 – Carte topologique recouvrante obtenue par l'algorithme *fast*-OSOM sur *Emotion* : la taille des cellules est proportionnelle au nombre d'individus qu'elles renferment ; la couleur indique la classe la plus représentative dans la cellule et les connexions entre cellules modélisent l'existence d'un recouvrement, dont l'importance est proportionnelle à l'épaisseur du lien.

2.3 Variantes robustes (OKM-L_1, T-OKM)

La robustesse des méthodologies de classification, en particulier non-supervisées, est une qualité essentielle permettant de limiter l'impact lié à la présence de bruit et/ou de données atypiques (*outliers*) sur la classification finale. Il est connu que l'algorithme des k-moyennes est peu robuste puisque la présence d'un seul outlier est susceptible de perturber totalement une classification en isolant l'outlier dans l'une des classes [García-Escudero and Gordaliza, 1999]. Nous illustrons par un exemple le fait que les méthodes de classification recouvrante sont encore davantage sensibles au bruit ou à la présence d'outliers.

Nous avons construit un jeu de données constitué de 24 points (univariés) correspondant aux performances des athlètes en finale du triple saut aux JO de Londres en 2012 : 12 hommes et 12 femmes. La répartition des points (illustrée en Fig. 6) est telle que k-moyennes (avec k=2) identifie systématiquement les deux classes d'athlètes de sexe opposé. L'application de OKM identifie également ces deux classes avec cependant un recouvrement sur les deux individus centraux, tel que l'athlète femme arrivée en tête et l'athlète homme arrivé 12ème sont tous les deux affectés aux deux clusters.

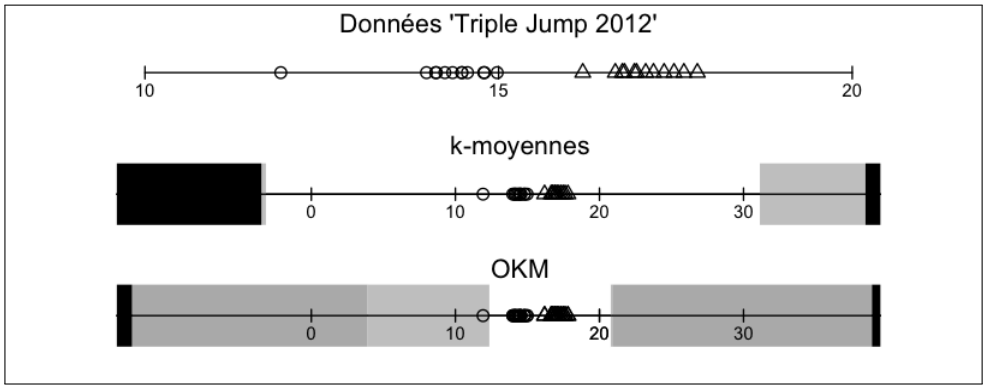


FIGURE 6 – Illustration de la robustesse de k-moyennes et de OKM sur les données "triple jump 2012" où les positions des △ et des ∘ désignent respectivement les performances des hommes et des femmes (en mètres). L'ajout d'une seule donnée dans une zone grise (resp. noire) induira une rupture partielle (resp. totale) de la classification.

En repartant des fonctions objectives utilisées dans k-moyennes et OKM, nous avons calculé analytiquement les points de "rupture partielle", c'est-à-dire les positions à partir desquelles l'ajout d'un individu entraînera une modification des classifications. Nous avons également observé, de manière empirique cette fois, les points de "rupture totale" à partir desquels l'individu supplémentaire sera isolé dans une classe (outlier), tandis que les 24 athlètes initiaux seront regroupés dans l'autre classe. Ces points de ruptures sont illustrés en Fig. 6.

On observe par exemple pour k-moyennes qu'il faudrait un nouveau saut comptabilisé à 31.2m pour altérer (partiellement) la classification et qu'à partir de 38.5m la rupture serait totale ; en revanche l'ajout d'une seule valeur nulle (e.g. un athlète qui se blesse) n'aurait pas de conséquence [9]. La sensibilité de OKM est nettement plus grande puisqu'on observe qu'un saut supplémentaire enregistré à moins de 12.5m perturbera (à la marge)

9. Nous avons cependant observé que deux valeurs nulles induirait une rupture totale.

la classification et qu'en dehors de l'intervalle $[3.9, 21]$ la rupture sera majeure (zone en gris foncé), avec l'organisation suivante : classe 1 = {1 type d'athlète + outlier} et classe 2 = {2 types d'athlètes} ; l'isolement complet de l'outlier étant quant à lui repoussé à l'extérieur de l'intervalle $[-12.5, 38.9]$. Cette sensibilité accrue est liée au fait que les méthodes recouvrantes (telles que OKM) utilisent non seulement les profils des clusters mais également des combinaisons de ces profils pour modéliser les zones de recouvrement. Ainsi l'ajout d'un outlier influe non plus seulement sur le profil de la plus proche classe mais sur l'ensemble des profils des classes qui partagent des individus avec cette dernière ; dans notre exemple précédent, les deux classes étant recouvrantes à la base, tout ajout de point aura une incidence sur la position des deux profils de classe.

Différentes stratégies peuvent être envisagées pour traiter le problème de l'hypersensibilité des approches de classification recouvrantes ; parmi celles-ci on trouve :

— les approches à base de **médoïdes** qui contraignent les profils à rester dans l'enveloppe convexe des clusters (parmi les données initiales),
— l'adaptation à des métriques réputées plus robustes telle que la **norme** L_1,
— des solutions algorithmiques comme le ***trimming*** consistant à élaguer une partie des données lors du processus d'apprentissage du modèle.

Nous avons proposé dans [Cleuziou, 2009a, Cleuziou, 2009b] une version de l'algorithme OKM considérant des médoïdes en guise de profils de clusters. Cette variante qui s'est avérée efficace en terme de robustesse reste néanmoins coûteuse puisqu'elle nécessite, lors du processus d'affectation, de rechercher parmi les données dans X, le meilleur représentant du recouvrement de clusters étudié.

Nous présentons ici deux adaptations de OKM correspondant à la mise en œuvre des deux dernières stratégies évoquées : OKM-L_1 pour l'adaptation à la norme L_1 ([Cleuziou et al., 2012]) et T-OKM pour la variante avec trimming ([Cleuziou and de Carvalho, 2014]).

Adaptation à la norme L_1. Le passage de la norme L_2 à la norme L_1 dans la méthode OKM ne remet pas en cause la forme du critère objectif (somme d'erreurs) et nous choisissons de conserver également la définition de la combinaison des représentants [10] :

$$J_{OKM-L_1}(\Pi, C) = \sum_{x_i \in X} \|x_i - \frac{\sum_k \pi_{i,k} c_k}{\sum_k \pi_{i,k}}\|_1 = \sum_{x_i \in X} \sum_{v=1}^{M} |x_{i,v} - \frac{\sum_k \pi_{i,k} c_{k,v}}{\sum_k \pi_{i,k}}| \qquad (17)$$

L'heuristique d'affectation dans OKM-L_1 restera identique en considérant pour chaque individu x_i les classes de la plus proche à la plus éloignée (cette fois au sens de $\|x_i - c_k\|_1$) et à l'affecter tant que l'erreur diminue. En revanche la mise à jour des profils de classes ne peut plus être obtenue par résolution d'un problème d'optimisation convexe et nécessite donc d'être redéfinie. Chaque mise à jour d'un représentant c_l consiste à minimiser indépendamment pour chaque variable $v = 1 \dots p$ l'expression

$$\sum_{x_i \in \pi_l} |x_{i,v} - \frac{\sum_k \pi_{i,k} c_{k,v}}{\sum_k \pi_{i,k}}| = \sum_{x_i \in \pi_l} |(x_{i,v} - \frac{\sum_{k \neq l} \pi_{i,k} c_{k,v}}{\sum_k \pi_{i,k}}) - \frac{c_{l,v}}{\sum_k \pi_{i,k}}|$$

10. Il est possible de modifier la fonction de combinaison en considérant un point médian plutôt que l'isobarycentre des profils de classes.

Il s'agit alors d'un problème d'optimisation en norme L_1 de la forme $\sum_{i=1}^{n} |y_i - az_i|$ (où a est la variable du problème) et dont une solution peut être obtenue par la méthode proposée par [Karst, 1958] :

1. Construire les variables $b_i = y_i/z_i$ $(i = 1 \dots n)$
2. Ordonner les z_i $(\tilde{z}_1 \dots \tilde{z}_n)$ selon l'ordre croissant des b_i $(\tilde{b}_1 \leq \dots \leq \tilde{b}_n)$
3. Déterminer le plus petit indice r tel que $\sum_{i=1}^{r} |\tilde{z}_i| > \sum_{i=r+1}^{N} |\tilde{z}_i|$ et retenir $a = \tilde{b}_r$

Lorsque tous les z_i sont égaux on voit aisément que cette méthode détermine la valeur médiane de l'ensemble des b_i ; dans notre contexte, les z_i sont tous positifs et expriment une forme de pondération inversement proportionnelle au nombre de classes de x_i tandis que les termes b_i désignent exactement la position $c_{l,v}$ optimale du point de vue de l'individu x_i (i.e. telle que $|x_{i,v} - \frac{\sum_k \pi_{i,k} c_{k,v}}{\sum_k \pi_{i,k}}| = 0$). Ainsi la mise à jour de c_l correspond à une forme de médiane pondérée des profils souhaités par chaque individu dans π_l. C'est précisément ce recours à la notion de médiane plutôt qu'à une moyenne qui permet aux approches de classification en norme L_1 d'être moins sensibles à la présence de bruit ou d'individus atypiques dans les données.

Approche par trimming. Le *trimming*, que nous traduirions par "émondage" en français, consiste à enlever une partie bien choisie des données afin qu'elles n'interviennent pas dans la mise à jour des profils de classes [García-Escudero et al., 2010]. Étant donné un paramètre d'émondage α et un ensemble d'individus X de taille N, on peut montrer qu'en choisissant à chaque itération dans k-moyennes les $[N(1-\alpha)]$ individus les plus proches des profils de classes, l'algorithme assurera la décroissance du critère objectif émondé. L'adaptation de l'émondage à OKM nécessite de choisir les $[N(1-\alpha)]$ individus les plus proches d'une combinaison de profils. Nous présentons ci-après un pseudo-code pour l'algorithme T-OKM (*Trimmed*-OKM) :

1. *Initialisation* : choisir aléatoirement K individus de X comme profils des classes C^0.
2. *Affectation* : structurer X en K classes recouvrantes par l'heuristique d'affectation de OKM à partir des profils C^t.
3. *Étape de concentration* :
   (a) Construire $H \subset X$ constitué des $[N(1-\alpha)]$ données les plus proches de l'une des (2^K-1) combinaisons de profils [11].
   (b) Mettre à jour successivement les profils de classes $c_1^t \to c_K^t$ par l'expression :

$$c_l^{t+1} = \frac{\sum\limits_{x_i \in \pi_l \cap H} \frac{1}{|\pi_i|^2} \left(|\pi_i| . x_i - \sum\limits_{k=1}^{l-1} \pi_{i,k} c_k^{t+1} - \sum\limits_{k=l+1}^{K} \pi_{i,k} c_k^t \right)}{\sum\limits_{x_i \in \pi_l \cap H} \frac{1}{|\pi_i|^2}}. \qquad (18)$$

4. Répéter les étapes 2 et 3 jusqu'à convergence du critère objectif émondé suivant :

$$J_{T-OKM}(\Pi, H, C) = \sum_{x_i \in H} \| x_i - \frac{\sum_k \pi_{i,k} c_k}{\sum_k \pi_{i,k}} \|^2. \qquad (19)$$

11. Notons qu'il n'est pas nécessaire de parcourir l'ensemble des combinaisons, mais qu'il suffit d'ordonner les individus selon leur erreur, calculée lors de l'étape précédente d'affectation.

Afin d'illustrer empiriquement l'impact de chacune des trois stratégies pour gérer la robustesse, nous proposons une expérience similaire à celle proposée récemment par [D'Urso et al., 2014]. Nous générons de manière artificielle un jeu de données en dimension 2, composé de deux classes de 100 individus chacune uniformément répartis dans deux carrés (ayant des côtés de longueur 1) centrés respectivement en (-2,-2) et (2,2). Des données bruitées ou atypiques sont ensuite générées en quantité variable (de 0 à 30% de données supplémentaires) suivant une gaussienne de paramètres $\mathcal{N}((5,5),5)$. Nous observons l'influence des données supplémentaires sur les classifications obtenues par Okm et ses différentes variantes en mesurant l'écartement des profils de classes par rapport à leur position initiale (mesure *rd()* pour *Robustness Detection* proposée par [D'Urso et al., 2014]).

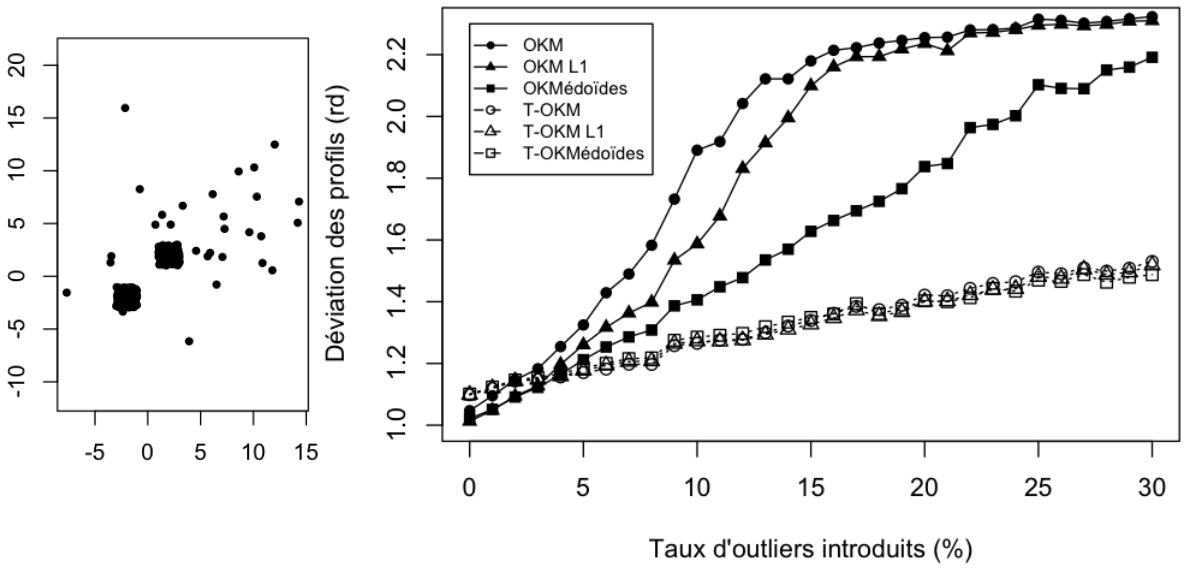


Figure 7 – Évaluation de la robustesse de Okm, Okm en norme L_1, Okm avec médoïdes et de leur version émondée (α=0.5) : visualisation d'une génération d'exemples (figure de gauche) et évaluation quantitative par la mesure *rd* (figure de droite).

Chaque expérience de génération de données a été répétée 100 fois et nous reportons en Fig.7 les moyennes obtenues pour la mesure *rd()* qui vaut 1 lorsque les profils des classes sont conformes aux attentes et augmente à mesure que les profils s'écartent de leur position initiale.

Les observations confirment les attentes concernant les versions non-émondées de Okm, à savoir que l'utilisation de la norme L_1 améliore la robustesse de Okm, tandis qu'une variante à base de médoïdes est significativement meilleure encore de ce point de vue. Enfin et surtout, l'intérêt de l'émondage est notable pour cette expérience puisque les trois versions de Okm avec un taux d'émondage de 50% [12] résistent très bien à l'ajout de données bruitées ou atypiques, sans qu'il n'y ait d'ailleurs de différences significatives entres elles.

12. Taux généralement utilisé pour les méthodes de *Trimming* dans la littérature.

2.4 Vers l'Analyse de Données Symboliques (I-OKM)

L'Analyse de Données Symboliques (ADS), apparue vers la fin des années 80 [Diday, 1989], est une extension de l'analyse de données classiques (de type attributs/valeurs) à des données plus complexes décrites par des variables symboliques telles que des intervalles, des ensembles, des histogrammes ou des distributions [Billard and Diday, 2007]. Plus qu'une simple extension, l'ADS peut être vue comme une manière de traiter des concepts de plus haut niveau en considérant des agrégations d'individus décrits de manière standard (attribut/valeur) ; typiquement l'ADS permet d'analyser et de comparer des espèces animales à partir de la description des animaux qui les composent, des villes à partir de leurs habitants, des hôpitaux connaissant leurs patients, des internautes étant données leurs traces numériques, etc.

Le traitement de ce type complexe de données a nécessité d'étendre les techniques usuelles en statistique exploratoire. Ainsi l'ADS couvre aujourd'hui des techniques de classification supervisée (e.g. arbres de décisions), de clustering (e.g. partitions ou hiérarchies), d'analyse factorielle, d'analyse discriminantes ou encore de visualisation. Pour la tâche qui nous concerne, le clustering, l'adaptation aux données symboliques nécessite généralement de définir une métrique appropriée et d'adapter le modèle en conséquence ; par exemple H.H. Bock propose dans [Bock, 2003] différentes manières de définir une distance entre deux intervalles multidimensionnels (hyper-rectangles) :

— la distance Euclidienne entre leur milieu,
— la somme des distances Euclidiennes entre leurs sommets,
— la distance de Hausdorff entre les deux hyper-rectangles

que nous illustrons graphiquement dans la Figure 8 pour des intervalles bi-dimensionnels.

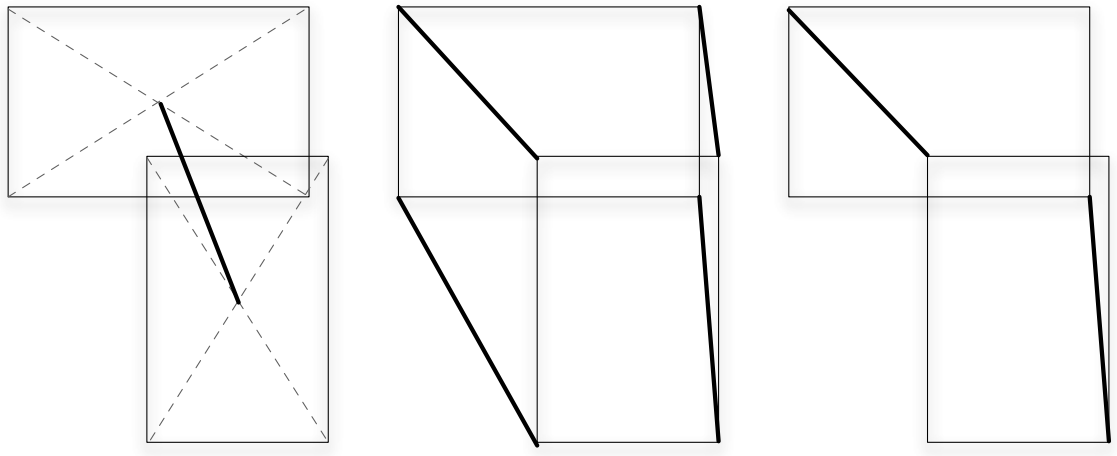


FIGURE 8 – Illustrations des éléments (segments) sur lesquels reposent le calcul de distance entre données intervalles : distance entre les milieux (à gauche), somme des distances entre les sommets (au centre) ou maximum des distances ensemblistes de Hausdorff (à droite).

L'utilisation de la méthode des milieux projette chaque intervalle en un simple point. Ce choix permet de recourir aisément à toutes les techniques d'analyse exploratoire mais reste insatisfaisant pour comparer la forme et le volume des intervalles considérés. Les

deux autres méthodes ont donné lieu à des adaptations du clustering dynamique dans lesquels les prototypes de classes sont également des intervalles.

Distance des sommets. Par exemple, [Bock, 2003] montre que le critère des moindres carrés utilisé dans l'algorithme k-moyennes se réécrit selon les bornes des intervalles lorsque l'on cherche à minimiser la somme des inerties des intervalles autour de leur intervalle prototypique via la méthode des sommets.

Soient X un ensemble de N hyper-rectangles dans $\mathbb{R}^M$ chacun décrit par un vecteur d'intervalles tel que $x_i = ([a_{i,1}, b_{i,1}], \ldots, [a_{i,M}, b_{i,M}])$ et C un ensemble de K prototypes définis dans le même espace ($c_k = ([\alpha_{k,1}, \beta_{k,1}], \ldots, [\alpha_{k,M}, \beta_{k,M}])$), l'algorithme des k-moyennes peut être adapté à ce type de données intervalles en considérant le critère objectif suivant :

$$J^{L_2}_{I-kmeans}(\Pi, C) = \sum_{k=1}^{K} \sum_{x_i \in \pi_k} \|a_i - \alpha_k\|^2 + \|b_i - \beta_k\|^2. \tag{20}$$

Le processus d'optimisation consiste alors classiquement à itérer les étapes :

1. d'affection de chaque observation au prototype le plus proche :

$$\pi(x_i) = \arg\min_{k=1\ldots K} \|a_i - \alpha_k\|^2 + \|b_i - \beta_k\|^2$$

2. puis de mise à jour optimale des prototypes, obtenue en calculant l'isobarycentre des bornes des intervalles de la classe :

$$\alpha^*_k = \frac{\sum_{x_i \in \pi_k} a_i}{|\pi_k|} \text{ et } \beta^*_k = \frac{\sum_{x_i \in \pi_k} b_i}{|\pi_k|}$$

Distance de Hausdorff (type L_1). De même, [Chavent et al., 2003] montrent astucieusement qu'en utilisant une distance de Hausdorff de type L_1 pour comparer les intervalles ($d_H(x_i, x_j) = \sum_{v=1}^{M} \max(|a_{i,v} - a_{j,v}|, |b_{i,v} - b_{j,v}|)$), le critère objectif

$$J^{Hausdorff-L_1}_{I-kmeans}(\Pi, C) = \sum_{k=1}^{K} \sum_{x_i \in \pi_k} d_H(x_i, c_k) \tag{21}$$

peut se réécrire selon les milieux et demi-longueurs des intervalles, de la manière suivante :

$$J^{Hausdorff-L_1}_{I-kmeans}(\Pi, C) = \sum_{v=1}^{M} \sum_{k=1}^{K} \sum_{x_i \in \pi_k} |m_{i,v} - \mu_{k,v}| + |l_{i,v} - \lambda_{k,v}| \tag{22}$$

où m_i désigne le milieu de l'intervalle x_i ($m_{i,v} = \frac{a_{i,v}+b_{i,v}}{2}$), μ_k le milieu de l'intervalle prototype c_k ($\mu_{k,v} = \frac{\alpha_{k,v}+\beta_{k,v}}{2}$), l_i la demi-longueur de l'intervalle x_i ($l_{i,v} = \frac{b_{i,v}-a_{i,v}}{2}$) et enfin λ_k la demi-longueur du prototype c_k ($\lambda_{k,v} = \frac{\beta_{k,v}-\alpha_{k,v}}{2}$).

Le problème d'optimisation lié à la mise à jour des prototypes se résout en prenant comme prototypes optimaux les intervalles définis par les médianes des milieux et des demi-longueurs des intervalles dans chaque classe.

On retient, de ce qui précède, que l'adaptation de l'algorithme k-moyennes à des données intervalles nécessite de choisir judicieusement la métrique pour l'espace des intervalles de manière à se ramener à une formulation d'un problème d'optimisation en norme L_2 ou L_1 déjà connu.

OKM **sur données intervalles.** Une façon naturelle d'étendre ces modèles à la classification recouvrante est de considérer des combinaisons de profils de classes dans l'espace des intervalles. Pour une métrique de type Euclidienne sur cet espace on posera alors le critère objectif suivant :

$$J^{L_2}_{I-okm}(\Pi, C) = \sum_{x_i \in X} \|a_i - \frac{\sum_k \pi_{i,k}\alpha_k}{\sum_k \pi_{i,k}}\|^2 + \|b_i - \frac{\sum_k \pi_{i,k}\beta_k}{\sum_k \pi_{i,k}}\|^2. \quad (23)$$

de telle sorte qu'une observation (intervalle) pourra appartenir à plusieurs classes si sa description est proche de l'intervalle défini par les bornes inférieures et supérieures moyennes des profils (intervalles) de ces classes, comme illustré en Figure 9.

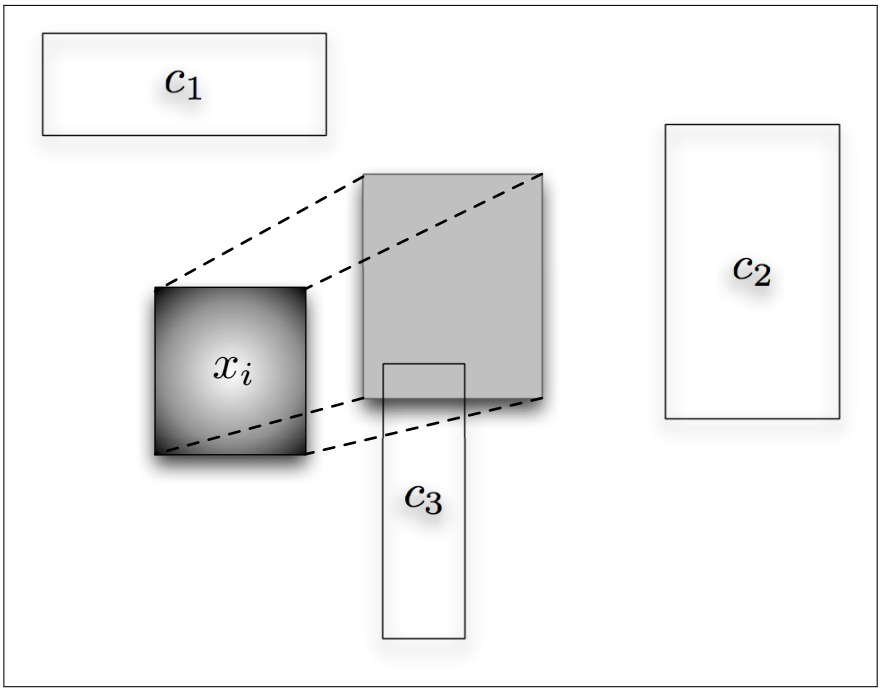


FIGURE 9 – Illustration de l'erreur induite par une observation (intervalle) x_i affectée à trois classes de profils (intervalles) respectifs c_1, c_2 et c_3. Les distances Euclidiennes sur les sommets de x_i et ceux de l'isobarycentre des trois profils (intervalle en gris) sont matérialisées par les segments en pointillés.

Le processus de clustering se traduirait alors par la minimisation du critère (23) via l'itération des deux étapes traditionnelles :

1. d'affection de chaque observation aux profils les plus proches en utilisant par exemple une heuristique d'affectation similaire à celle proposée dans OKM (cf. Algorithme 1),
2. puis de mise à jour successive des profils tels que α^*_k correspond au barycentre

du nuage des déviations [13] des bornes inférieures $\{(\hat{a}_i^l, \frac{1}{|\pi_i|^2})\}$ et β_k^* à celui des déviations des bornes supérieures $\{(\hat{b}_i^l, \frac{1}{|\pi_i|^2})\}$.

Sans précautions particulières, la mise à jour des profils peut cependant conduire à des incohérences.

Cohérence des profils intervalles. En effet il s'agit de veiller à ce que les profils de classes correspondent à des intervalles cohérents, c'est-à-dire des intervalles de longueurs positives ($\forall k, v,\ \alpha_{k,v} \leq \beta_{k,v}$) ce qui ne serait pas garanti par le processus d'optimisation décrit précédemment, contrairement aux modèles initiaux non-recouvrants.

Prenons l'exemple univarié de trois intervalles $x_1 = [0, 1]$, $x_2 = [2, 3]$ et $x_3 = [4, 12]$ à organiser en deux classes éventuellement recouvrantes π_1 et π_2 dans l'espace des intervalles muni d'une métrique Euclidienne. Nous présentons en Figure 10 une représentation des trois observations, et du processus de clustering induit par l'extension de Okm aux données intervalles (I-Okm), tel que décrit précédemment. On observe qu'en choisissant

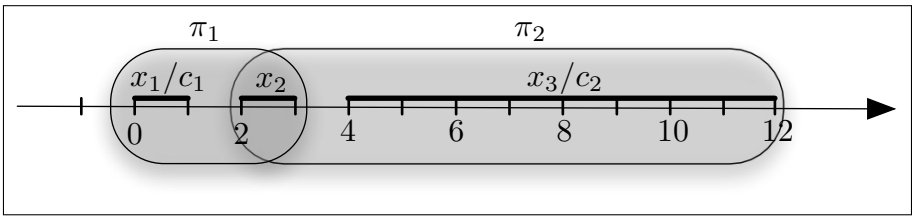


Figure 10 – Exemple d'une classification recouvrante sur données intervalles où la mise à jour du profil c_1 est susceptible de conduire à un profil intervalle incohérent ($\beta_1 < \alpha_1$).

initialement x_1 et x_3 comme profils initiaux (c_1 et c_2), la première étape d'affectation conduira à deux classes, recouvrantes sur x_2 ($\pi_1 = \{x_1, x_2\}$ et $\pi_2 = \{x_2, x_3\}$). La mise à jour du profil c_1 est réalisée indépendamment pour chaque borne :

- α_1^* est donné par le barycentre du nuage pondéré $\{(a_1, 1), (\hat{a}_2^1, \frac{1}{4})\} = \{(0, 1), (0, \frac{1}{4})\}$ d'où $\alpha_1^* = 0$,
- β_1^* est donné par le barycentre du nuage pondéré $\{(b_1, 1), (\hat{b}_2^1, \frac{1}{4})\} = \{(1, 1), (-6, \frac{1}{4})\}$ d'où $\beta_1^* = -0.4$

conduisant à la situation d'un profil intervalle dégénéré tel que la borne supérieure β_1 est plus petite que la borne inférieure α_1.

Notons que ce risque d'incohérences n'est pas spécifique à la modélisation géométrique des recouvrements telle que proposée dans Okm mais peut survenir du moment que les recouvrements sont modélisés par une combinaison quelconque de profils de classes (par exemple pour le modèle cumulatif utilisé dans ALS).

Résolution par approche sous-optimale. Afin de résoudre le problème d'incohérence présenté précédemment, nous proposons de modifier de manière générale l'étape de mise à jour des profils de classes par ajout d'une contrainte de non-inversion sur les profils intervalles générés. En observant que le critère objectif (23) peut se réécrire en fonction

13. Position d'une borne d'un profil de classe c_l permettant à un individu d'annuler l'erreur qu'il induit sur le système selon cette borne.

des milieux et demi-longueurs des intervalles :

$$J^{L_2}_{I-okm}(\Pi, C) = 2. \sum_{x_i \in X} \|m_i - \frac{\sum_k \pi_{i,k}\mu_k}{\sum_k \pi_{i,k}}\|^2 + \|l_i - \frac{\sum_k \pi_{i,k}\lambda_k}{\sum_k \pi_{i,k}}\|^2, \quad (24)$$

il en résulte que la mise à jour de c_k, défini par le couple (μ_k, λ_k), revient à considérer deux problèmes d'optimisation convexe indépendants :

1. la recherche du milieu μ_k optimal (non contraint) et dont la solution est donnée par le barycentre du nuage pondéré $\{(\hat{m}^k_i, \frac{1}{|\pi_i|^2})\}$,
2. le calcul de la demi-longueur λ^k optimale sous la contrainte d'être positive ou nulle et dont la solution est obtenue en considérant, pour chaque dimension v, soit le barycentre du nuage pondéré $\{(\hat{l}^k_{i,v}, \frac{1}{|\pi_i|^2})\}$ si celui-ci est positif, soit une demi-longueur nulle dans les autres cas.

Cette manière de procéder ne permet pas d'atteindre l'optimalité des mises à jour mais assure la décroissante du critère objectif et donc la convergence du processus de clustering tout en respectant la cohérence des profils intervalles, qui demeurent donc exploitables a posteriori.

OKM **pour la distance de Hausdorff sur les intervalles.** À la manière de [Chavent et al., 2003] on peut montrer que le critère objectif de OKM adapté à la distance de Hausdorff sur les données intervalles se réécrit comme une somme de deux termes en norme L_1 sur les milieux et demi-longueurs. La minimisation du critère ainsi modélisé nécessite en particulier la mise à jour des profils (intervalles) en considérant deux problèmes indépendants d'optimisation en norme L_1 déjà résolus en section 2.3. La contrainte de cohérence sur les intervalles nécessite alors pour la mise à jour des demi-longueurs des profils, une stratégie sous-optimale identique à celle proposée précédemment, cette fois sur le polygone convexe [Karst, 1958].

Notons que, dans ce qui précède nous avons considéré une modélisation simple des recouvrements, via l'isobarycentre des bornes des profils intervalles. D'autres combinaisons intéressantes, propres aux données intervalles, pourraient être explorées telles que le plus petit intervalle englobant (*join*) ou inclus (*meet*) susceptible d'une part de conduire à de nouvelles classifications recouvrantes pertinentes et d'autre part d'introduire un élagage des combinaisons de profils à explorer.

Par cette extension du modèle OKM à l'analyse de données symboliques nous avons montré à nouveau la généricité du modèle initial qui s'adapte plutôt bien à la plupart des contextes exploités en analyse de données mais qui nécessite néanmoins des précautions importantes pour assurer la cohérence du processus et des résultats qui en découlent.

2.5 Variante ensembliste à noyau (K-OKSETS)

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 projection que ceux qui ont fait leurs preuves en classification traditionnelle (non-recouvrante)

et d'envisager ainsi 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.

Clustering et noyaux. 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 $X = (x_1, \ldots, x_N)^T$, un ensemble de N observations décrites dans $\mathbb{R}^M$ et $\mathcal{K}$ une matrice noyau induite par une projection implicite $\phi(.)$ telle que $\mathcal{K}_{i,j} = \langle \phi(x_i), \phi(x_j) \rangle$. A priori, la méthode des k-moyennes ne peut pas être "kernélisée" directement puisqu'elle considèrerait la minimisation du critère d'inertie dans l'espace projeté :

$$J(\Pi, C) = \sum_{k=1}^{K} \sum_{x_i \in \pi_k} \|\phi(x_i) - \phi(c_k)\|^2 \tag{25}$$

et que ce critère s'appuie sur K centre mobiles (modélisant les profils des clusters) $\{c_k\}_{k=1}^K$ dont on ne souhaite pas calculer explicitement les projections $\phi(c_k)$. 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 :

$$J_{kkmeans}(\Pi) = \sum_{k=1}^{K} \sum_{x_i \in \pi_k} \mathcal{K}_{i,i} - \frac{2 \sum_{x_j \in \pi_k} \mathcal{K}_{i,j}}{|\pi_k|} + \frac{\sum_{x_j, x_l \in \pi_k} \mathcal{K}_{j,l}}{|\pi_k|^2}. \tag{26}$$

Cette astuce permet d'envisager un clustering totalement identique à k-moyennes dans n'importe quel espace induit par un noyau $\mathcal{K}$, cependant il est important d'observer que cette transformation a un coût non négligeable puisqu'elle induit un 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, on serait amené à considérer le critère objectif suivant

$$J_{OKM}(\Pi, C) = \sum_{x_i \in X} \|\phi(x_i) - \frac{\sum_k \pi_{i,k} \phi(c_k)}{\sum_k \pi_{i,k}}\|^2 \tag{27}$$

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).

Approche ensembliste. Nous avons proposé [Cleuziou, 2014] une modélisation ensembliste des recouvrements, qui se rapproche du modèle géométrique utilisé dans Okm mais permet un passage aux noyaux. Nous commençons par introduire la notion de *nuage* dans une classification.

Définition 2 *Étant donné un sous-ensemble de clusters $P \subseteq \Pi$, on appellera "nuage" de P l'application $N(.) : \mathcal{P}(\Pi) \rightarrow \mathcal{P}(X)$ qui lui associe l'union de ses extensions : $N(P) = \bigcup_{\pi_k \in P} \pi_k$.*

On parlera également de "nuage" associé à un individu $N(x_i)$ pour indiquer de façon analogue l'union des clusters auxquels il appartient : $N(x_i) = \bigcup_{\pi_k | x_i \in \pi_k} \pi_k$.

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

$$J_{OKSets}(\Pi) = \sum_{x_i \in X} \| x_i - \sum_{x_j \in N(x_i)} \frac{x_j}{|N(x_i)|} \|^2 \tag{28}$$

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 :

$$\begin{aligned} J_{KOKSets}(\Pi) &= \sum_{x_i \in X} \| \phi(x_i) - \sum_{x_j \in N(x_i)} \frac{\phi(x_j)}{|N(x_i)|} \|^2 \\ &= \sum_{x_i \in X} \mathcal{K}_{i,i} - \frac{2 \sum_{x_j \in N(x_i)} \mathcal{K}_{i,j}}{|N(x_i)|} + \frac{\sum_{x_j, x_l \in N(x_i)} \mathcal{K}_{j,l}}{|N(x_i)|^2} \end{aligned} \tag{29}$$

et on observe que dans le cas d'un clustering non-recouvrant, chaque individu appartient à un seul cluster $x_i \in \pi_k \Leftrightarrow N(x_i) = \pi_k$, ce qui nous ramène exactement au critère objectif de l'algorithme des k-moyennes à noyaux (26). Le modèle K-OKSETS défini précédemment est donc une généralisation recouvrante du modèle des k-moyennes à noyaux.

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 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 (29) et n'aurait donc aucune raison de converger vers un recouvrement et des profils de clusters stables. Nous avons proposé un algorithme adaptatif guidé par le critère $J_{KOKSets}$ (Figure 3).

L'étape cruciale de l'algorithme réside dans la procédure d'affectation d'un individu à un ou plusieurs clusters. Cette procédure doit être réalisée efficacement et assurer la décroissance du critère $J_{KOKSets}$. 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 (cf. Section 1.2) et d'autre part de stocker dans une structure

Algorithme 3 K-Okset

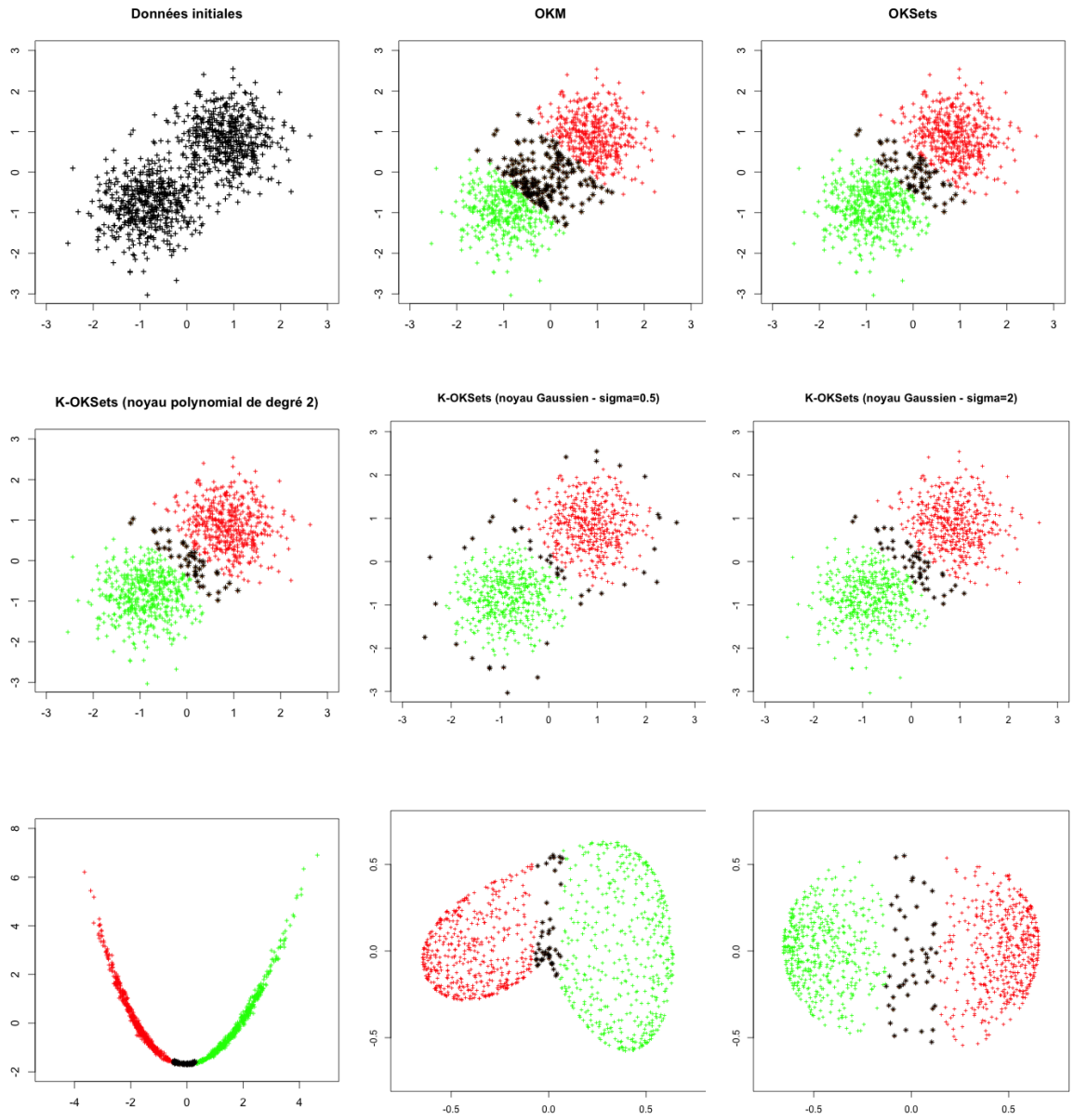
procedure K-Okset($X, \mathcal{K}, K, T$)
  $t \leftarrow 0$
  Initialisation aléatoire des clusters : $\pi_k^{(0)} \leftarrow Random(x_1, \ldots, x_N), \quad k = 1 \ldots K$
  for $x_i \notin \bigcup_k \pi_k^{(0)}$ **do** Affectation des individus restant
    $\pi_i^{(0)} \leftarrow$ Affectation-Kokset(x_i, $\mathcal{K}$, $\Pi^{(0)}$)
  end for
  Calculer $J_{KOKSets}^{(0)}$
  repeat
    $t \leftarrow t + 1$
    for $x_i \in X$ **do** Affectation des individus
      $\pi_i^{(t)} \leftarrow$ Affectation-Kokset(x_i, $\mathcal{K}$, $\Pi^{(t)}$)
    end for
    Calculer $J_{KOKSets}^{(t)}$
  until $J_{KOKSets}^{(t)} \geq J_{KOKSets}^{(t-1)}$ ou $t = T$ Conditions d'arrêt
  Return $\Pi^{(t)}$
end procedure

appropriée les nuages associés à chacune des combinaisons existantes de manière à éviter de les recalculer à chaque nouvelle affectation (voir [Cleuziou, 2014] pour plus de précisions sur cette structure).

Compte-tenu de l'augmentation exponentielle du nombre théorique de combinaisons 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 avons observé empiriquement les premiers gages de confirmation de cette hypothèse : par exemple pour 15 clusters sur la base Iris, au maximum 52 combinaisons seront considérées et stockées dans la structure, contre plus de 32 000 combinaisons théoriques envisageables.

Aperçu du comportement de KOKSets**.** Nous proposons une illustration du comportement de l'approche de classification recouvrante ensembliste OKSets et de son exploitation dans le cadre du clustering à noyau (KOKSets). Nous proposons en Figure 11 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) 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. En effet et comme nous l'avons déjà évoqué dans le chapitre précédent, dans les approches de classification recouvrante originelles (e.g. ALS, MOC ou Okm), la notion de "profils" de clusters ne correspond plus à l'idée intuitive de "centres" et l'optimisation du critère objectif de ces méthodes conduit à des profils pouvant être éloignés des données des clusters qu'ils

FIGURE 11 – Visualisations des recouvrements sur données artificielles : sans noyau (ligne supérieure), avec noyau (ligne du milieu) et dans l'espace de projection reconstruit (ligne inférieure).



représentent. Ce phénomène peut provoquer une réaction en chaîne : plus deux clusters se recouvriront, plus leurs profils seront éloignés de leur centre de gravité et plus ils risquent d'engendrer des recouvrements supplémentaires à l'itération suivante, etc. Ce phénomène est corrigé dans l'approche ensembliste OKSets du fait de la disparition des profils mobiles de clusters.

Les autres visualisations rendent compte des classifications obtenues par KOKSets avec différents noyaux : polynomial de degré 2 et gaussiens avec variances de 0.5 puis 2 ; nous visualisons les classifications et en particulier les recouvrements (points en noir) à la fois dans l'espace initial (ligne centrale) et dans l'espace de projection approximé par un positionnement multidimensionnel (MDS) à partir des distances euclidiennes induites par le noyau (ligne inférieure). On notera de façon générale la cohérence des classes et recouvrements générés et dans le cas particulier d'un noyau gaussien à faible variance la possibilité de détecter, via les recouvrements, le contour des deux distributions gaussiennes.

Bien que le choix du noyau reste un problème ouvert, en particulier dans le contexte de la classification non-supervisée, nous avons montré d'un point de vue qualitatif sur différents jeux de données recouvrants (multi-labels) que la variante ensembliste elle-même et son extension à noyau permettent de générer des classifications recouvrantes davantage conformes aux classifications recouvrantes par rapport à Okm (en particulier en ce qui concerne la taille des recouvrements) ou par rapport à d'autres approches non recouvrantes à noyau.

2.6 Vers la classification recouvrante conceptuelle

Dans une problématique assez différente mais toujours liée à la classification recouvrante, nous nous sommes intéressés à l'étude et la modélisation des recouvrements pour la tâche de classification conceptuelle dans le formalisme de l'analyse formelle de concepts (AFC).

Étant donné un ensemble d'objets G décrit par une relation binaire I sur un ensemble d'attributs M, la tâche de classification conceptuelle consiste à produire de façon simultanée un ensemble de clusters sur les objets et une description/définition intensionnelle de ces clusters. Plusieurs formes de clusterings conceptuels peuvent être recherchées : des structures hiérarchiques [Fisher, 1987] définissant un arbre de classification des données ou des structures "à plat" [Michalski and Stepp, 1983]. Dans ce deuxième cas, l'espace de recherche est généralement défini par l'ensemble des classifications dont les clusters sont des concepts formels (fermés) réalisant une couverture des données. À cet espace de recherche (trop large) est adjoint une ou plusieurs contraintes structurelles ou qualitatives permettant respectivement de restreindre l'espace de solutions ou d'orienter la recherche vers une solution de "bonne qualité". Par exemple, la contrainte structurelle de partitionnement (strict) étant généralement trop forte, elle est souvent délaissée au profit d'une contrainte de tolérance sur la taille des recouvrements entre concepts (e.g. chevauchement d'au plus n objets entre deux clusters/concepts).

Modélisation des classifications conceptuelles recouvrantes. Nous avons introduit dans [Cleuziou and Crémilleux, 2015] une nouvelle contrainte structurelle qui renverse la problématique des recouvrements entre concepts en passant de recouvrements

"subis" ou "tolérés" à des recouvrements "choisis" et susceptibles de conduire à de nouvelles solutions conceptuellement pertinentes.

Dans la suite de ce résumé, (G, M, I) désigne un contexte formel tel que $(g, m) \in I$ (aussi noté gIm) si et seulement si l'objet g possède l'attribut m. On utilisera également les opérateurs de dérivation usuels A' et B' pour tout $A \subseteq G$ et $B \subseteq M$ qui correspondent respectivement à l'intension de A ($A' = \bigcap_{g \in A}\{m \in M | gIm\}$) et à l'extension de B ($B' = \bigcap_{m \in B}\{g \in G | gIm\}$).

Définition 3 *Soit (G, M, I) un contexte formel, une* **classification conceptuelle** $\mathcal{C} = \{(A_1, B_1), \ldots, (A_p, B_p)\}$ *désigne un ensemble de concepts dans (G, M, I) tels que :*

i) $\forall i,\ A'_i = B_i$ *et* $B'_i = A_i$,

ii) $\forall i,\ A_i \neq \emptyset$,

iii) $\forall i, j,\ A_i \cap A_j \notin \{A_i, A_j\}$,

iv) $\bigcup_{i=1}^{p} A_i = G$

Les quatre propriétés vérifiées par une classification conceptuelle assurent que chaque classe est un concept formel (i), dont l'extension est non-vide (ii), qu'il n'y a pas d'inclusions entre classes (iii) et enfin que la classification réalise bien une couverture de l'ensemble des objets (iv).

Les propriétés précédentes sont certes contraignantes mais insuffisamment pour espérer extraire des classifications conceptuelles intéressantes d'un espace de recherche très vaste. En effet aucune contrainte relative (entre concepts), autre que l'inclusion, n'étant spécifiée, plusieurs concepts ne se différenciant que par un seul objet pourront par exemple apparaître dans une solution sans apporter d'intérêt particulier à sa sémantique mais tout en augmentant inutilement la taille de cette solution et par la même la taille de l'espace de recherche.

Nous introduisons une nouvelle contrainte structurelle visant à modéliser conceptuellement les recouvrements/chevauchements entre classes. Nous considérons que **pour être pertinent, un recouvrement entre deux ou plusieurs concepts formels doit pouvoir s'expliquer conceptuellement par les attributs des concepts concernés, et uniquement ceux-ci**. Ainsi, considérant un objet g, situé par exemple à l'intersection de deux concepts formels A_1 et A_2 alors cet objet possède au minimum tous les attributs de $A'_1 \cup A'_2$; si g possède un attribut supplémentaire m, celui-ci n'étant pas caractéristique des concepts initiaux ($m \notin A'_1 \cup A'_2$) il ne devrait pas non plus être caractéristique d'un objet de l'intersection, autrement dit il doit exister des objets dans les concepts initiaux qui possèdent également cet attribut. Cette nouvelle contrainte structurelle est formalisée par la propriété (v) suivante

$$v)\ \forall m \in M,\ \forall \mathcal{S} \subset \mathcal{C},\ \forall g \in \bigcap_{\mathcal{S}} A_i,\quad gIm \Rightarrow \forall A_i \in \mathcal{S},\ \exists g_i \in A_i \setminus \bigcap_{\mathcal{S}} A_i,\ \text{t.q. } g_i Im \quad (30)$$

obligeant chaque attribut m d'un objet g situé à l'intersection d'une famille $\mathcal{S}$ de concepts, à être possédé par au moins un objet de chaque concept (en dehors de l'intersection). Cette nouvelle contrainte, ajoutée à celles d'une classification conceptuelle (Définition 3) nous permet d'introduire la notion de *couverture conceptuelle*.

Définition 4 *Soit* $\mathcal{C} = \{(A_1, B_1), \ldots, (A_p, B_p)\}$ *un ensemble de* p *concepts associé à un contexte formel* (G, M, I), $\mathcal{C}$ *est une* **couverture conceptuelle** *si et seulement si* $\mathcal{C}$ *est une classification conceptuelle satisfaisant également la propriété* v*).*

Étant donné un ensemble $\mathcal{C}$ de p concepts, vérifier que $\mathcal{C}$ satisfait aux exigences d'une couverture conceptuelle est a priori coûteux. Les propriétés i) à iv) peuvent être vérifiées simplement en observant chaque concept (pour les propriétés i), ii) et iv)) et chaque paire de concepts (pour la propriété iii)) ; en revanche la propriété v) que nous avons introduite nécessiterait a priori de considérer l'ensemble des 2^p sous-familles de concepts. Cependant, on peut montrer que cette contrainte présente une bonne propriété de monotonie permettant de vérifier la satisfaction en ne considérant que les paires de concepts (proposition 1).

Proposition 1 *Soient un contexte formel* (G, M, I) *et* $\mathcal{S}$ *une famille d'au moins* $k+1$ *concepts formels extraite de ce contexte. Si* $\mathcal{S}$ *satisfait la propriété* v*) pour tout recouvrement de* k *de ses concepts alors* v*) est également satisfaite pour tout recouvrement de* $k+1$ *concepts de* $\mathcal{S}$*.*

Sur la base Iris [D.J. Newman and Merz, 1998] discrétisée de manière à décrire les 150 objets iris selon 8 attributs binaires, nous avons généré l'ensemble des solutions possibles en considérant :

- les classifications conceptuelles non-contraintes (propriétés i) à iv))
- les partitions conceptuelles (idem en interdisant tout recouvrement entre concepts)
- les couvertures conceptuelles (cf. définition 4)

La Figure 12 (graphique de gauche) présente la distribution des solutions relativement au nombre de concepts qu'elles contiennent. On note alors que le filtre opéré par la contrainte structurelle v) permet de réduire significativement l'espace des solutions puisque seulement un tiers des classifications satisfont les propriétés d'une couverture conceptuelle [14]. Le graphique de droite fournit une analyse qualitative de ces trois différents ensembles de solutions par rapport à la classification de référence à l'aide de la mesure F-Bcubed. Il est alors très intéressant d'observer la qualité du filtre opéré par la contrainte v), en effet les couvertures conceptuelles apparaissent en moyenne davantage en correspondance avec la classification de référence que les classifications non-contraintes, mais également par rapport aux partitions conceptuelles.

Stratégie de recherche. L'exploration de l'espace des couvertures conceptuelles reste un problème difficile au même titre que la recherche d'une partition conceptuelle. Une manière de procéder consisterait à explorer l'espace des classifications conceptuelles (non contraintes) par fusions successives de concepts en partant d'une solution particulière dite "base minimale", définie par l'ensemble des concepts formels obtenus par les extensions élémentaires sur chaque objet de G. Partant de cette base minimale, nous proposons une approche hiérarchique bottom-up qui consiste à chaque étape de fusion à (1) favoriser l'émergence d'une couverture conceptuelle (ou à conserver ce statut lorsque

14. On compte au total 69,882 couvertures conceptuelles contre 207,259 classifications conceptuelles et 767 partitions conceptuelles.

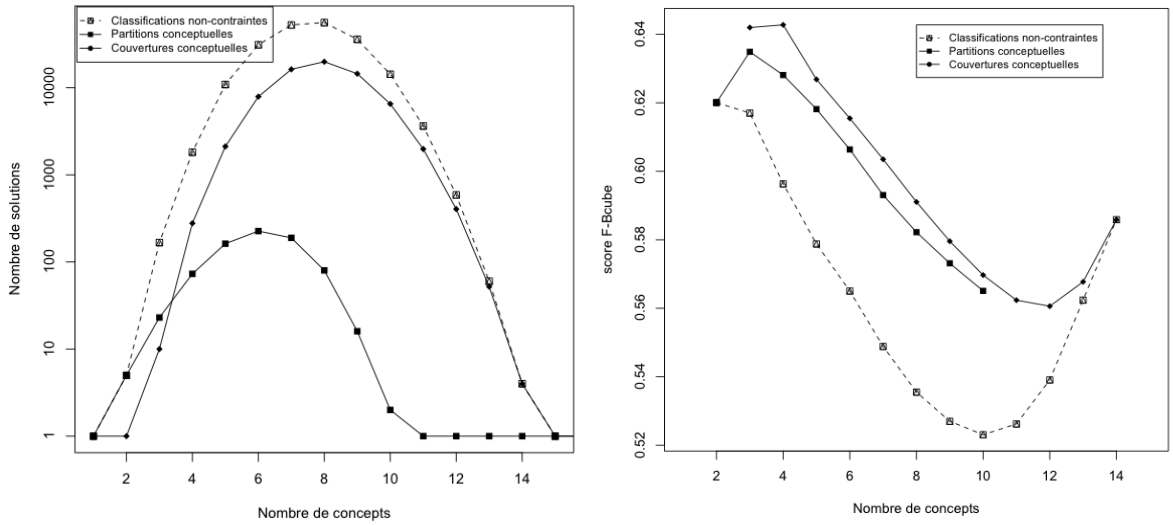


FIGURE 12 – Taille de l'espace de recherche (gauche) et évaluation externe (droite) des classifications sur la base *Iris* discrétisée.

la solution en cours correspond déjà à une couverture conceptuelle) et (2) préférer la génération de concepts plus spécifiques.

Cette stratégie n'offre pas en théorie l'assurance d'atteindre une couverture conceptuelle ; en pratique néanmoins, on observe sur diverses expérimentations que cette approche permet effectivement d'extraire un ensemble de couvertures conceptuelles.

Exemple. Afin d'illustrer le modèle précédent, nous considérons l'exemple ("*Living Context*") fréquemment utilisé en AFC, composé de huit êtres vivants (végétaux ou animaux) décrits selon neuf attributs booléens. Le contexte formel correspondant est proposé en Table 3. En complément, la Figure 13 présente le treillis des concepts formels associé à cet exemple ; on retrouve par exemple les concepts numérotés 15 et 16, facilement identifiables par les deux rectangles maximaux situés sur les trois premières colonnes de la Table 3.

Ces deux concepts sont particulièrement intéressants car ils forment à eux deux une couverture conceptuelle :

i) il s'agit de rectangles maximaux (concepts formels),

ii) leur extension n'est pas vide (ils couvrent chacun 5 objets),

iii) il ne sont pas inclus l'un dans l'autre,

iv) ils couvrent à eux deux l'ensemble des objets,

v) les trois premiers attributs suffisent à définir l'intersection de ces deux concepts.

Plus précisément, la pertinence conceptuelle du recouvrement des concepts 15 et 16 (propriété *v*)) est vérifiée en observant que les attributs ne participant pas à la définition de ces deux concepts ne sont pas de bons descripteurs du recouvrement ($\{$*Reed, Frog*$\}$) :

— le roseau (*reed*) a certes besoin de chlorophylle (*needs chlorophyll*), mais le maïs (*maize*) du concept 15 et le chardon d'eau (*spike-weed*) du concept 16 en ont également besoin,

	lives in water	needs water to live	lives on land	needs chlorophyll	two seed leaves	one seed leaf	can move around	has limbs	suckles its offspring
Leech	×	×					×		
Bream	×	×					×	×	
Spike-weed	×	×		×		×			
Reed	×	×	×	×		×			
Frog	×	×	×				×	×	
Dog		×	×				×	×	×
Bean		×	×	×	×				
Maize		×	×	×		×			

TABLE 3 – Contexte formel de huit êtres vivants ("*Living Context*").

— le roseau est monocotylédone (*one seed leaf*), mais une fois encore le maïs et le chardon d'eau le sont aussi,
— enfin la grenouille (*frog*) peut se mouvoir (*can move around*) et possède des membres (*has limbs*), mais la brème (*bream*) du concept 16 et le chien (*dog*) du concept 15 partagent également ces deux attributs.

Enfin, sémantiquement, cette couverture conceptuelle composée des concepts 15 et 16 et tout à fait intéressante à faire émerger dans la mesure où elle distingue les espèces vivant sur terre de celles vivant dans l'eau et identifie dans le même temps, par l'intermédiaire du recouvrement, les deux espèces "amphibiennes" que l'on peut observer à la fois sur terre et dans l'eau.

En guise de contre exemple, la collection composée des concepts 9, 14 et 15 ne correspond pas à la définition d'une couverture conceptuelle car le recouvrement sur les concepts 14 et 15 ne satisfait pas la contrainte de pertinence conceptuelle v). En effet d'après le treillis en Figure 13, on observe que le chien (qui est partagé par ces deux concepts) possède des membres tandis qu'aucune espèce propre au concept 15 n'en possède. Cet attribut possède donc un certain pouvoir discriminant vis-à-vis du recouvrement, alors qu'il n'intervient pas dans l'intension des concepts associés, ce qui contredit la propriété v).

Cet exemple de couverture non conceptuelle est également intéressant sémantiquement car il nous informe sur les organisations qui sont ou non autorisées. Au sommet du treillis, il faudra ainsi choisir entre une classification stricte ou recouvrante des espèces selon leur lieu de vie (terrestre-concept 15 vs. aquatique-concept 16) ou leur règne (animal-concept 14 vs. végétal-concept 17) mais aucune classification mélangeant ces deux modes de structuration ne seront possibles.

Dans le treillis (Figure 13), les concepts 1, 11, 3, 8 et 9 constituent la base minimale, autrement dit il est inutile de considérer les concepts situés sous cette base minimale pour la tâche de classification conceptuelle. Par exemple, puisque le concept 11 ou l'un de

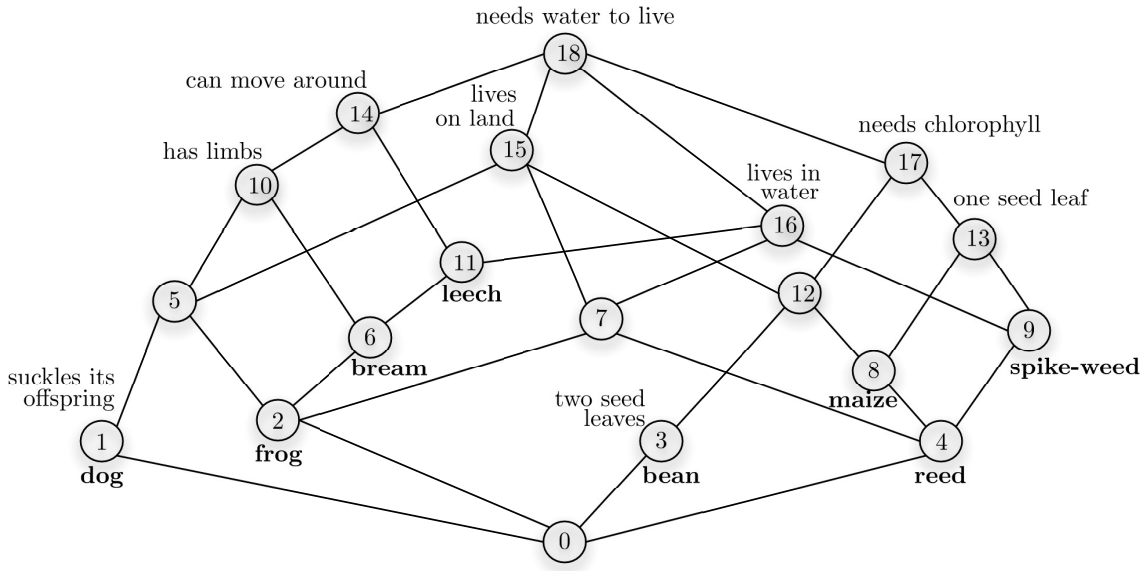


FIGURE 13 – Treillis des concepts formels (*Line diagram*) de l'exemple "*Living Context*".

ses sur-concepts doit nécessairement apparaître dans toute classification pour satisfaire la propriété (*iv*) de couverture de G (en particulier l'objet *leech* ici), les sous-concepts de 11 ne pourront pas apparaître dans une classification conceptuelle sans contredire la propriété (*iii*) de non-inclusion entre concepts.

Enfin nous présentons sur la Figure 14, deux structures hiérarchiques obtenues par fusions itératives de concepts à partir de la base minimale et selon la stratégie évoquée précédemment (préférence sur les fusions conduisant à des concepts plus spécifiques et favorisant l'émergence de couvertures conceptuelles). Pour les deux structures, on a constaté que chacun des niveaux des hiérarchies satisfont les propriétés d'une couverture conceptuelle. Enfin, ces deux exemples de structures sont représentatifs de l'ensemble des solutions envisageables par l'approche proposée puisqu'en étudiant toutes les stratégies de fusions possibles, elles conduisent toutes aux deux principales organisations déjà identifiées, le partage des espèces au sommet de la hiérarchie en deux grandes classes :

— espèces terrestres vs. espèces aquatiques avec un recouvrement sur les espèces amphibiennes (concepts 15 et 16),
— espèces animales vs. espèces végétales correspondant cette fois à une partition conceptuelle (concepts 14 et 17).

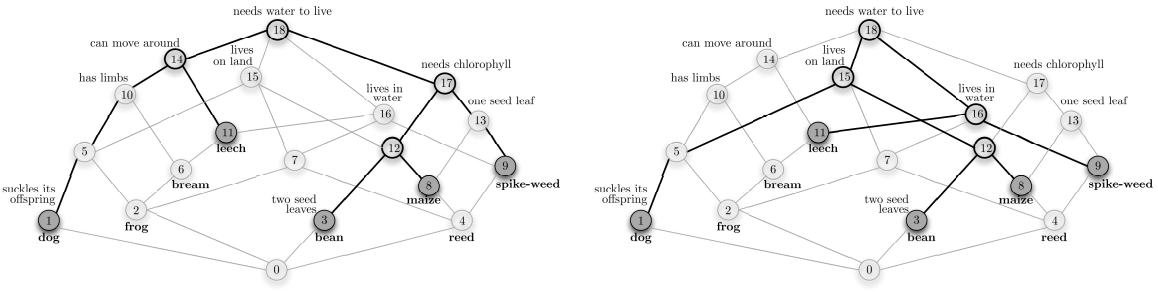


FIGURE 14 – Projection sur le treillis de deux structures hiérarchiques obtenues par fusions itératives de concepts à partir de la base minimale.

Notons que, même si nous avons utilisé le treillis de concepts pour illustrer le fonctionnement de notre approche, en pratique celle-ci ne nécessite nullement la construction d'un tel treillis, ni même le calcul a priori de l'ensemble des concepts formels induit par un contexte (G, M, I) donné.

Après avoir introduit, dans le premier chapitre, un nouveau modèle de classification recouvrante, nous avons consacré ce second chapitre d'une part à l'enrichissement de ce modèle et d'autre part à la transposition de celui-ci pour répondre à des préoccupations à la fois théoriques et applicatives en analyse de données. Pour cela nous avons été amené à explorer des sous-domaines et des théories plus variés tels que les approches neuronales, l'analyse de données symboliques et l'analyse formelle de concepts. Le chapitre à venir viendra clôturer cette première partie dédiée aux contributions sur la classification non-supervisée, en abordant un autre type complexe de structuration que constitue la reconnaissance de formes symétriques dans les données.

Publications associées au chapitre 2

[Ben N'cir et al., 2014] Ben N'cir, C., Cleuziou, G., and Essoussi, N., (2014). Generalization of c-means for identifying non-disjoint clusters with overlap regulation. In *Pattern Recognition Letters*, 45 :92–98.

[Ben N'Cir et al., 2014b] Ben N'cir, C., Cleuziou, G., and Essoussi, N., (2014). Généralisation des k-moyennes pour produire des recouvrements ajustables. In *14èmes Journées Francophones Extraction et Gestion des Connaissances, EGC 2014*, Rennes, France, 28-32 Janvier, 2014, pages 209–220.

[Cleuziou, 2007] Cleuziou, G. (2007). Classification recouvrante avec pondération locale des attributs. In *14èmes rencontres de la Société Francophone de Classification*, Paris, France, pages 58–61.

[Cleuziou, 2009a] Cleuziou, G., (2009). Okmed et wokm : deux variantes de okm pour la classification recouvrante. In *9èmes Journées Francophones d'Extraction et de Gestion des Connaissances*, volume 2. Revue des Nouvelles Technologies de l'Information, Cépaduès-Edition.

[Cleuziou, 2009b] Cleuziou, G., (2009). Two variants of the okm for overlapping clustering. In Guillet, F., Ritschard, G., Zighed, D. A., and Briand, H., editors, EGC (best of volume), volume 292 of *Studies in Computational Intelligence*, pages 149–166, Springer.

[Cleuziou, 2009c] Cleuziou, G. (2009). Adaptation des modèles d'auto-organisation pour la classification recouvrante. In *16èmes rencontres de la Société Francophone de Classification*, Grenoble, France, pages 11–14.

[Cleuziou, 2010] Cleuziou, G. (2010). OSOM : un algorithme de construction de cartes topologiques recouvrantes. In *10èmes Journées Francophones Extraction et Gestion des Connaissances, EGC 2010*, pages 97–108.

[Cleuziou, 2013] Cleuziou, G., (2013). Osom : A method for building overlapping topological maps. In *Pattern Recognition Letters*, 34(3) :239–246.

[Cleuziou, 2014] Cleuziou, G. (2014). Passage aux noyaux en classification recouvrante. In *14èmes Journées Francophones Extraction et Gestion des Connaissances, EGC 2014*, Rennes, France, 28-32 Janvier, 2014, pages 209–220.

[Cleuziou and Crémilleux, 2015] Cleuziou, G. and Crémilleux, B. (2015). Vers une classification conceptuelle recouvrante. In *22èmes rencontres de la Société Francophone de Classification*, Nantes.

[Cleuziou and de Carvalho, 2014] Cleuziou, G. and de Carvalho, F. (2014). Robustesse en classification recouvrante : une approche par trimming. In *21èmes rencontres de la Société Francophone de Classification*, Rabbat, Maroc.

[Cleuziou et al., 2012] Cleuziou, G., de Carvalho, F., and Rousseau, L. (2012). Okm-L1 et comparaison de recouvrements. In *19èmes rencontres de la Société Francophone de Classification*, Marseille.

3 Reconnaissance de formes symétriques

3.1 Introduction au clustering point-symétrique

Ce chapitre s'intéresse à la problématique de reconnaissance de formes symétriques dans les données à partir de techniques de clustering. Cette tâche trouve naturellement ses applications dans le domaine de la vision par ordinateur et en analyse d'images où il peut être très utile de détecter des objets physiques de formes souvent symétriques ou approximativement symétriques (e.g. bâtiments, visages, aliments, etc.). Le clustering a été utilisé pour cette tâche depuis le début des années 2000, avec une première contribution portant sur la définition d'une première distance (non-métrique) dite "point-symétrique" exploitée par une variante de k-moyennes afin de construire des clusters présentant une symétrie centrale par rapport à leur prototype (cercles, anneaux, bandes, étoiles, etc.) [Su and Chou, 2001]. Par la suite, des améliorations ont été apportées tant sur la définition de la distance que sur le processus de clustering afin de : (1) traiter des situations où les clusters sont eux-mêmes positionnés de manière symétrique dans l'espace [Chou et al., 2002, Chung and Lin, 2007] et (2) mieux explorer l'espace des solutions, sous exploité par l'approche originelle [Chung and Lin, 2007, Bandyopadhyay and Saha, 2007].

D'une manière générale les approches proposées jusqu'ici parviennent à reconnaitre des clusters symétriques, comme le montrent les expérimentations conduites aussi bien sur des données générées artificiellement que sur des images réelles. Cependant, toutes ces approches considèrent une représentation des données dans un espace Euclidien et n'envisagent pas la possibilité d'appliquer ces techniques sur des données pour lesquelles la distance Euclidienne n'est pas adaptée alors que des organisations en clusters symétriques peuvent pourtant apparaître dans des espaces de représentation alternatifs.

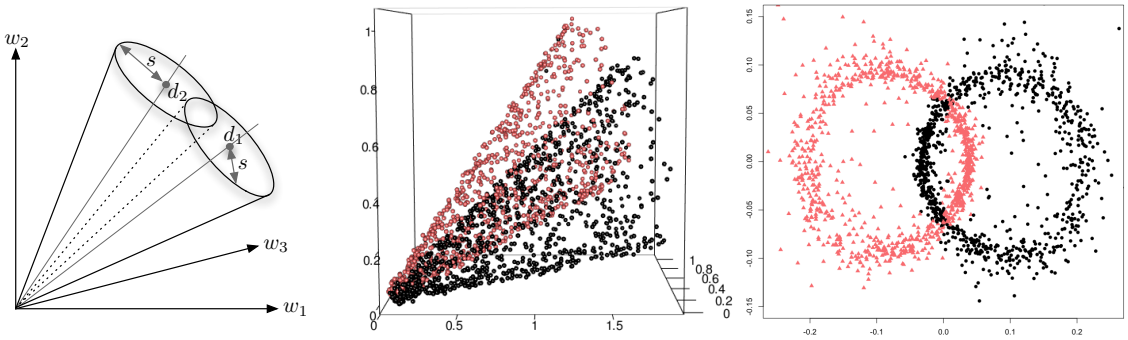


FIGURE 15 – Génération artificielle d'un jeu de données dans $\mathbb{R}^3$ avec deux ensembles de points, chacun étant composé de 10,000 données situées à une similarité du cosinus $s \pm \epsilon$ d'une donnée (d_1 ou d_2) : principe de génération (à gauche), visualisation du jeu de données dans l'espace Euclidien (au centre), visualisation 2D dans un espace reconstruit par MDS à partir de la similarité du cosinus (à droite).

Par exemple en Recherche d'Information, et cela permettra de réaliser un premier lien avec la seconde partie de ce mémoire, il peut être intéressant de rechercher dans un corpus un ensemble D de documents textuels assez similaires à un document de référence d (même sujet), mais également suffisamment éloignés (informations complémentaires). L'utilisation de la similarité du cosinus étant recommandée pour les données textuelles,

l'ensemble D aura une forme d'hyper-cône dans un espace Euclidien et ne pourra pas être reconnu par les approches actuelles, alors qu'il s'agirait bien d'une hyper-sphère d'une certaine épaisseur dans l'espace muni de la distance issue de la similarité du cosinus (cf. Figure 15).

L'astuce du noyau, comme nous l'avons vu au chapitre précédent, est un procédé mathématique puissant permettant de réaliser un traitement sur les données (par exemple un processus de clustering) dans un espace de projection implicite. Afin de lever les limites mentionnées des approches de clustering point-symétrique, nous avons proposé dans [Cleuziou and Moreno, 2015] une formulation à noyau de ces méthodes, ce qui permet de les rendre compatibles avec tout type de similarité pouvant être formalisée par une matrice noyau (semi-définie positive).

3.2 Approches existantes de clustering point-symétrique

Le génération de clusters symétriques s'appuie sur la définition d'une mesure de distance entre une donnée x_i et un cluster π_k, permettant de quantifier la contribution de x_i contribue à une forme symétrique pour π_k. Les distances proposées pour rechercher des formes présentant une symétrie centrale reposent sur l'idée que x_i contribuera fortement à la forme symétrique de π_k (faible distance) s'il existe une donnée proche du point symétrique de x_i par rapport au centre du cluster π_k (cf. Figure 16).

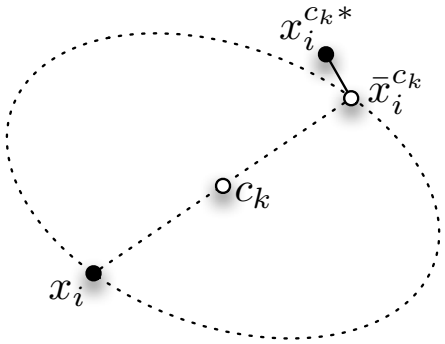


FIGURE 16 – Illustration d'un objet x_i, de son point symétrique $\bar{x}_i^{c_k}$ par rapport au centre c_k (d'un cluster π_k) et de son plus proche objet symétrique $x_i^{c_k*}$.

Trois distances "point-symétrique" ont été proposées. Nous les présentons dans la Table 4 telles que formalisées par [Bandyopadhyay and Saha, 2007]. La première distance d_1 fait d'un objet x_i un bon candidat pour appartenir à un cluster π_k non-seulement si il est possible de trouver dans X un objet $x_i^{c_k*}$ qui soit à une petite distance (Euclidienne) du point symétrique $\bar{x}_i^{c_k}$ (numérateur), mais surtout cette distance doit être d'autant plus petite que les objets (x_i et $x_i^{c_k*}$) sont éloignés du centre c_k (dénominateur).

La seconde distance d_2 corrige un biais de la première qui consiste à préférer l'affectation des objets à des clusters éloignés, en pondérant d_1 par la distance Euclidienne entre l'objet x_i et le centre c_k du cluster. Enfin, la dernière proposition d_3 se veut plus

[Su and Chou, 2001]	$d_1(x_i, c_k) = \frac{\|\bar{x}_i^{c_k} - x_i^{c_k*}\|}{\|x_i - c_k\| + \|x_i^{c_k*} - c_k\|}$
[Chou et al., 2002]	$d_2(x_i, c_k) = \frac{\|\bar{x}_i^{c_k} - x_i^{c_k*}\|}{\|x_i - c_k\| + \|x_i^{c_k*} - c_k\|}.\|x_i - c_k\|$
[Bandyopadhyay and Saha, 2007]	$d_3(x_i, c_k) = \frac{\|\bar{x}_i^{c_k} - x_i^{c_k*}\| + \|\bar{x}_i^{c_k} - x_i^{c_k**}\|}{2}.\|x_i - c_k\|$

TABLE 4 – Différentes propositions de distances point-symétrique.

robuste en considérant les deux plus proches voisins ($x_i^{c_k*}$ et $x_i^{c_k**}$) du point symétrique, tout en supprimant le terme de normalisation utilisé dans les deux premières distances.

Étant donnée une distance point-symétrique $d_{sym} : X \times \mathbb{R}^P \to \mathbb{R}^+$, le processus de clustering sera ensuite guidé par la minimisation d'un critère de moindres carrés basé sur la distance d_{sym} :

$$J_{SBKM}(\Pi, C) = \sum_{k=1}^{K} \sum_{x_i \in \pi_k} d_{sym}(x_i, c_k)^2. \tag{31}$$

Notons que cette formalisation s'écarte du cadre théorique habituel des méthodes de réallocations dynamiques présentées jusqu'ici : la fonction objective (31) ne s'appuie plus sur une décomposition de l'inertie des données (théorème de Huygens) et il ne s'agit plus d'un problème d'optimisation convexe ; il ne sera donc pas recherché une minimisation stricte de cette fonction. L'algorithme de base SBKM *Symmetry-Based K-Means* [Su and Chou, 2001, Chou et al., 2002, Chung and Lin, 2007] procède par l'itération classique de deux étapes :

1. affectation de chaque objet x_i au plus proche cluster [15] selon d_{sym} :

$$\pi(x_i) = \arg\min_k d_{sym}(x_i, c_k),$$

2. mise à jour des prototypes de classes par l'isobarycentre des objets :

$$c_k = \frac{\sum_{x_i \in \pi_k} x_i}{|\pi_k|}$$

tandis que [Bandyopadhyay and Saha, 2007] proposent d'explorer l'espace des solutions par méta-heuristique via l'algorithme GAPS (*Genetic Algorithm with Point Symmetry*).

Nous présentons en Figure 17 un aperçu de quelques résultats accessibles par les approches présentées ci-dessus sur des jeux de données générés artificiellement.

15. Pour éviter la génération de clusters ayant une forme géométrique trop dégradée, une constante θ bien choisie est généralement utilisée de sorte que pour des distances point-symétriques toutes supérieures à θ, un objet est affecté sur la base de la distance Euclidienne.

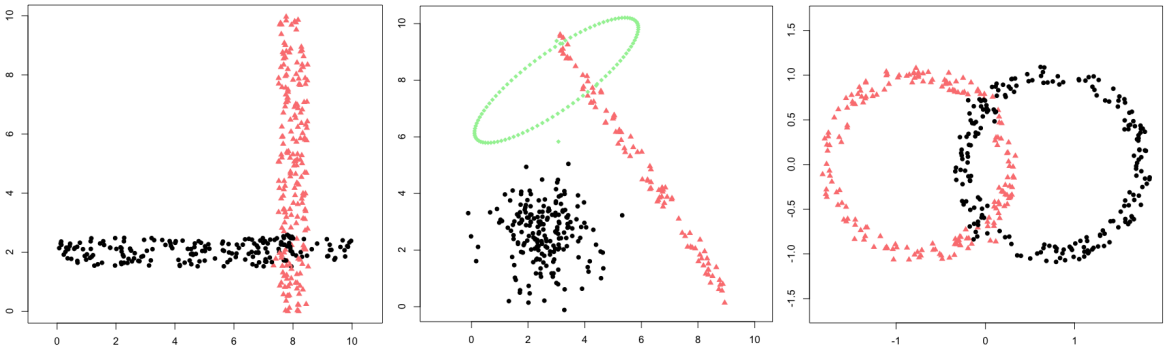


FIGURE 17 – Clusterings réalisés avec l'algorithme SBKM et les distances d_2 (à gauche et au centre) et d_1 (à droite).

Afin d'illustrer leurs limites, nous proposons enfin en Figure 18 de nouveaux jeux de données ainsi que les meilleurs résultats observés en utilisant les différentes combinaisons de distances et d'algorithmes proposés jusqu'ici.

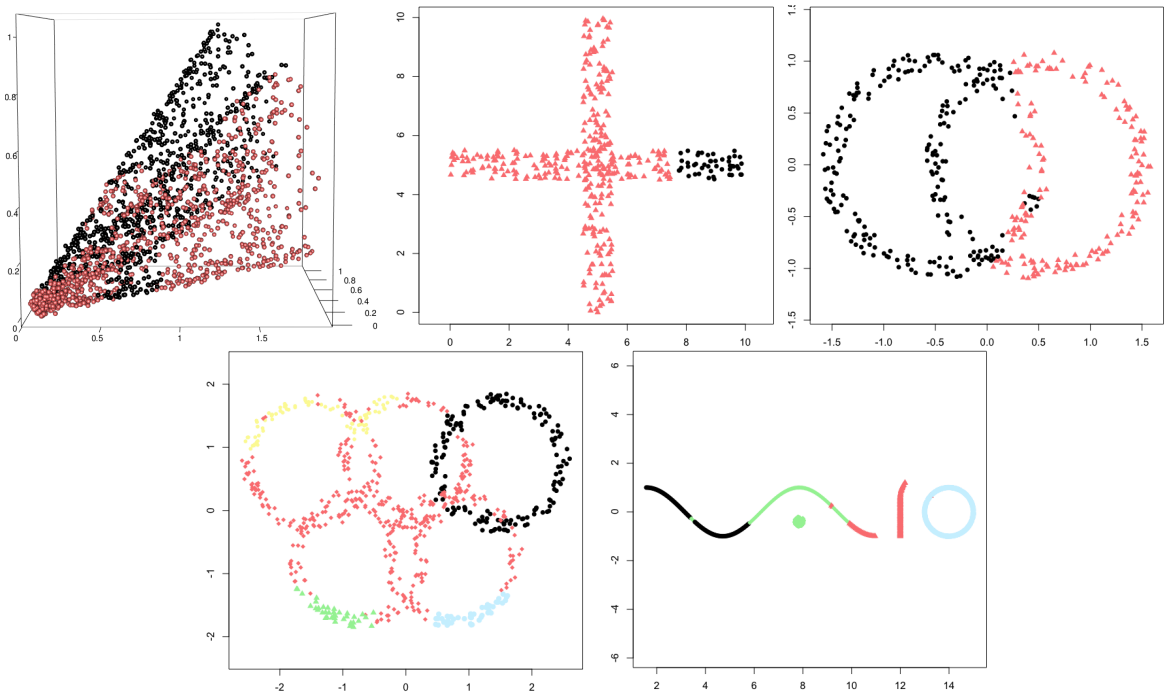


FIGURE 18 – Meilleurs clusterings obtenus par les approches point-symétriques existantes : (figure en haut à gauche) SBKM avec d_1, (autres figures) SBKM avec d_3.

3.3 Kernélisation du clustering point-symétrique

Comme nous l'avons déjà évoqué dans le chapitre précédent, le passage au noyau nécessite une réécriture du critère objectif de sorte qu'il s'appuie uniquement sur les produits scalaires entre objets, seule information disponible en entrée de l'algorithme. Pour les trois distances point-symétriques déjà évoquées dans la Table 4, cette réécriture n'est absolument pas triviale. En effet, contrairement au terme $\|x_i - c_k\|$ qui peut se réécrire aisément en introduisant la définition de c_k comme l'isobarycentre de la classe, le terme $\|x_i^{c_k*} - c_k\|$ utilisé dans d_1 et d_2 nécessiterait d'être en mesure d'identifier le

plus proche voisin du point symétrique, quant aux termes de la forme $\|\bar{x}_i^{c_k} - x_i^{c_k*}\|$ qui rendent compte de la distance entre un point symétrique (qui n'appartient pas à X a priori) et son plus proche voisin dans X, une réécriture analytique à partir des produits scalaires sur X semble encore plus difficile.

Un raisonnement plus approfondi faisant intervenir des propriétés de géométrie Euclidienne (théorème des médianes dans les triangles en particulier) nous a permis d'établir le théorème suivant :

Théorème 1 *Soient $X = \{x_1, \ldots, x_N\}$, $\pi_k \subseteq X$ et $\bar{x}_i^{c_k}$ le point symétrique de x_i par rapport à l'isobarycentre c_k de π_k (tel que $\bar{x}_i^{c_k} = 2c_k - x_i$), la distance Euclidienne entre $\bar{x}_i^{c_k}$ et tout élément $x_j \in X$ est donnée par*

$$\|\bar{x}_i^{c_k} - x_j\| = \left(2d_e(x_i, \pi_k)^2 + 2d_e(x_j, \pi_k)^2 - d_e(x_i, x_j)^2\right)^{\frac{1}{2}}$$

où $d_e(x_i, \pi_k)$ désigne la distance Euclidienne entre l'objet x_i et le (centre de gravité du) cluster π_k :

$$d_e(x_i, \pi_k) = \left(\langle x_i, x_i \rangle - 2.\frac{\sum\limits_{x_j \in \pi_k} \langle x_i, x_j \rangle}{|\pi_k|} + \frac{\sum\limits_{x_j \in \pi_k} \sum\limits_{x_l \in \pi_k} \langle x_j, x_l \rangle}{|\pi_k|^2} \right)^{\frac{1}{2}} \tag{32}$$

et $d_e(x_i, x_j)$ désigne la distance Euclidienne usuelle entre les deux objets x_i et x_j :

$$d_e(x_i, x_j) = (\langle x_i, x_i \rangle + \langle x_j, x_j \rangle - 2.\langle x_i, x_j \rangle)^{\frac{1}{2}} \tag{33}$$

Grâce au Théorème 1, il est possible d'exprimer la distance point-symétrique (selon d_1, d_2 et d_3) entre tout objet x_i et tout cluster $\pi_k \subseteq X$, dans l'espace de projection induit par un noyau $\mathcal{K} = \langle \phi(.), \phi(.) \rangle$ quelconque défini sur $X \times X$. Par exemple, pour la distance d_2 on procèdera de la manière suivante :

1. commencer par calculer d'après (32)

$$\|\phi(x_i) - c_k\| = d_e(\phi(x_i), \pi_k) = \left(\mathcal{K}_{i,i} - 2.\frac{\sum_{x_j \in \pi_k} \mathcal{K}_{i,j}}{|\pi_k|} + \frac{\sum_{x_j \in \pi_k} \sum_{x_l \in \pi_k} \mathcal{K}_{j,l}}{|\pi_k|^2} \right)^{\frac{1}{2}},$$

2. identifier le plus proche voisin du point symétrique de x_i par rapport au cluster π_k dans l'espace de projection (théorème 1) :

$$x_i^{c_k*} = \arg\min_{x_j \in X} 2d_e(\phi(x_i), \pi_k)^2 + 2d_e(\phi(x_j), \pi_k)^2 - d_e(\phi(x_i), \phi(x_j))^2$$

3. calculer les distances
   - $\|\phi(x_i^{c_k*}) - c_k\| = d_e(\phi(x_i^{c_k*}), \pi_k)$ d'après (32)
   - $\|\bar{x}_i^{c_k} - \phi(x_i^{c_k*})\| = (2d_e(\phi(x_i), \pi_k)^2 + 2d_e(\phi(x_i^{c_k*}), \pi_k)^2 - d_e(\phi(x_i), \phi(x_i^{c_k*}))^2)^{\frac{1}{2}}$ d'après le théorème 1

   puis en déduire la distance point-symétrique $d_2(\phi(x_i), \pi_k)$.

Une version kernélisée de l'algorithme SBKM a donc été proposée afin de réaliser un clustering point-symétrique dans un espace de projection induit par un noyau quelconque (Algorithme 4). Cet algorithme est guidé par une version kernélisée de la fonction objective originelle :

$$J_{K-SBKM}(\Pi) = \sum_{k=1}^{K} \sum_{x_i \in \pi_k} d_{sym}(\phi(x_i), \pi_k)^2. \tag{34}$$

Algorithme 4 K-SBKM

procedure K-SBKM($X, \mathcal{K}, K, T, d_{sym}$)
  $t \leftarrow 0$
  Initialisation aléatoire des clusters : $\pi_k^{(0)} \leftarrow Random(x_1, \ldots, x_N), \quad k = 1 \ldots K$
  for $x_i \notin \bigcup_k \pi_k^{(0)}$ **do** Affectation des individus au plus proche cluster
    $\pi_i^{(0)} \leftarrow \arg\min_{k=1\ldots K} d_e(\phi(x_i), \pi_k)$
  end for
  Calculer $J_{K-SBKM}^{(0)}$
  repeat
    $t \leftarrow t + 1$
    for $x_i \in X$ **do** Réaffectation des individus au plus proche cluster symétrique
      $\pi_i^{(t)} \leftarrow \arg\min_{k=1\ldots K} d_{sym}(\phi(x_i), \pi_k^{(t-1)})$
    end for
    Calculer $J_{K-SBKM}^{(t)}$
  until $J_{K-SBKM}^{(t)} \geq J_{K-SBKM}^{(t-1)}$ ou $t = T$ Conditions d'arrêt
  Return $\Pi^{(t)}$
end procedure

Cet algorithme prend en entrée l'ensemble des objets à traiter X, un noyau $\mathcal{K}$ sur ces objets, un scalaire K correspondant au nombre de clusters attendus, un nombre d'itérations maximum T ainsi qu'un choix de distance point-symétrique d_{sym} (typiquement d_1, d_2 ou d_3). L'algorithme débute par une initialisation des clusters par tirage aléatoire de K objets (graines) suivie d'une affectation de chacun des autres objets à leur plus proche graine au sens de la distance Euclidienne, dans l'espace de projection induit par $\mathcal{K}$. Le processus de construction des clusters symétriques peut alors commencer, celui-ci consiste en une réallocation dynamique de l'ensemble des objets au cluster le plus proche cette fois au sens de la distance d_{sym} et en considérant à chaque fois les clusters obtenus à l'itération précédente.

En terme de complexité, une mutualisation du calcul de certains termes permet d'atteindre facilement une complexité en $O(TKN^3)$ pour K-SBKM, à comparer avec une complexité en $O(TKPN^2)$ pour l'algorithme originel non-kernélisé SBKM, où T désigne le nombre maximum d'itérations, K le nombre de clusters, P la dimensionalité des données et N le nombre d'objets à traiter. Ainsi, outre les avantages liés à la généricité de l'algorithme $K-SBKM$, il est important d'observer que même sans changer d'espace de représentation, il est plus efficace d'avoir recours à la version kernélisée lorsque la dimensionalité des données est plus grande que le nombre d'objets (ce qui peut être le cas pour les données textuelles par exemple).

3.4 Résultats et discussion

Sur le jeu de données construit artificiellement et composé de deux cônes de 10,000 objets chacun, nous avons exécuté l'algorithme K-SBKM à partir d'un noyau défini par la similarité du cosinus :

$$\mathcal{K}(x_i, x_j) = \frac{\langle x_i, x_j \rangle}{\|x_i\| \|x_j\|}.$$

Nous montrons en Figure 19 une visualisation des clusters symétriques obtenus par K-SBKM et qui correspondent avec une précision de plus de 90% aux deux classes symétriques attendues.

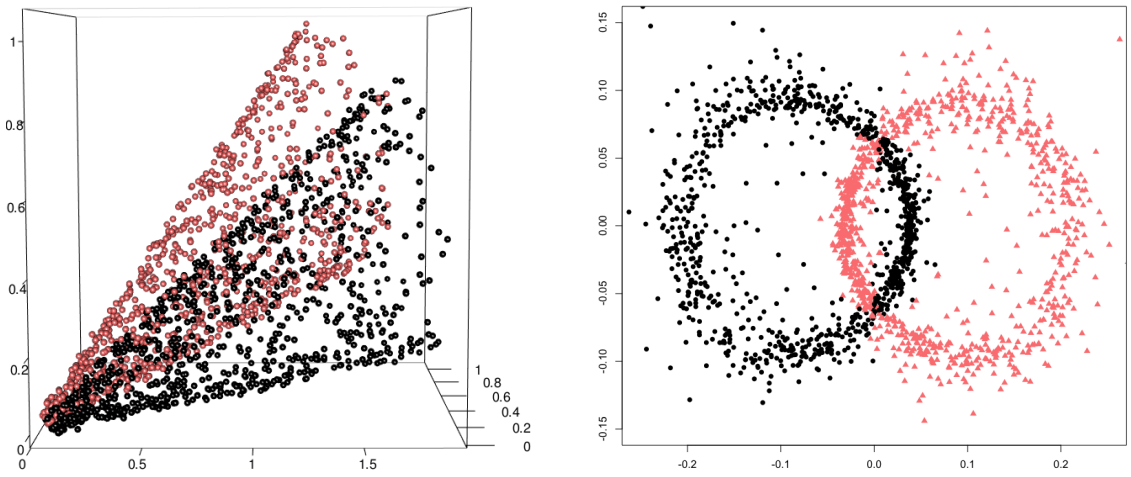


FIGURE 19 – Clusterings réalisés avec l'algorithme K-SBKM utilisant un noyau défini par la similarité du cosinus et la distances point-symétrique d_2. Visualisations dans l'espace Euclidien (à gauche) et dans l'espace induit par le noyau, reconstruit par MDS (à droite).

Sur les quatre autres jeux de données présentés en Figure 18 et pour lesquels les méthodes actuelles sont inefficaces, nous avons testé l'utilisation de noyaux usuels (gaussiens et polynomiaux) avec différents paramétrages afin d'observer si la projection des données dans un nouvel espace pouvait aider à reconnaître des formes symétriques difficiles à capturer dans l'espace de représentation initial. Nous proposons en Figure 20 les meilleurs résultats obtenus.

Il apparaît en effet que certaines projections usuelles permettent d'identifier des clusters symétriques complexes, du fait de leur entrelacement, de leur position relative ou de la variabilité de leur forme. L'amélioration est généralement observée pour des projections induisant une distorsion légère des données (faible degré pour un noyau polynomial, variance élevée pour un noyau gaussien), de sorte que les caractéristiques de symétrie demeurent mais deviennent plus faciles à capturer dans le nouvel espace. La problématique du choix d'un bon noyau peut de plus être partiellement résolue dans le context présent, puisque dans la plupart des situations, la valeur de la fonction objective (31) dans l'espace initial reste un bon critère pour comparer a posteriori et de manière non-supervisée différents clusterings symétriques. Cette idée constitue une première piste en vue d'une étude plus approfondie sur la sélection de noyaux pour le clustering symétrique.

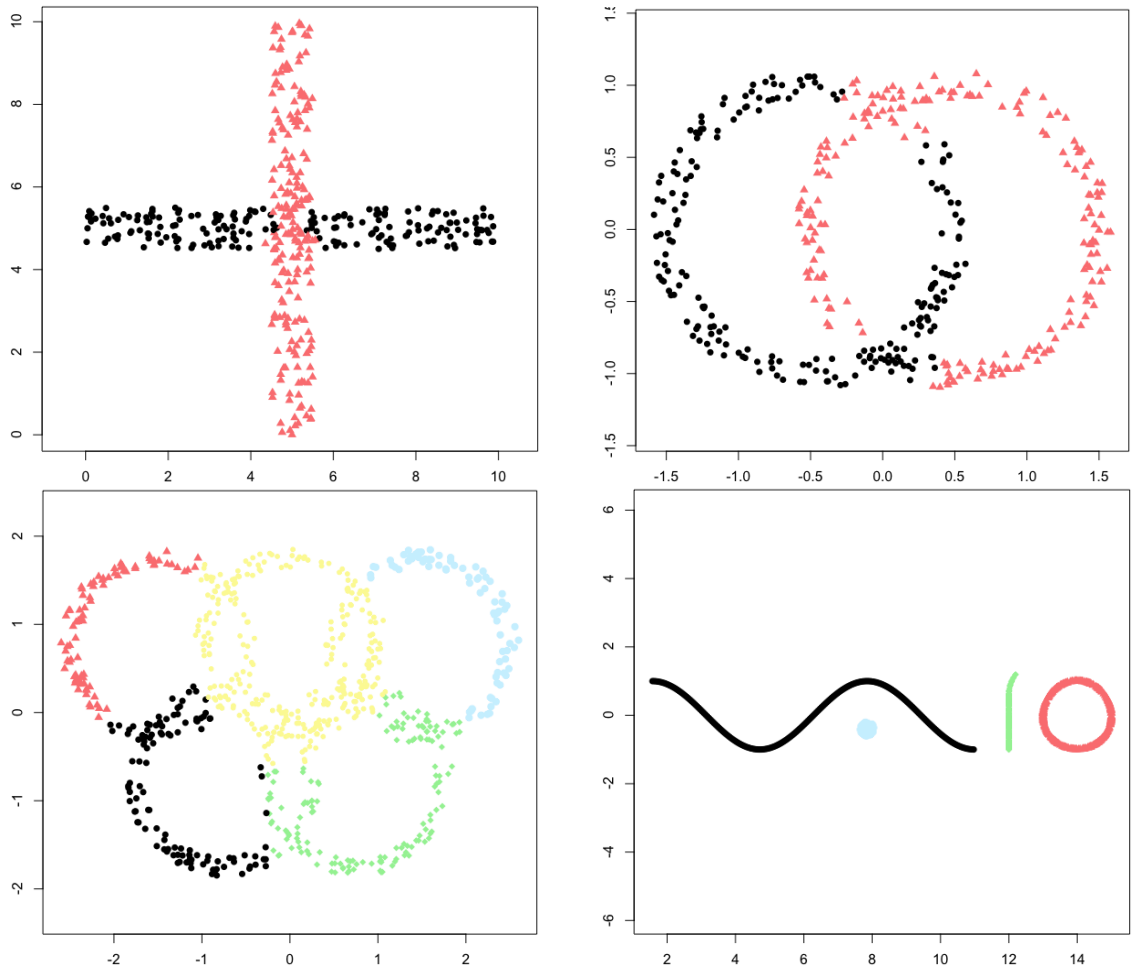


FIGURE 20 – Meilleurs clusterings obtenus avec l'algorithme K-SBKM : (figure en haut à gauche) avec un noyau gaussien ($\sigma = 2$) et la distance d_1, (figure en haut à droite) avec un noyau gaussien ($\sigma = 3$) et la distance d_1, (figure en bas à gauche) avec un noyau polynomial (de degré 2) et la distance d_3, (figure en bas à droite) avec un noyau gaussien ($\sigma = 1$) et la distance d_1.

Publications associées au chapitre 3

[Cleuziou and Moreno, 2015] Cleuziou, G. and Moreno, J. G. (2015). Kernel Methods for Point Symmetry-based Clustering. *Pattern Recognition*, 48 :2812–2830.

La première partie de ce mémoire a permis de synthétiser, dans un formalisme unifié, un ensemble de contributions dans le domaine du clustering, principalement ancrées dans la famille des méthodes dites de réallocations dynamiques, et guidées par le souci d'apporter à la communauté scientifique et aux utilisateurs des solutions intuitives et bien fondées à des problématiques réelles.

Chaque chapitre ou sous-chapitre pourrait donner lieu (et parfois donnera lieu) à de nouvelles perspectives de recherches propres. Par exemple on pourrait ne pas se contenter du contrôle offert à l'utilisateur par le modèle R-Okm sur les recouvrements et souhaiter aller plus loin en proposant un apprentissage automatique de ce paramétrage. On pourrait travailler à l'exploitation des cartes topologiques recouvrantes par des segmentations hiérarchiques recouvrantes et étudier les connexions théoriques d'ordre structurel avec les approches de classification pseudo-hiérarchiques (e.g. pyramides, hiérarchies faibles). Nous devrons étendre l'usage des méthodologies de classification recouvrante à de nouveaux types de données symboliques (e.g. données histogrammes) afin d'envisager une utilisabilité pratique de cet outil d'analyse dans le contexte de l'ADS. Enfin, l'ouverture proposée vers la classification conceptuelle recouvrante constitue une piste de recherche prometteuse qui devra être notamment poursuivie par la proposition de stratégies d'exploration précises et confirmées expérimentalement, ainsi que par l'étude de nouvelles contraintes structurelles complémentaires sur les couvertures conceptuelles.

Mais au-delà de ces perspectives "locales", il demeure une question plus "globale", qui couvre l'ensemble des trois chapitres précédents et qui n'a pas été abordée jusqu'ici. Il s'agit de la problématique de la sélection de modèles, non-seulement au sens usuel du choix du nombre de classes qui demeure un problème ouvert et particulièrement dans le cas du clustering recouvrant ou symétrique, mais aussi au sens du choix du modèle de recouvrement adapté aux données. Une façon d'appréhender cette problématique difficile dans un contexte purement non-supervisé tel qu'envisagé jusqu'ici, serait de profiter des formalisations à noyau que nous avons développé pour chaque type de clustering (K-OKSet pour le clustering recouvrant et K-SBKM pour le clustering symétrique) et de les utiliser dans un contexte semi-supervisé où l'utilisateur interagit de façon légère sur la classification par le biais de contraintes formulées au niveau des objets et intégrées dans le noyau.

Deuxième partie

Structuration de données textuelles

Sommaire

La structuration de données textuelles est un domaine d'étude très vaste qui rassemble (et parfois oppose) depuis plusieurs décennies des chercheurs issues de communautés diversifiées, notamment des mathématiciens, des informaticiens, des linguistes et des cogniticiens. La magie opère lorsque les modèles et algorithmes proposés par les uns rencontrent harmonieusement les représentations et théories élaborées par les autres.

Ces différentes communautés se retrouvent sur des problématiques de traitement du language naturel, d'ingénierie des connaissances ou encore de recherche d'information afin de répondre à des problématiques aux enjeux sociétaux déterminants tels que : la traduction automatique, la veille scientifique, l'élaboration de bases de connaissances (dictionnaires, thésaurus, ontologies), le traitement de la parole, etc. Pour chacune de ces problématiques, l'exploitation de la ressource textuelle passe généralement par une phase de structuration : il peut s'agir par exemple de structurer les mots d'une phrase pour en exploiter la structure syntaxique voire sémantique (structure d'arbre), de structurer les termes d'un domaine afin d'en dégager une organisation conceptuelle (structure de graphe) ou encore de structurer une collection de documents dans le but d'en faire émerger une typologie thématique (structure de classes).

Qu'il s'agisse d'une finalité ou d'un préalable à d'autres traitements, la structuration doit être automatisée dès lors que le traitement doit être fait en temps réel et/ou la quantité d'informations textuelles est importante. Si dans un premier temps il a semblé naturel de mettre au point des méthodologies figées (e.g. grammaires formelles, systèmes experts) sur des domaines (e.g. la médecine) ou des langues (principalement l'anglais) de pointe, il s'est avéré indispensable d'envisager une acquisition automatique de ces modèles afin d'universaliser les domaines et les langues exploitables.

La seconde partie de ce mémoire synthétise un ensemble de contributions dans le domaine de la structuration automatique de données textuelles par apprentissage et à partir de l'observation de textes bruts. Nos travaux ont donné lieu à des contributions dans chacune des trois problématiques mentionnées précédemment à savoir : la recherche d'information, l'ingénierie des connaissances et le traitement automatique des langues. Les deux premières problématiques sont abordées respectivement dans les deux chapitres à venir, tandis que l'étude menée, dans le cadre d'un encadrement doctoral, sur l'apprentissage de structures de dépendances est évoquée dans l'Annexe B.

Le chapitre 4 s'intéresse à la tâche spécifique, et relativement récente, de structuration des résultats de recherches Web. La spécificité de cette tâche réside, non pas dans la quantité de données à traiter, mais plutôt dans la difficulté à appréhender, en temps réel, des données textuelles très courtes et donc peu informatives (snippets), à les organiser en classes thématiques homogènes et à proposer une étiquette représentative dans une forme facilement compréhensible par un internaute. Les approches développées adaptent des méthodologies de classification traditionnelles de manière à réaliser, à l'aide d'une mesure de proximité fine sur les données, la classification et l'étiquetage des classes (approche dite polythetique) de façon simultanée et dans des espaces de représentation éventuellement séparés.

Enfin, le dernier chapitre porte sur l'acquisition automatique de taxonomies lexicales par une approche hybride, c'est-à-dire exploitant dans un même formalisme, des informations à la fois linguistiques et statistiques émanant de corpus textuels du do-

maine considéré. La tâche se ramène à la structuration d'un ensemble de termes en un graphe orienté sans cycles, modélisant des relations de subsomption (e.g. hyperonyme/hyponyme) entre les termes. Nous avons proposé d'une part d'utiliser la théorie de la prétopologie afin de modéliser la relation de subsomption par un processus de propagation complexe et adapté et d'autre part d'étendre cette théorie par la proposition d'une nouvelle classe d'espaces prétopologiques sur laquelle nous proposons une méthode d'apprentissage par une approche semi-supervisée.

4 Structuration des résultats de recherches Web

L'augmentation spectaculaire de la quantité d'information mise à disposition sur le Web (réseaux sociaux, blogs, etc.) nécessite de reconsidérer la manière dont le Web est interrogé et en particulier la manière de présenter les résultats d'une requête à un utilisateur. Aujourd'hui en 2015, plus des deux tiers des internautes utilisent le moteur de recherche Google© et récupèrent, pour chaque requête posée, une liste de documents - représentés par leur titre ainsi qu'un extrait textuel (*snippet*) - et ordonnés par importance selon l'algorithme du PageRank [Page et al., 1999]. Une telle liste ne permet pas d'obtenir une vue globale des documents retournés et requiert pour l'utilisateur un travail complémentaire d'analyse qui promet de s'intensifier avec l'évolution du Web. Les acteurs du Web travaillent depuis une vingtaine d'années au développement de techniques de structuration des résultats d'une requête en classes thématiques labélisées, comme le montrent certaines études publiées par Xerox© [Hearst and Pedersen, 1996], Microsoft© [Zeng et al., 2004] et bien sûr Google© [Scaiella et al., 2012], ainsi que des moteurs de recherche tels que Yippy (http ://www.yippy.com), Carrot (http ://carrot2.org), ou encore iBoogie (http ://www.iboogie.com).

Ce domaine d'étude est identifié sous le nom de *Search Results Clustering* (SRC) en Recherche d'Information. C'est précisément dans cette thématique que s'inscrivent les contributions du présent chapitre, nous les présenterons à l'image du mémoire, en nous concentrant davantage sur les nouvelles méthodologies de structuration qui relèvent véritablement de la Fouille de Données, que sur les aspects de pré-traitement des données textuelles ou d'évaluations des résultats qui relèvent plutôt du Traitement Automatique de la Langue (TAL) et de la Recherche d'Information (RI) respectivement.

4.1 Spécificités et problématiques du domaine SRC

Le clustering de résultats de recherches web (SRC) consiste à structurer un ensemble de documents (ou de snippets) X, récupérés par un moteur de recherche en réponse à une requête q, en une collection de classes thématiques $\Pi = \{\pi_1, \ldots, \pi_K\}$ réalisant une couverture de X et telle qu'à chaque classe π_k est associé un label l_k définissant un topic représentatif des éléments de la classe. Il est nécessaire de préciser de surcroît que le processus de structuration doit être efficace, de sorte qu'un ensemble de quelques centaines d'éléments puisse être structuré en une poignée de secondes seulement ; ceci afin d'offrir un service attractif à l'internaute.

Nous avons répertorié dans [Dias et al., 2011] quinze méthodes proposées pour la tâche de SRC sur les vingt dernières années. Ces méthodes se différencient selon trois propriétés principales :

- **Le type de structure qu'elles produisent :** certaines générant des structures à plat (e.g. partitions) et d'autres des structures hiérarchiques ; les secondes étant davantage propices à une exploration interactive par l'utilisateur des thèmes et sous-thèmes associés à sa requête.
- **La méthodologie de construction des classes :** selon que le processus de clustering est guidé par les documents [16], par les labels (approches dites mono-

16. Typiquement les méthodologies d'agrégation de documents par thèmes suivi de la génération des labels.

thétiques) [17], ou par les deux simultanément (approches polythétiques).

— **Le type d'entrées utilisées :** des méthodes considérant les mots ou unités polylexicales observés dans les documents entiers tandis que d'autres se fondent sur ces mêmes observations mais plutôt sur des représentations réduites que constituent les snippets.

Ces différentes expérimentations et comparaisons empiriques aboutissent à certains choix méthodologiques qui orientent les contributions à venir. Il ressort d'une part que les **structurations hiérarchiques** sont plus à même de répondre aux besoins de la tâche, d'autre part que les **unités polylexicales** portent une information utile à la caractérisation thématique des résultats d'une requête et enfin que l'usage des **snippets** plutôt que des documents entiers n'est pas préjudiciable à la qualité des résultats tout en apportant une garantie d'efficacité à la méthode.

A l'heure actuelle, deux problématiques majeures persistent et limitent l'exploitation des méthodes de SRC à grande échelle, à savoir :

1. la nécessité de présenter à l'utilisateur un ensemble réduit de topics de bonne qualité,
2. la qualité/représentativité des labels proposés.

Pour illustrer ces difficultés, nous prenons l'exemple de la requête "*interpol*" pour laquelle, l'analyse de l'ensemble des pages web retournées par n'importe quel moteur de recherche suggèrerait une structuration en deux principaux topics correspondant à l'*organisation internationale de police* d'une part et au *groupe de musique* (connu des amateurs de rock) d'autre part. Pourtant, comme illustré en Figure 21, la plupart des systèmes implémentés présentent à l'utilisateur une collection de 15 à 30 classes et dont les labels sont parfois peu indicatifs voire redondants.

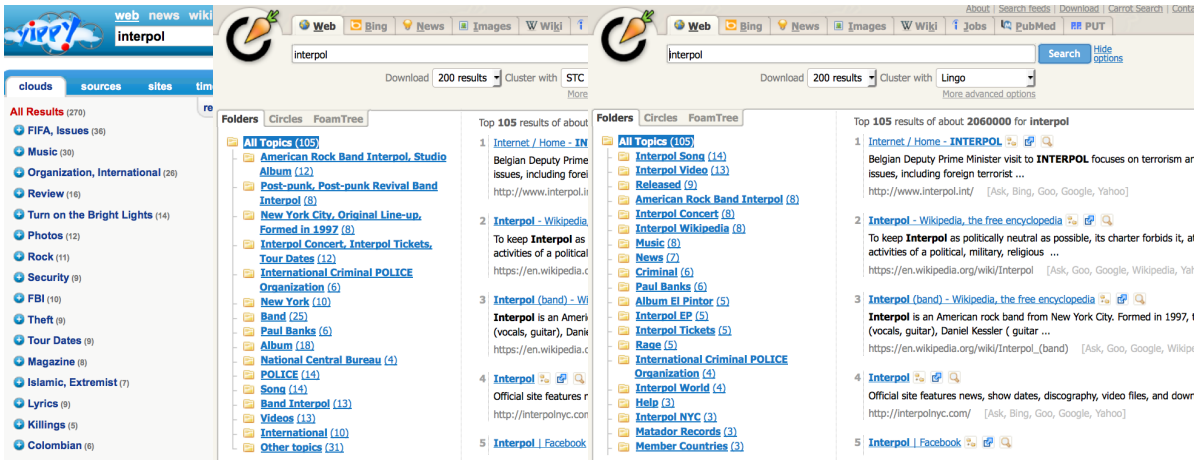


FIGURE 21 – Structuration des résultats de recherche sur la requête *interpol* par les méthodes Yippy (à gauche), STC [Zamir and Etzioni, 1998] (au centre), et LINGO [Osiński et al., 2004] (à droite).

Notre hypothèse concernant les limites observées des systèmes actuels de SRC sont que leur difficulté à identifier les bons topics est dû au fait que la manière dont ils modélisent la proximité thématique entre documents n'est pas suffisamment précise. Il s'agit en effet le plus souvent de mesures de similarité dites de premier ordre fondées sur

17. Recherche d'un ensemble de labels dont les couvertures définissent les thèmes.

l'observation des unités textuelles communes à deux documents. Pour les mêmes raisons, l'hypothèse, souvent admise par les systèmes actuels, spécifiant qu'un bon label de classe doit être composé d'unités textuelles partagées par tous les documents de la classe n'est manifestement pas pleinement satisfaisante et peut conduire à un sur-découpage tel qu'observé sur l'exemple précédent.

Les méthodes que nous présentons dans la suite de ce chapitre s'appuient sur des mesures de similarités textuelles d'ordre supérieur, plus précises, et offrant par la même occasion la possibilité de considérer des labels de classe davantage représentatifs des documents associés. Ces méthodes de clustering dédiées au SRC utilisent les snippets comme entrée des algorithmes et intègrent à la fois les critères d'homogénéité des classes et de représentativité de leur label dans un processus d'optimisation simple et intuitif.

4.2 Clustering de *snippets* et similarité d'ordre supérieur

Similarité entre mots. L'hypothèse distributionnelle de Harris [Harris, 1954] selon laquelle deux mots fréquemment observés dans des contextes similaires sont sémantiquement proches, est à la base de nombreux travaux en fouille de textes et en analyse statistique de corpus. La notion de "contexte" est relativement flexible et peut correspondre tantôt à un contexte syntaxique, tantôt à une fenêtre graphique sur un nombre limité de mots ou plus généralement au niveau de la phrase, du paragraphe ou du document entier. La qualité de la similarité sémantique est dépendante du choix de contexte utilisé, néanmoins si l'utilisation d'un contexte syntaxique (voire sémantique) pourrait sembler plus précis, leur identification nécessite une analyse parfois coûteuse des phrases qui pourrait conduire à limiter fortement la quantité d'observations et à remettre en cause le bien fondé de l'hypothèse distributionnelle, au profit de contextes (graphiques) certes plus simplistes mais générateurs d'exemples en quantité suffisante pour une analyse statistique.

Ainsi dans le contexte du SRC, on utilisera des mesures de similarité entre mots à partir de leur observation dans un ensemble de snippets X, où, pour des raisons d'efficacité, les contextes considérés seront les snippets entiers. Nous présentons dans la Table 5 quelques mesures classiques fondées sur l'hypothèse distributionnelle de Harris pour calculer la similarité sémantique entre mots. Dans les définitions de ces mesures, $p(w_i, w_j)$ désigne la probabilité d'observer les deux mots w_i et w_j ensemble dans un snippet de X qui est estimée par la proportion de snippets qui contiennent les deux mots dans X :

$$p(w_i, w_j) = \frac{f(w_i, w_j)}{|X|} = \frac{|\{x \in X | w_i \in x \wedge w_j \in x\}|}{|X|}.$$

La mesure SCP (ou *Symetric Conditional Probability*) mesure le produit des probabilités conditionnelles d'observer l'un des mot sachant que l'autre est présent dans le snippet. La mesure d'information mutuelle spécifique PMI (*Pointwise Mutual Information*) quantifie l'information apportée par une variable (un mot) sur l'autre. L'indice de $DICE$ évalue la similarité entre w_i et w_j par la proportion de snippets contenant les deux mots parmi le sous-ensemble contenant l'un des deux mots au moins. Enfin, la mesure Φ^2, moins connue, est une statistique de type χ^2 mesurant la force de l'association

Mesure	Définition
$SCP(w_i, w_j)$	$\frac{p(w_i, w_j)^2}{p(w_i).p(w_j)}$
$PMI(w_i, w_j)$	$\log_2 \frac{p(w_i, w_j)}{p(w_i).p(w_j)}$
$DICE(w_i, w_j)$	$\frac{2.f(w_i, w_j)}{f(w_i) + f(w_j)}$
$\Phi^2(w_i, w_j)$	$\frac{p(w_i, w_j) - p(w_i).p(w_j)}{p(w_i).p(w_j).(1 - p(w_i)).(1 - p(w_j))}$

TABLE 5 – Quelques mesures de similarités entre mots, fondées sur l'hypothèse distributionnelle de Harris [Harris, 1954].

entre les deux mots en tenant compte également des snippets qui ne les contiennent pas [Gale and Church, 1991].

Soit V le vocabulaire constitué des M termes (mots ou unités-polylexicales) observés dans l'ensemble de snippets X, il est aisé par les mesures présentées ci-dessus, de dériver une matrice de proximités entre mots P de taille $(M \times M)$ et symétrique.

Similarité entre snippets. Comme évoqué précédemment, la plupart des approches proposées en SRC utilisent pour quantifier la proximité thématique entre deux snippets, des mesures dites de premier ordre, c'est-à-dire utilisant les caractéristiques (mots) communes aux deux documents. Si X désigne toujours l'ensemble des snippets à structurer et V le vocabulaire de taille M extrait de X, chaque snippet x_i peut-être décrit par un vecteur binaire dans $\{0, 1\}^M$ où chaque composante binaire $x_{i,s}$ indique la présence (1) ou l'absence (0) du terme w_s dans le snippet x_i. Avec ce type de mesure, les deux snippets suivants, observés en réponse à la requête "interpol" sur Google© ne partagent aucun terme en commun autre que "the" (qui serait filtré lors de la construction du vocabulaire V), ils aboutiront à une similarité nulle quelque soit la mesure de premier ordre utilisée (similarité du cosinus, indice de Jaccard, distance Euclidienne, etc.) et n'auront aucune chance d'apparaître dans une même classe thématique.

Snippet x_1 :

> *Interpol is an American rock band from New York City. Formed in 1997, the band's original line-up consisted of Paul Banks (vocals, guitar), Daniel Kessler (guitar ...*

Snippet x_2 :

> *Interpol's profile including the latest music, albums, songs, music videos and more updates*

Il est pourtant clair que x_1 et x_2 concernent tous les deux le topic du groupe musical du fait par exemple de la présence des termes *rock, vocals, guitar* dans x_1 et de *music, albums, songs* dans x_2.

L'incapacité des méthodes actuelles à comparer finement deux documents qui n'ont aucun terme en commun est d'autant plus forte que les documents sont courts (tels les snippets). L'utilisation de mesures de similarité d'ordre supérieur, qui intègrent la proximité sémantique entre les mots observés dans les snippets, est donc indispensable pour répondre à ce type de problème.

Nous nous sommes intéressés dans un premier temps à une mesure particulière de deuxième ordre, la mesure InfoSimba (notée IS dans la suite) proposée initialement dans [Dias et al., 2007] et dont nous avons cherché à expérimenter le passage aux ordres supérieurs dans [Cleuziou and Dias, 2008]. Une forme simplifiée de la mesure IS consiste à calculer la moyenne des proximités entre les termes de chaque snippet :

$$IS(x_i, x_j) = \frac{1}{\|x_i\|_1.\|x_j\|_1} \sum_{w_s \in x_i} \sum_{w_t \in x_j} x_{i,s}.x_{j,t}.P(w_s, w_t), \tag{35}$$

où $P()$ désigne n'importe quelle mesure de proximité entre mots telles que présentées précédemment, et $x_{i,s}$ le poids du terme w_s dans le snippet x_i le cas échéant (par défaut on utilisera une représentation binaire telle que présentée initialement). On observera que, dans le cas d'une représentation binaire des snippets et moyennant une normalisation des vecteurs x_i ($\tilde{x}_{i,s} = \frac{x_{i,s}}{\sum_t x_{i,t}}$), la mesure IS se réécrit plus simplement comme un produit matriciel :

$$IS(x_i, x_j) = \tilde{x_i}^T P \tilde{x_j}. \tag{36}$$

Ce type de mesure n'a, à notre connaissance, jamais été utilisée pour comparer et structurer les résultats d'une recherche Web. Son intégration dans un processus de clustering nécessite en effet, soit d'avoir recours à un algorithme de clustering fondé sur une entrée relationnelle (matrice de similarités entre snippets), soit de redéfinir un processus de clustering spécifique à la mesure souhaitée. Les méthodes de clustering prenant en entrée une mesure de similarité entre objets (e.g. méthodes hiérarchiques, clustering à base de médoïdes, etc.) sont généralement coûteuses et ne permettent pas d'extraire efficacement un label représentatif autrement que par le choix de l'un des snippets de la classe. Notre expertise dans les méthodologies de clustering nous a amenée à proposer un algorithme de type réallocation dynamique, fondé sur la mesure InfoSimba et permettant d'extraire simultanément une collection de classes thématiquement homogènes et un ensemble de labels représentatifs associés.

Clustering de snippets dans un contexte SRC. La méthode que nous avons proposé dans [Dias et al., 2011, Moreno et al., 2013] s'appuie sur une adaptation de l'algorithme k-moyennes guidée par la mesure de similarité IS en lieu et place de la distance Euclidienne usuelle. Le processus est réalisé non plus dans l'espace $\mathbb{R}^M$, mais dans l'espace de représentation des snippets, soit $\{0,1\}^M$ et consiste à rechercher une collection $\Pi = \{\pi_1, \ldots, \pi_K\}$ de K classes et un ensemble L de K labels $l_1, \ldots, l_K$ également dans $\{0,1\}^M$ de manière à maximiser la fonction objective suivante :

$$J_{ISKM}(\Pi, L) = \sum_{k=1}^{K} \sum_{x_i \in \pi_k} IS(x_i, l_k). \tag{37}$$

Les labels l_k joueront alors le rôle des prototypes de classes, pouvant être interprétés comme de simples ensembles de mots (composantes non nulles). Ces labels pourront prendre une forme particulière en fonction des attentes dans le contexte applicatif SRC, en particulier nous imposerons en général des labels composés d'un nombre p fixé de composantes non-nulles : $l_k \in \{\{0,1\}^M | \|l_k\|_1 = p\},\ \forall k$.

Avant de chercher un algorithme d'optimisation garantissant la convergence du processus de clustering, deux remarques fondamentales s'imposent. Tout d'abord, le théorème de Huygens sur lequel repose k-moyennes n'est bien sûr plus vérifié par la mesure IS. Autrement dit, et contrairement à k-moyennes, la maximisation des similarités intra-classe (37) n'impliquera pas automatiquement la minimisation des similarités inter-classes. Enfin, la mesure IS ne satisfaisant pas non plus l'inégalité triangulaire, il n'est en théorie pas garantie que deux snippets x_i et x_j proches d'un même prototype l_k soient proches entre eux. Cependant, d'une part on peut montrer que la mesure IS satisfait une forme d'inégalité triangulaire plus faible qui accrédite la tendance du critère (37) à générer des classes de snippets similaires entre eux, et d'autre part cela nous offre la possibilité d'utiliser la similarité intra-classe (moyenne des similarités sur les paires de snippets) comme critère complémentaire afin de valider a posteriori la qualité interne d'une collection de classes de snippets.

Éléments de raisonnement

Si x_i et x_j sont tous les deux proches de l_k selon IS, cela signifie que les termes dans x_i et dans x_j sont globalement proches des termes de l_k selon P ; en calculant pour chaque mesure P proposée la densité de probabilité associée à la variable aléatoire $P_{s,t}$ on peut montrer que celle-ci est d'autant plus élevée que $P_{s,u}$ et $P_{t,u}$ sont elles-mêmes élevées ; autrement-dit, si deux termes w_t et w_s sont similaires à un même terme w_u selon P, cela augmente la probabilité pour que w_t et w_s soient également similaires entre eux. Et cette inégalité triangulaire "faible" se transpose aux similarités entre snippets par définition de la mesure IS.

L'algorithme ISKM (*InfoSimba-based k-means*) (Algorithme 5) que nous proposons pour maximiser la fonction objective (37) réalise un partitionnement de X par le procédé classique de réallocation dynamique. Dans ce procédé, l'affectation de chaque snippet au prototype le plus similaire (38) permet effectivement de maximiser la fonction objective lors de l'étape d'allocation :

$$\pi(x_i) = \arg\max_{k=1,\ldots,K} IS(x_i, l_k). \tag{38}$$

L'étape de mise à jour des prototypes est quant à elle moins triviale puisqu'il s'agit de rechercher de façon indépendante pour chaque classe π_k une combinaison l_k composée de p termes de V et maximisant l'objectif local $\sum_{x_i \in \pi_k} IS(x_i, l_k)$. Cependant, le caractère additif de la mesure IS nous évite de considérer les $\binom{M}{p}$ prototypes possibles puisqu'il suffira de retenir pour chaque classe π_k les p termes de plus fort intérêt (39) pour assurer un choix optimal de prototypes :

$$interet(w_s, \pi_k) = \sum_{x_i \in \pi_k} \frac{IS(x_i, \mathbf{e}_s)}{\|x_i\|_1}, \tag{39}$$

où $\mathbf{e}_s$ désigne le vecteur unitaire dans $\{0,1\}^M$ ayant seulement la composante s à 1.

Algorithme 5 ISKM

procedure ISKM(X, K, T, p)
  $t \leftarrow 0$
  Initialisation aléatoire des profils : $l_k^{(t)} \leftarrow Random(x_1, \ldots, x_N), \quad k = 1 \ldots K$
  Initialisation de la classification : $\forall i,\ \pi^{(t)}(x_i) = \arg\max_{k=1,\ldots,K} IS(x_i, l_k^{(t-1)})$
  repeat
    $t \leftarrow t + 1$

    for $k = 1 \ldots K$ **do** Mise à jour des profils de classes par (39)
      Calculer $interet(w_s, \pi_k)$ $(s = 1 \ldots M)$
      Retenir les p termes de plus fort intérêt $(w_{z_1}, \ldots, w_{z_p})$
      Construire le prototype optimal $l_k^{(t)} = \sum_{v=1}^{p} \mathbf{e}_{z_v}$
    end for

    for $i = 1 \ldots N$ **do** Affectation de chaque individu par (38)
      $\pi^{(t)}(x_i) = \arg\max_{k=1,\ldots,K} IS(x_i, l_k^{(t)})$
    end for

  until $\Pi^{(t)} = \Pi^{(t-1)}$ ou $t = T$ Conditions d'arrêt
  Return $\Pi^{(t)}$
end procedure

Discussion et évolution du modèle. Tout d'abord observons que l'algorithme ISKM est assuré de converger du fait de la maximisation du critère objectif à chaque étape du processus. Le partitionnement généré correspond à un optimum au moins local de la fonction objective et dépend de l'étape d'initialisation (aléatoire). Cet algorithme répond aux principales exigences du domaine d'application (SRC) en structurant efficacement un ensemble de snippets en K classes telles que :

1. les snippets dans chaque classe sont similaires entre eux (homogénéité thématique), sans exiger pour autant qu'ils partagent des éléments textuels,
2. à chaque classe est associé un label de p termes, dont tous les snippets de la classe sont proches (représentativité), sans exiger non plus que les termes du label soient présents dans les snippets de la classe (généricité).

L'algorithme ISKM constitue la brique fondamentale d'un processus de SRC plus complet présenté dans [Dias et al., 2011] avec une structuration hiérarchique de la collection de classes, puis dans [Moreno et al., 2013] en détaillant le procédé de sélection de modèle (choix du nombre de classes). En effet, afin de répondre au besoin de structuration hiérarchique des snippets en une arborescence non-nécessairement binaire, l'algorithme ISKM a été utilisé dans une stratégie plus générale de clustering hiérarchique divisif. Dans cette stratégie de structuration, l'ensemble de snippets X est récursivement partitionné par l'algorithme ISKM, dont le choix du paramètre K est réalisé de manière globale par l'étude de la fonction objective sur les partitions générées selon différentes valeurs de K.

Cette contribution a été validée par des expérimentations réalisées sur les quelques benchmarks du domaine SRC (e.g. ODP-239 [Carpineto and Romano, 2010], MORESQUE [Di Marco and Navigli, 2013]) et de manière comparative avec les méthodes

de référence telles que LINGO [Osiński et al., 2004], STC [Zamir and Etzioni, 1998] et OPTIMSRC [Carpineto and Romano, 2010]. Y compris pour des labels de petite taille (e.g. $p = 2$ ou 3), les collections de snippets proposées par notre approche utilisant des similarités d'ordre supérieur se sont révélées chaque fois de meilleur qualité [18] que les approches de l'état de l'art. À titre d'exemple, nous avons montré dans [Dias et al., 2011], que l'approche proposée permet de structurer les snippets de la requête *"interpol"* en deux classes au sommet de la structure hiérarchique, dont les labels (prototypes) proposés à l'utilisateurs sont (avec p=3) :

1. *icpo* [19]*, access, police organization*
2. *news, music, interpol music*

et correspondent parfaitement aux deux thématiques attendues et souhaitables pour aider l'utilisateur à appréhender la globalité des réponses à sa requête.

4.3 Clustering de *snippets* dans des espaces duales

Nous avons cherché par la suite à généraliser le processus SRC proposé précédemment de manière à améliorer la qualité des labels, difficiles et donc rarement évalués dans les protocoles d'expérimentation en SRC. En particulier nous avons envisagé la possibilité que ces labels ne soient plus choisis nécessairement dans l'espace de représentation des snippets mais qu'ils proviennent de sources externes susceptibles de proposer des labels dans un registre plus adapté aux attentes des utilisateurs. Typiquement, Google© propose des suggestions de termes pour compléter une requête utilisateur, à partir des requêtes fréquemment formulées sur le moteur de recherche (*query-logs*). Par exemple, la requête "*interpol*" suscite de nombreuses suggestions parmi lesquelles on retrouve {*interpol pop group, interpol music group, interpol rock band, interpol police, interpol organization crime, ...*}. Ces *query logs*, pré-enregistrés, pourraient alors servir de bibliothèque de labels pour représenter les classes de snippets en SRC.

Par exemple nous illustrons en Figure 22 l'intérêt d'utiliser des *query logs* dans un contexte SRC appliqué à la recherche d'images (décrites par le contenu textuel qui les entourent) et développé dans la thèse de J. G. Moreno [Moreno, 2014]. L'approche ISKM a été utilisée pour le développement d'une application smartphone dans laquelle l'utilisateur peut saisir une requête (e.g. *jaguar*) et accéder en réponse à une interface dans laquelle il peut faire défiler les clusters labélisés (verticalement) et les images contenues dans les clusters (horizontalement). Si la comparaison entre le processus SRC obtenu par l'algorithme ISKM et une recherche guidée par les *query logs* révèle une bonne homogénéité des classes produites par ISKM, elle révèle également un déficit d'explicitation des labels.

Afin de tirer profit des deux aspects (homogénéité des classes obtenues par un processus SRC et qualité des labels suggérés par les *query logs*), nous avons proposé dans [Moreno et al., 2014] l'algorithme *Dual C-Means* (DCM) qui, étant donné un ensemble d'objets (snippet) X décrit dans un espace E_1 induit par un vocabulaire V_1, réalise un

18. Par rapport à des classifications souhaitées (évaluation externe).
19. International Criminal Police Organization (ICPO).

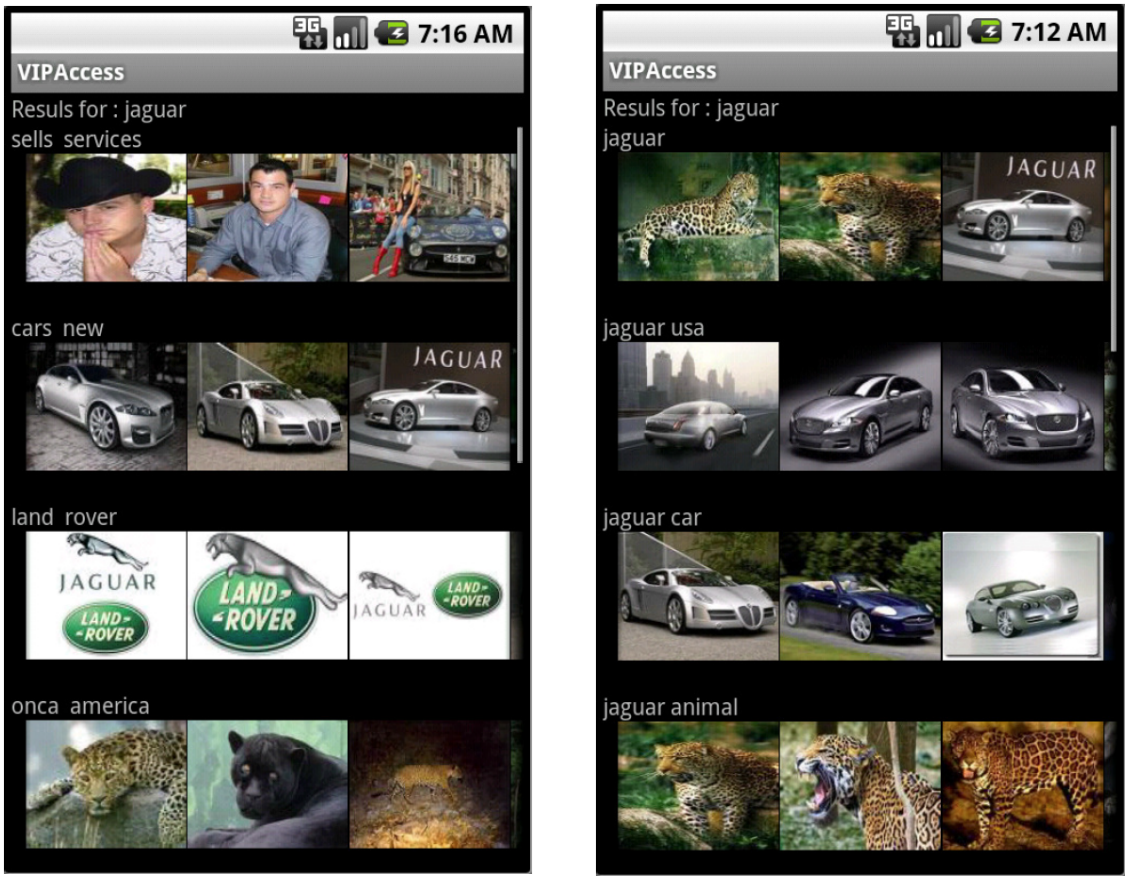


FIGURE 22 – Processus SRC développé sur smartphone : exemple des résultats de la requête *jaguar* organisés par ISKM (à gauche) et par les *query logs* (à droite).

partitionnement de X guidé par des prototypes (labels) décrits dans un espace E_2 induit par un vocabulaire éventuellement différent V_2 (illustration en Figure 23).

L'hypothèse forte sur laquelle repose cette stratégie est l'existence d'une mesure de proximité permettant de comparer tout objet de E_1 avec tout label de E_2. Dans le contexte applicatif du SRC, les snippets et les labels étant définis par deux vocabulaires V_1 et V_2 de tailles respectives M_1 et M_2, l'hypothèse précédente peut se limiter à l'existence d'une matrice de "passage" P $(M_1 \times M_2)$ mettant en relation chaque terme de V_1 avec chaque terme de V_2 et en imposant la forme suivante pour la mesure de similarité (40) :

$$S(x_i, l_k) = \tilde{x}_i^T P \tilde{l}_k, \tag{40}$$

où $\tilde{x}_i$ et $\tilde{l}_k$ désignent les formes normalisées des vecteurs binaires $x_i \in \{0,1\}^{M_1}$ et $l_k \in \{0,1\}^{M_2}$ respectivement. La forme de la mesure S reste générique et correspond à une mesure de similarité d'ordre supérieur entre éléments textuels avec la possibilité d'instancier P par toute mesure basée par exemple sur des collocations de termes telles que celles déjà évoquées précédemment (Table 5).

Dans un processus d'optimisation analogue à l'algorithme ISKM précédent, les étapes d'affectation et de mise à jour des prototypes dans DCM consistent respectivement à :

— affecter chaque snippet au prototype (label) le plus similaire (41) :

$$\pi(x_i) = \underset{k=1,\ldots,K}{\arg\max}\, S(x_i, l_k), \tag{41}$$

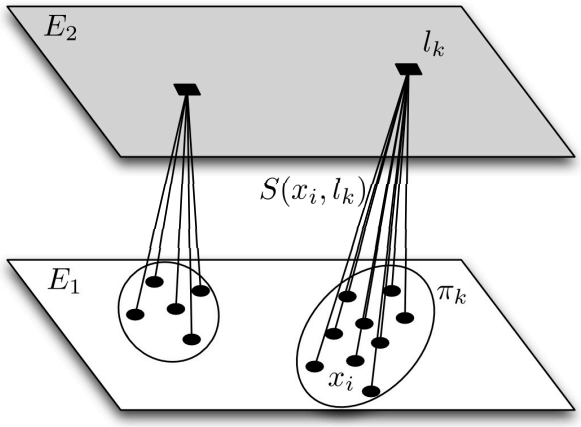


FIGURE 23 – Illustration de l'approche *Dual C-Means* : le partitionnement des objets définis sur E_1 est guidé par des prototypes définis sur un autre espace E_2.

— recherche dans l'espace des labels E_2 le meilleur prototype pour chaque cluster (42) :

$$l_k = \arg\max_{l \in E_2} \sum_{x_i \in \pi_k} S(x_i, l), \tag{42}$$

de manière à assurer, à chaque itération, la maximisation de la fonction objective suivante

$$J_{DCM}(\Pi, L) = \sum_{k=1}^{K} \sum_{x_i \in \pi_k} S(x_i, l_k). \tag{43}$$

On observera que dans le cas particulier où les labels sont définis sans restriction dans le même espace que les snippets ($E_1 = E_2$) alors la mesure de similarité S correspond à la mesure INFOSIMBA et l'algorithme DCM se ramène exactement à l'algorithme ISKM initial. En ceci, DCM constitue bien une généralisation de l'algorithme ISKM.

Nous avons cherché à évaluer globalement la qualité du processus SRC proposé par l'algorithme DCM. Pour cela un jeu de donnée spécifique WEBSRC401 [20] a été construit et se présente comme un premier benchmark offrant à la communauté la possibilité d'évaluer un processus SRC en termes de qualités des clusters et des labels simultanément. WEBSRC est composé de 5,560 snippets pré-organisés en 50 topics, correspondant chacun à une requête, et en sous-topics (entre 3 et 6 sous-topics par topic), à chaque sous-topic étant associé un *query-log*. Nous avons alors testé l'algorithme DCM paramétré avec 10 classes (K=10) dans deux configurations :

1. sans tenir compte des *query logs* dans le processus de clustering : DCM est alors exécuté avec $E_1 = E_2$ (soit la configuration ISKM), puis un *query log* est affecté a posteriori à chaque classe par la règle (42) parmi l'ensemble des *query logs* fournis dans [Sakai et al., 2013] ;
2. en utilisant les *query logs* dans le processus de clustering : DCM est exécuté dans sa configuration générique avec $E_1 \neq E_2$ et E_2 correspondant à l'ensemble des *query logs* fournis dans [Sakai et al., 2013].

20. http ://websrc401.greyc.fr/

Pour chacune des 50 requêtes/topics, la collection de classes structurant la centaine de snippets correspondant est évaluée selon deux critères : d'une part la qualité du clustering est mesurée par la mesure FBcubed quantifiant la correspondance avec l'organisation de référence en sous-topics (cf. chapitre 1), et d'autre part la qualité des labels est évaluée par une mesure de rappel indiquant le taux de labels attendus (*query logs* associés aux sous-topics) effectivement retrouvés parmi les dix labels proposés par DCM (*query logs* proposés comme labels des clusters). Enfin, plusieurs formes de matrices de proximités entre termes P ont été testées en considérant les quatre mesures de collocation déjà présentées (SCP, PMI, DICE et Φ^2). Les qualités moyennes (sur 50 requêtes) des clusters et des labels sont présentées dans les diagrammes proposés en Figure 24.

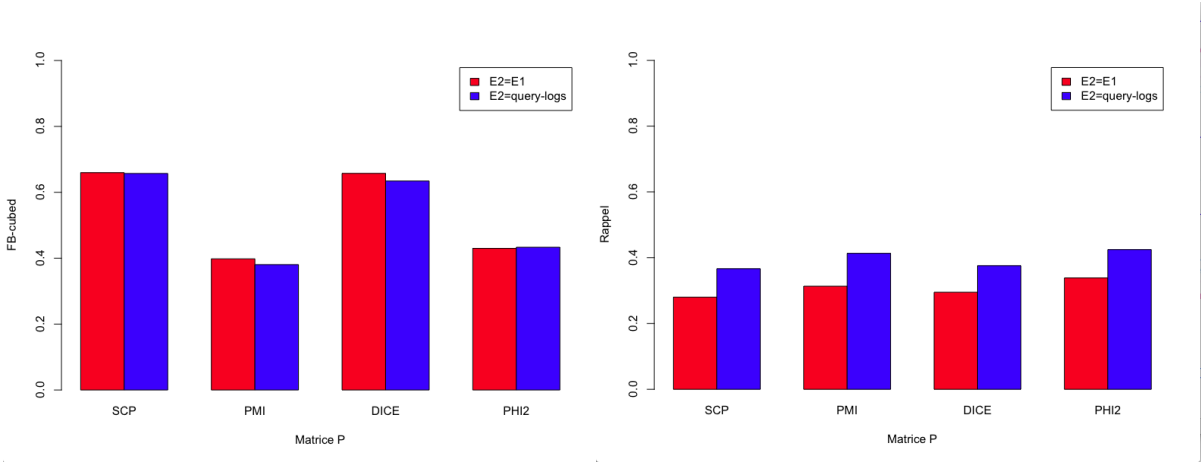


FIGURE 24 – Évaluation du processus SRC proposé par DCM sur WEBSRC401 avec ou sans dissociation des espaces de représentation des snippets (E_1) et des labels (E_2) et pour différentes matrices de similarités entre termes P : qualité des clusters (à gauche) et taux de rappel sur les labels (à droite).

Il en ressort que (1) la qualité des clusters, comme on pouvait s'y attendre, est plutôt altérée par la dissociation des espaces de représentation mais de manière non-significative (test statistique à l'appui) mais (2) en revanche, et comme espéré, la pertinence des labels est significativement améliorée par DCM lorsque celui-ci est configuré de manière à structurer les snippets "autour" des *query logs* directement. Des expérimentations plus complètes ont été présentées dans [Moreno et al., 2014], mais les résultats partiels présentés ci-dessus suffisent à illustrer d'une part l'intérêt de l'algorithme DCM, d'autre part le fait que la qualité des clusters soit très dépendante de la manière dont est construite la matrice P et enfin du chemin qu'il reste à parcourir pour la problématique SRC tant les scores de qualités, bien qu'améliorés par les contributions proposées, restent perfectibles.

La tâche de SRC constitue une problématique récente en Recherche d'Information qui a déjà donné lieu à d'intenses recherches mais qui promettent de s'intensifier dans l'avenir jusqu'à l'émergence d'une solution efficace et adoptée des internautes. Notre contribution dans ce domaine a consisté à explorer et valider d'une part l'hypothèse selon laquelle des mesures de similarités d'ordre supérieur sont mieux à même de structurer des snippets en une collection de classes thématiques homogènes et d'autre part qu'un processus de génération centré sur des formes spécifiques et disponibles de labels (e.g.

query logs) conduirait à des labels plus explicites (intensions) sans détériorer la qualité de la collection (extensions).

D'autres hypothèses devraient être envisagées, notamment sur la forme des collections de snippets générées. En particulier il est admis qu'une structuration en classes disjointes n'est pas adaptée à la problématique et qu'une organisation des snippets en classes recouvrantes, rendant accessible une page Web multi-thématique *via* plusieurs classes, conviendrait mieux [Zamir and Etzioni, 1998]. Fort des contributions proposées dans les deux premiers chapitres de ce mémoire, nous pourrions effectivement envisager une formulation "recouvrante" des algorithmes ISKM et DCM ; néanmoins la modélisation d'un recouvrement par combinaison de labels dans un espace discret de termes n'est pas anodine et suscite réflexions. Quelques modélisations ont déjà été proposées dans ce sens et offrent des perspectives d'études intéressantes.

Plus fondamentalement, nous avons observé que la matrice P, modélisant les proximités entre termes, est déterminante pour le succès d'un processus basé sur une similarité d'ordre supérieur entre éléments textuels. Jusqu'à présent nous avons considéré des matrices construites à partir de mesures reposant sur l'hypothèse distributionnelle des termes et en observant leurs co-occurrences dans des contextes de bas niveau (les snippets). Il serait assurément profitable d'un point de vue SRC de travailler à l'amélioration de la similarité entre termes, soit en proposant des contextes distributionnels plus précis, soit en ayant recours à des connaissances sémantiques externes.

Nous entendons par "contextes distributionnels plus précis", le fait par exemple d'observer deux termes dans un même contexte graphique (fenêtre de quelques mots) ou syntaxique (rôle grammatical au sein de la phrase). L'analyse syntaxique de chaque phrase n'étant pas envisageable dans un contexte SRC pour des raisons d'efficacité, des formes de représentation intermédiaires pourraient être envisagées comme par exemple des structures de dépendance. L'analyse non-supervisée de dépendances dans les textes bruts ayant fait l'objet d'un travail d'encadrement doctoral, nous consacrons l'Annexe B de ce mémoire à une description synthétique de cette étude.

Enfin et surtout, le recours à des connaissances sémantiques externes pourrait être d'une grande aide dans la construction d'une matrice de proximité P de qualité. De nombreux travaux ont par exemple proposé des mesures de proximités sémantiques entre termes en analysant les positions relatives des termes dans un thésaurus tel que WordNet [Pedersen et al., 2004]. Ces structures étant généralement construites manuellement par des linguistes, la qualité des mesures de proximités qui en découlent sont effectivement bien supérieures à toute mesure basée uniquement sur des collocations. Néanmoins ces ressources linguistiques s'avèrent rares, ce qui limite leur utilisation dans un contexte SRC à des thématiques générales et des langues richement dotées. La génération de telles ressources linguistiques de manière automatique ou semi-automatique est une autre problématique en Recherche d'Information et plus généralement en Linguistique Computationnelle qui dépasse largement le spectre du SRC et à laquelle nous nous intéressons dans le prochain et dernier chapitre du mémoire.

Publications associées au chapitre 4

[Cleuziou and Dias, 2008] Cleuziou, G. and Dias, G. (2008). Apprentissage de mesures de similarité sémantiques : étude d'une variante de la mesure infosimba. In *first joint meeting of the Société Francophone de Classification and the Classification and Data Analysis Group of the Italian Statistical Society*, pages 233–236, Caserta, Italy.

[Dias et al., 2011] Dias, G., Cleuziou, G., and Machado, D. (2011). Informative polythetic hierarchical ephemeral clustering. In *Web Intelligence and Intelligent Agent Technology (WI-IAT), 2011 IEEE/WIC/ACM International Conference*, volume 1, pages 104–111.

[Moreno et al., 2013] Moreno, J. G., Dias, G., and Cleuziou, G. (2013). Post-retrieval clustering using third-order similarity measures. In *ACL* (2), pages 153–158.

[Moreno et al., 2014] Moreno, J. G., Dias, G., and Cleuziou, G. (2014). Query log driven web search results clustering. In *Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval*, pages 777–786. ACM.

5 Acquisition de taxonomies lexicales

En codant les relations sémantiques entre termes, les Taxonomies Lexicales telles que le thésaurus WordNet [Miller et al., 1990], enrichissent le potentiel de raisonnement des applications en Recherche d'Information et en Traitement Automatique des Langues. Cependant, comme nous l'avons évoqué dans le précédant chapitre, le développement à grande échelle de telles ressources est impossible du fait des efforts que nécessite leur construction [Kozareva and Hovy, 2013]. En conséquence, de nombreuses recherches sont apparues depuis une quinzaine d'années, qui considèrent l'apprentissage automatique de taxonomies lexicales [Cimiano et al., 2009] ; c'est dans ce domaine d'étude que se situe la contribution proposée dans cet ultime chapitre.

5.1 Introduction à l'acquisition de taxonomies lexicales

Apprendre des taxonomies lexicales à partir de textes, plutôt que de les construire "à la main", présenterait d'indéniables avantages. Cela permettrait d'abord d'obtenir directement des ressources adaptées à un domaine particulier mais aussi de générer des taxonomies à large couverture étant donné le faible coût humain requis.

Les deux principales étapes dans la construction automatique de taxonomies lexicales sont : (1) l'extraction terminologique et (2) la structuration des termes. De nombreux travaux ont été menés autour de l'extraction terminologique [Navigli and Velardi, 2004, Kozareva et al., 2008] et la présente contribution se consacre exclusivement à la seconde étape, la structuration des termes, qui est généralement réalisée par l'un des quatre grands types d'approches suivants : orientée similarités, orientée patrons, utilisant les concepts formels ou enfin les approches associatives. Ainsi, dans la suite de ce chapitre on considèrera que l'ensemble E des termes à structurer en une taxonomie est connu et qu'il aura par exemple été obtenu à partir d'un corpus de textes par un algorithme d'extraction terminologique [21].

Les approches orientées similarités ou clustering procèdent par classification hiérarchique des termes [Pereira et al., 1993, Paaß et al., 2004] à partir d'une mesure de similarité portant sur leur sens et généralement fondée sur des représentations vectorielles comme les mesures de collocations déjà évoquées dans le chapitre précédant. Ces approches présentent l'avantage considérable de pouvoir identifier des relations sémantiques qui n'apparaissent pas explicitement dans les textes. Elles permettent également d'éviter le problème des chaînes inconsistantes [22] du fait du processus hiérarchique de structuration. Cependant, il est connu que ces méthodes ne peuvent garantir la génération de relations aussi précisément que les approches à base de patrons linguistiques.

Les approches orientées patrons [Kozareva and Hovy, 2013, Velardi et al., 2013] définissent des patrons lexico-syntaxiques sensés identifier des relations sémantiques entre termes. Ces patrons peuvent être construits manuellement ou générés automatiquement par bootstrapping ; ils sont réputés pour leur grande précision dans la reconnaissance d'instances de relations. Mais ces approches ont une couverture généralement limitée et

21. Le lecteur intéressé pourra trouver quelques exemples d'outils d'extraction terminologique sur http ://linguagreca.com/blog/2013/09/automatic-terminology-extraction/

22. Typiquement les cycles de relations sémantiques du genre : w_1 is a w_2, w_2 is a w_3 et w_3 is a w_1.

ne permettent d'extraire qu'un nombre réduit de relations, à plus forte raison dans des corpus spécifiques comme par exemple des domaines techniques. De plus des chaînes inconsistantes peuvent apparaître puisque les relations sont extraites par paires et de manière indépendante les unes des autres.

[Cimiano et al., 2005] utilisent l'analyse formelle de concepts (déjà évoquée en première partie) pour structurer les termes dans un treillis de concepts formels à partir de l'observation des contextes, par exemple syntaxiques, dans lesquels ils apparaissent (typiquement les verbes dont les termes composent le complément). Néanmoins ce type de structuration ne correspond pas vraiment aux spécifications d'une taxonomie lexicale censée fournir des relations sémantiques entre termes plutôt que des relations d'inclusions entre concepts formels.

Enfin, les méthodes dites associatives [Sanderson and Croft, 1999], exploitent des mesures de similarité asymétriques entre termes afin de modéliser des relations de subsomption identifiant qu'un terme w_i est plus général/spécifique qu'un autre terme w_j. Les méthodes associatives proposées jusqu'ici font une hypothèse implicite forte selon laquelle les termes généraux sont toujours plus fréquents que leurs termes plus spécifiques, ce qui n'est biensûr pas toujours vérifié en pratique.

La plupart des travaux dans ce domaine s'orientent vers l'une des approches précédentes exclusivement. Cependant quelques études récentes ce sont intéressées à l'hybridation de ces méthodologies de manière à tirer profit des avantages de chaque approche. Deux contributions importantes ont été réalisées dans cette voie [Snow et al., 2006, Yang and Callan, 2009]. Dans ces deux propositions une métrique est apprise qui modélise les relations d'hyperonymie (relations *is-a*) dans un espace de représentation construit par des critères potentiellement discriminants (e.g. contextes, co-occurrences, dépendances syntaxiques ou patrons) ; l'apprentissage est réalisé de manière supervisée (par régression) à partir de taxonomies existantes. La métrique apprise est ensuite utilisée dans un processus incrémental d'acquisition de la taxonomie, formalisé comme un problème d'optimisation mono- ou multi-objectifs. Le principal avantage dans ces approches réside dans leur capacité à intégrer plusieurs critères pour modéliser les relations *is-a* de manière à combiner des informations précises mais à faible couverture, avec d'autres d'avantage bruitées mais à fort rappel. Malheureusement ces deux propositions se placent dans un contexte d'apprentissage supervisé et nécessitent de disposer de larges taxonomies pré-établies telles que WordNdet ou ODP (*Open Directory Project*). Or ces ressources ne sont pas universellement adaptées à tout domaine et ne sont disponibles que dans un nombre limité de langues.

Dans ce chapitre, nous proposons une nouvelle stratégie multi-critères d'induction de taxonomies lexicales dans un contexte d'apprentissage semi-supervisé, voire "auto-supervisé" (nous reviendrons sur ce terme le moment venu). Cette étude repose sur la théorie de la prétopologie [Belmandt, 1993] qui offre un cadre formel adapté pour la définition d'opérateurs de propagation complexes. L'idée générale consiste à (1) apprendre, à partir d'un ensemble très partiel de relations (éventuellement déjà connues), une fonction de propagation pour la relation de subsomption entre termes par combinaison de descripteurs pertinents (e.g. issus des approches associatives ou de patrons) puis (2) déduire par l'analyse des fermés de l'espace prétopologique induit par la fonction de propagation apprise, une structuration en un DAG (*Directed Acyclic Graph*) correspondant à la taxonomie finale retournée. Plus précisément, et comme nous allons le décrire

par la suite, les deux principes de propagation et de structuration ne sont pas réalisés indépendamment mais de façon unifiée, l'apprentissage de la fonction de propagation étant guidée par l'évaluation de la structure de DAG qu'elle induit.

Nous présentons dans la section suivante quelques rappels indispensables de prétopologie avant d'illustrer l'utilité de ces concepts pour la tâche d'acquisition de taxonomies lexicales.

5.2 Rappels de prétopologie

La prétopologie est une théorie dont les fondements trouvent leurs racines au début du XXème siècle motivés par un besoin de construire une topologie peu contraignante et plus adaptée à la modélisation de phénomènes complexes. Nous reprenons ici les principales notions dans le formalisme notationel tel que proposé dans l'ouvrage de référence [Belmandt, 1993]. Nous considèrerons dans ce qui suit un ensemble fini E d'éléments (termes) à structurer.

Définition 5 *On appelle espace prétopologique, un couple (E, a) où $a()$ est une application de $\mathcal{P}(E)$ dans $\mathcal{P}(E)$ telle que :*

i) $a(\emptyset) = \emptyset$,

ii) $\forall A \in \mathcal{P}(E), A \subseteq a(A)$.

La fonction $a()$ est appelée *adhérence*, elle modélise un phénomène d'extension. L'adhérence est indissociable de de son application duale d'*intérieur*[23] qui modélise quant à elle un phénomène d'érosion mais que nous n'évoquerons plus par la suite par soucis de simplification.

Définition 6 *Soit (E, a) un espace prétopologique, K sera un fermé de E si et seulement si $a(K) = K$. De même on appelle fermeture de A (notée $F(A)$), le sous-ensemble obtenu par applications successives de $a()$ sur A jusqu'à obtention d'un point fixe.*

La Figure 25 vient illustrer le phénomène de fermeture d'un sous-ensemble dans un espace prétopologique.

La spécificité de l'adhérence en prétopologie est que, contrairement à la topologie traditionnelle, elle n'est pas contrainte à satisfaire l'idempotence, de sorte que l'extension d'un sous-ensemble A peut être réalisé en plusieurs étapes (e.g. $A \subseteq a(A) \subseteq a(a(A)) \subseteq \cdots \subseteq a^n(A)$) et non en une seule fois comme en topologie où l'adhérence de A désigne directement sa fermeture ($a(a()) = a()$).

Dans la suite nous considèrerons uniquement des espaces prétopologiques de type V, tels que l'application d'adhérence satisfait en plus la propriété d'isotonie (44) :

$$\forall A, B \subseteq E,\ A \subseteq B \Rightarrow a(A) \subseteq a(B) \tag{44}$$

23. L'intérieur d'un sous-ensemble A étant définie par le complémentaire de l'adhérence du complémentaire de A : $i(A) = E \setminus a(E \setminus A)$ également noté $a(A^c)^c$.

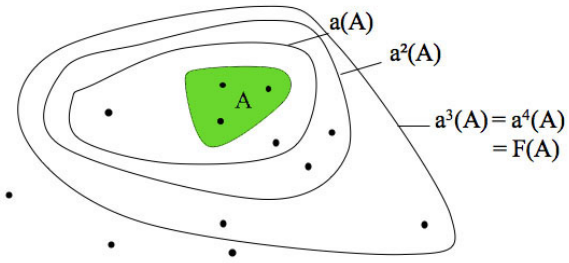


FIGURE 25 – Fermeture de A par adhérences successives en prétopologie.

Les espaces prétopologiques de type V vérifient certaines propriétés intéressantes, notamment le fait que l'intersection de fermés est un fermé. Mais avant tout, cette classe d'espace prétopologiques permet de faire un lien direct avec les notions de préfiltres et de voisinages.

Définition 7 *Un préfiltre $\mathcal{F}$ sur E est une famille de parties de E ($\mathcal{F} \subseteq \mathcal{P}(E)$), stable par passage à tout sur-ensemble :*

$$\forall F \in \mathcal{F},\ \forall H \in \mathcal{P}(E),\ F \subseteq H \Rightarrow H \in \mathcal{F}.$$

Proposition 2 *Soient $\mathcal{V} = \{\mathcal{V}(x),\ \forall x \in E\}$ une famille de préfiltres sur E tels que chaque sous-ensemble de $\mathcal{V}(x)$ contient x, et $a()$ l'application de $\mathcal{P}(E)$ dans $\mathcal{P}(E)$ définie par*

$$a(A) = \{x \in E | \forall V \in \mathcal{V}(x),\ V \cap A \neq \emptyset\},$$

alors (E, a) est un espace prétopologique de type V (engendré par $\mathcal{V}$).

La famille de préfiltres $\mathcal{V}$, telle que définie dans le Proposition 2 constitue une famille de voisinages prétopologiques sur E. Pour résumer, à partir d'une famille de voisinages prétopologiques sur E, la Proposition 2 définit une application $a()$ permettant d'engendrer un espace prétopologique de type V. De manière analogue, si on restreint $\mathcal{V}$ a correspondre à une famille de filtres [24] (ou voisinages topologiques), on peut montrer que l'espace prétopologique engendré correspond à un espace topologique (idempotence).

Avant de justifier l'intérêt de la prétopologie pour modéliser les relations de subsomption entre termes, nous terminons ces rappels en introduisant l'algorithme de structuration prétopologique proposé par [Largeron and Bonnevay, 2002]. Étant donné un espace prétopologique (E, a) de type V, cet algorithme structure les éléments de E en un DAG en considérant les inclusions entre fermés élémentaires [25]. Nous illustrons le principe de l'algorithme (en version descendante) sur la Figure 26. Sur cet exemple de 7 termes, l'espace prétopologique est structuré en 5 fermés élémentaires :

- $F(\{vehicle\}) = E$,
- $F(\{automotive\}) = F(\{car\}) = F(\{truck\}) = \{autom., car, truck, axle, wheel\}$,
- $F(\{bicycle\}) = \{bicycle, wheel\}$,

24. Un *filtre* étant un préfiltre contenant un plus petit élément.
25. Les fermés élémentaires de E sont les fermés issus des singletons de E : $\{F(\{x\}), \forall x \in E\}$.

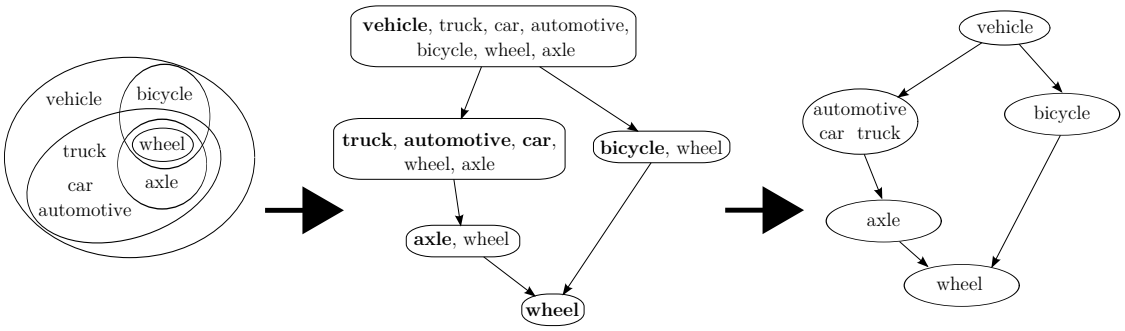


FIGURE 26 – Exemple de structuration en DAG (à droite) induite par un espace prétopologique de type V (à gauche).

— $F(\{axle\}) = \{axle, wheel\}$,
— $F(\{wheel\}) = \{wheel\}$.

La structure de DAG est générée en recherchant le ou les éléments conduisant à un fermé élémentaire maximal par inclusion, en l'occurrence le terme *vehicle* a pour fermé (élémentaire) l'ensemble E tout entier, il est maximal et unique, ainsi *vehicle* est placé, seul, au sommet du DAG. La structuration se poursuit alors récursivement sur l'ensemble $E \setminus \{vehicle\}$.

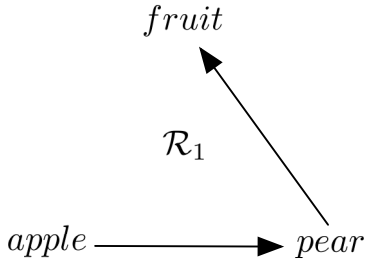
5.3 Modélisation prétopologique pour l'acquisition de taxonomies lexicales

La prétopologie n'a, à notre connaissance, jamais été utilisée pour modéliser des phénomènes de propagation dans des données textuelles. Pourtant cette théorie, et en particulier les espaces de types V, nous ont semblé particulièrement adaptés pour l'acquisition de taxonomies lexicales par une approche hybride (patrons et mesures associatives), et ceci pour plusieurs raisons déterminantes :

1. d'une part l'usage de préfiltres offre un formalisme élégant pour une **analyse multi-critères** permettant de combiner des critères linguistiques (e.g. patrons) et statistiques (e.g. mesures associatives) sur les termes, dans un contexte hybride ;
2. d'autre part, la non-indempotence de l'application d'adhérence (qui découle de l'usage des préfiltres) permet une **modélisation fine de la relation de subsomption**, qui s'avère particulièrement complexe du fait de l'ambiguité sémantique de la langue ;
3. enfin, cette théorie est dotée d'outils de structuration permettant de dériver directement d'un espace prétopologique de type V, une **structure correspondant exactement à la forme d'un thésaurus** (du type de WordNet) : ensemble de concepts (de un ou plusieurs termes), organisés hiérarchiquement par des relations de subsomption tel que chaque concept dispose d'au plus un descendant et de zéro ou plusieurs ascendants.

Nous développons les arguments précédents sur un petit exemple de trois termes $E = \{fruit, apple, pear\}$ que nous cherchons à structurer en une taxonomie par des relations de subsomption. Supposons que la recherche de patrons linguistiques (du genre

"w_i *is a* w_j") pour chaque paire dans E nous renvoie une première relation binaire $\mathcal{R}_1$ suggérant une première série d'instances de subsomption sur E :



Patrons observés :

"*pear* is a *fruit*"
"*apple* is a *pear*"

FIGURE 27 – Exemple de relation binaire sur E retournée par les patrons linguistiques.

On pourrait s'étonner d'observer la relation "*apple* is a *pear*" et qu'en revanche "*apple* is a *fruit*" n'ait pas été observé. Or, une simple recherche sur Google© nous permettrait de comprendre que les mots *apple* et *pear* peuvent jouer des rôles sémantiques complexes dans la langue. Il est par exemple fréquent de les utiliser dans des expressions telles que "*it is rather like saying an* **apple is a pear**" ou encore d'utiliser *pear* pour illustrer une forme si particulière : "Cashew **apple is a pear**-shaped fruit" ; enfin *apple* sert souvent comme l'exemple de fruit par excellence : "*my favorite* **fruit is an apple**" ou encore "the program [...] determines if the **fruit is an apple**". Ainsi la relation $\mathcal{R}_1$ capture à la fois une relation de domination (de *fruit* sur *pear*) souhaitée, mais aussi une relation dite horizontale (entre *apple* et *pear*) qui n'est pas totalement infondée bien qu'elle ne corresponde pas à une instance de subsomption.

Supposons également que de façon complémentaire, une analyse associative à la manière de [Sanderson and Croft, 1999], nous renvoie une seconde relation binaire $\mathcal{R}_2$ et donc un autre ensemble d'instances de subsomption sur E :

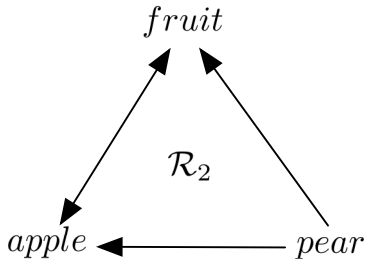


Tableau des probabilités conditionnelles estimées sur en.wikipedia.org :

$p(w_i\|w_j)$		w_j		
		fruit	*apple*	*pear*
w_i	*fruit*	1	**0.11**	**0.24**
	apple	**0.12**	1	**0.26**
	pear	0.04	0.04	1

FIGURE 28 – Exemple de relation binaire sur E retournée par une approche associative.

La relative domination de *apple* dans cette relation est dûe cette fois à la polysémie du mot qui peut, par exemple, désigner le fruit comme l'entreprise Américaine, ce qui se traduit, dans le corpus génériques, par une utilisation de *apple* plus fréquente que $fruit$, lui-même étant présent beaucoup plus massivement que *pear*.

En considérant pour chaque terme w_i de E, le sous-ensemble composé du terme w_i et des termes avec lesquels il est en relation $w_i\mathcal{R}w_j$, une base de voisinage topologique (filtre) dans $\mathcal{P}(E)$ peut être définie pour chacune des deux relations binaires $\mathcal{R}_1$ et $\mathcal{R}_2$:

— Pour la relation $\mathcal{R}_1$:
- $B_1(fruit) = \{fruit\}$,
- $B_1(apple) = \{apple, pear\}$,

- $B_1(pear) = \{pear, fruit\}$;

— Pour la relation $\mathcal{R}_2$:
- $B_2(fruit) = \{fruit, apple\}$,
- $B_2(apple) = \{apple, fruit\}$,
- $B_2(pear) = E$.

Les deux familles de filtres engendrées par B_1 et B_2 peuvent être réunies, dans un cadre multi-critères, afin de générer une famille de préfiltres $\mathcal{V} = \{\mathcal{V}(x), \forall x \in E\}$ où $\mathcal{V}(x)$ correspond au préfiltre engendré par la base $\{B_1(x), B_2(x)\}$:

$$\forall x \in E,\ \mathcal{V}(x) = B_1^S \cup B_2^S \tag{45}$$

où la notation A^S identifie A et tous ses sur-ensembles dans $\mathcal{P}(E)$.

Sachant que les espaces prétopologiques engendrés par une famille de préfiltres $\mathcal{V}$ ou par toute famille de bases de ces préfiltres sont identiques, nous ne considèrerons dans la suite que les bases de préfiltres et non les préfiltres complets.

Ainsi, la construction précédente nous conduit à la famille suivante de bases de préflitres [26] sur E :
- $\mathcal{B}(fruit) = \{\{fruit\}, \{fruit, apple\}\}$,
- $\mathcal{B}(apple) = \{\{apple, pear\}, \{apple, fruit\}\}$,
- $\mathcal{B}(pear) = \{\{pear, fruit\}, E\}$,

puis par la Proposition 2 à un espace prétopologique (E, a) de type V défini par

$$\forall A \in \mathcal{P}(E),\ a(A) = \{x \in E \mid \forall B \in \mathcal{B}(x),\ B \cap A \neq \emptyset\},$$

où $a(A)$ modélise, par combinaison des relations $\mathcal{R}_1$ et $\mathcal{R}_2$, un ensemble de termes immédiatement dominés (par subsomption) avec un élément de A. Nous détaillons en Table 6 la structure de l'espace prétopologique obtenu, en énumérant chaque partie de E et en indiquant son adhérence et sa fermeture.

$\mathcal{P}(E)$	a()	F()
$\emptyset$	$\emptyset$	$\emptyset$
$\{fruit\}$	$\{fruit, pear\}$	E
$\{apple\}$	$\{apple\}$	$\{apple\}$
$\{pear\}$	$\{pear\}$	$\{pear\}$
$\{fruit, apple\}$	E	E
$\{fruit, pear\}$	E	E
$\{apple, pear\}$	$\{apple, pear\}$	$\{apple, pear\}$
E	E	E

TABLE 6 – Structure de l'espace prétopologique issu des relations binaires $\mathcal{R}_1$ et $\mathcal{R}_2$.

La non-idempotence de l'application $a()$ se manifeste ici en observant que le terme $fruit$ propage sa domination en deux étapes, d'abord en intégrant *pear* puis ensuite *apple* :

$$\{fruit\} \subset a(\{fruit\}) = \{fruit, pear\} \subset a(a(\{fruit\})) = E.$$

26. On observera sur cet exemple que seul $\mathcal{B}(apple)$ ne contient pas de plus petit élément (unique) et n'est donc pas un filtre.

Cet exemple simplifié illustre l'intérêt de la prétopologie pour modéliser la manière complexe dont les relations sémantiques s'établissent entre les termes. En effet, si les deux approches (par patrons et associatives) s'accordent sur le fait que $fruit$ subsume *pear*, il n'y a pas de consensus concernant le rôle de *apple* vis-à-vis des deux premiers termes étant donnée l'usage qui peut-être observé de ce mot. En revanche, les deux relations étant en accord sur la relation existant entre *apple* et le sous-ensemble $\{fruit, pear\}$, la domination de $fruit$ peut alors être propagée à *apple*.

Enfin, nous présentons en Figure 29 la structure des fermés élémentaires de l'espace prétopologique construit, puis la taxonomie finale obtenue par l'algorithme de [Largeron and Bonnevay, 2002]. Il est important d'observer, qu'aucune topologie induite

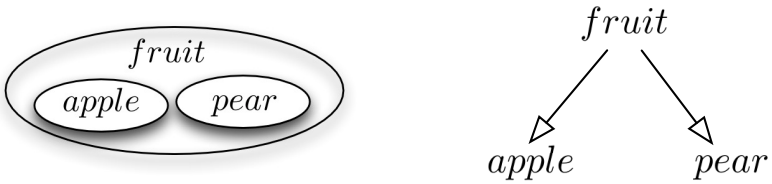


FIGURE 29 – Structure induite par les fermés élémentaires de (E, a) (à gauche) et taxonomie finale obtenue (à droite).

par l'une des deux relations initiales $\mathcal{R}_1$ ou $\mathcal{R}_2$, ni par combinaison simple de ces relations (e.g. union ou intersection) n'aurait conduit à la taxonomie finale souhaitée. En ceci, le processus que nous venons de présenter sur un exemple pourra être vu comme un nouvel "opérateur" d'agrégation et donc de structuration multi-critères.

5.4 Formalisation et discussions sur le modèle

L'utilisation de la prétopologie pour l'acquisition de taxonomies lexicales, telle qu'illustrée dans la section précédente a été présentée dans [Cleuziou et al., 2011], et se résume ainsi :

Étant donnée un famille de relations binaires réflexives $\{\mathcal{R}_1, \ldots \mathcal{R}_K\}$ sur un ensemble de termes E,

1. construire, pour chaque relation, une famille de (bases de) voisinages topologiques associés à chaque élément $x \in E$

$$\forall k,\ \forall x,\ B_k(x) = \{y \in E | x \mathcal{R}_k y\}$$

2. construire une famille $\mathcal{B}$, de (bases de) voisinages prétopologiques

$$\forall x,\ \mathcal{B}(x) = \{B_1(x) \ldots, B_K(x)\}$$

3. appliquer l'algorithme de structuration de [Largeron and Bonnevay, 2002] sur l'espace prétopologique engendré par $\mathcal{B}$ (Proposition 2).

Cette première contribution a été évaluée sur une tâche de reconstruction de taxonomies existantes. Pour cela nous avons considéré des sous-domaines du thésaurus WordNet : pour chaque taxonomie à reconstruire, nous avons (1) récupéré la liste des termes

TABLE 7 – Comparaison d'approches : associatives, orientées patrons et prétopologiques, pour la tâche de reconstruction des taxonomies *Vehicles* et *Plants* extraites de WordNet.

	Vehicles			**Plants**		
Approche	Prec.	Rec.	FMesure	Prec.	Rec.	FMesure
[Sanderson and Croft, 1999] associative	0.75	0.35	**0.48**	0.55	0.32	**0.40**
[Cleuziou et al., 2011] prétopologique	0.45	**0.36**	0.40	0.16	**0.34**	0.22
[Kozareva and Hovy, 2013] patrons	**0.79**	0.18	0.29	**0.62**	0.10	0.18

E puis (2) calculé des relations binaires sur $E \times E$ en utilisant, sur des corpus adaptés, d'une part des patrons linguistiques et d'autre part des mesures associatives seuillées avant de (3) mettre en œuvre la structuration prétopologique.

Nous avons alors pu mettre en avant la généricité de l'approche mais aussi en observer ses limites. L'approche est générique car elle permet d'instancier la plupart des méthodes existantes parmi les approches associatives ou à base de patrons et de les combiner entre elles : en effet ces dernières étant généralement fondées sur une modélisation d'une relation sémantique orientée (e.g. hyperonymie ou méronymie) aboutissant à une relation binaire, il suffit de considérer cette relation comme entrée (unique ou combinée à d'autres relations) de notre méthode de structuration prétopologique. En comparant [27] les taxonomies ainsi reconstruites avec les taxonomies attendues sur les sous-domaines *Vehicles* (108 termes) et *Plants* (554 termes) de WordNet, nous avons obtenu les résultats présentés dans la Table 7.

Ces premiers résultats illustrent bien le compromis opéré sur les approches individuelles, lorsqu'elles sont combinées via la structuration prétopologique proposée. Même si nous avons observé que ce compromis était de meilleur qualité qu'une agrégation naïve (par unions ou intersections) des approches individuelles, nous pouvons en déduire que la modélisation actuelle ne permet pas de tirer profit efficacement des informations relationnelles (topologies) contenues dans chaque entrée.

Ce résultat est dû à la définition même de l'espace prétopologique, c'est-à-dire à la manière dont il est engendré à partir d'une famille de préfiltres (Proposition 2). En effet, si la recherche d'accords entre les différentes topologies doit rester un principe fondamental en analyse multi-critères, l'application d'adhérence telle qu'elle est définie, conditionne la propagation de la subsomption à partir d'un singleton, à l'existence d'un accord parfait au commencement du processus de fermeture. Autrement dit, si toutes les relations en entrée ne s'accordent pas sur le fait qu'un élément x est en relation avec un élément y ($\forall k,\ x\mathcal{R}_k y$), alors $\{y\}$ sera un fermé et ne dominera aucun autre élément dans la structure finale. Or un tel accord parfait est non-seulement contraignant mais surtout d'autant moins envisageable que les relations en entrée sont nombreuses. Ce phénomène est clairement visible sur l'exemple proposé dans la section précédente, où l'on peut observer que la fermeture de $\{fruit\}$ ne permet d'atteindre E tout entier que parceque les deux relations sont d'accord sur la relation $pear \rightarrow fruit$, à l'origine du processus de propagation.

À partir des limites que nous venons de mettre en évidence, nous avons proposé une

27. Évaluation par F-Mesure sur les relations (paires ordonnées de termes) extraites.

nouvelle manière d'engendrer des espaces prétopologiques de type V, moins contraints, à partir d'une famille de voisinages prétopologiques, puis nous avons envisagé d'apprendre automatiquement ces nouveaux espaces par une approche semi-supervisée.

5.5 Apprentissage semi-supervisé d'espaces prétopologiques (LPS)

Généralisation. À partir d'une famille $\{\mathcal{B}(x)\}_{x\in E}$ de voisinages prétopologiques sur E, qui seraient obtenus comme précédemment par agrégation de voisinages topologiques tels que $\mathcal{B}(x) = \{B_1(x), \ldots, B_K(x)\}$, nous envisageons plusieurs manières de combiner les voisinages $B_1, \ldots, B_K$ au sein de la fonction d'adhérence. Chaque combinaison Q est une fonction logique définie sur un langage $\mathcal{Q}_K$ constitué de K fonctions propositionnelles $\{q_1, \ldots, q_K\}$ elles-mêmes définies sur $\mathcal{P}(E) \times E$ et telles que

$$q_k(A, x) \equiv (B_k(x) \cap A \neq \emptyset),$$

informant sur le fait que le voisinage de x selon le $k$ième critère intersecte ou non l'ensemble A. On notera $\mathcal{Q}_K^+$ l'ensemble des fonctions de combinaisons restreint aux DNF strictement positives, c'est-à-dire aux formules logiques en forme normale disjonctive, non-vides et sans négation.

Actuellement, par la Proposition 2, l'espace prétopologique engendré par $\mathcal{B}$ est définit par l'unique fonction d'adhérence

$$a(A) = \{x \in E | \forall B \in \mathcal{B}(x),\ B \cap A \neq \emptyset\},$$

autrement dit, dans le formalisme que venons d'introduire

$$a(A) = \{x \in E | \bigwedge_{k=1}^{K} q_k(A, x)\}.$$

Nous proposons alors de ne plus imposer une forme spécifique de combinaison (trop contraignante) mais de considérer au contraire la classe $\mathcal{C}_\mathcal{B}$ des espaces prétopologiques engendrés par $\mathcal{B}$ pour toute combinaison logique dans $\mathcal{Q}_K^+$.

Nous avons montré que les nouvelles formes de combinaisons proposées nous assurent de conserver les propriétés des espaces prétopologiques de type V, et en particulier la propriété d'isotonie (Proposition 3).

Proposition 3 *Soit, pour tout élément x de E, $\mathcal{B}(x)$ la base d'un préfiltre de parties de E qui contiennent x, obtenue par combinaison de K voisinages topologiques*[28]*.*

Soient Q une formule logique définie sur $\mathcal{Q}_K^+$ et $a()$ l'application de $\mathcal{P}(E)$ dans $\mathcal{P}(E)$ définie pour toute partie A de E par :

$$a(A) = \{x \in E \mid Q(A, x)\}$$

Alors le couple (E, a) est un espace prétopologique de type V. On dit qu'il est engendré "de manière logique" par $(\mathcal{B}, Q)$ et on le note (E, a_Q).

28. $\forall x \in E,\ \mathcal{B}(x) = \{B_1(x), \ldots, B_K(x)\}$ sont des (bases de) voisinages topologiques (filtres).

Ce résultat est fondamental puisqu'il nous assure, entre autre, de pouvoir continuer à utiliser l'algorithme de structuration prétopologique sur ces nouveaux espaces.

Exemple. Afin d'illustrer les nouvelles possibilités de structuration, nous proposons un exemple venant compléter celui utilisé dans la Section 5.3. Considérons un ensemble de quatre termes $E = \{food, fruit, apple, golden\}$ ainsi que trois relations binaire réflexives définies sur E (Figure 30).

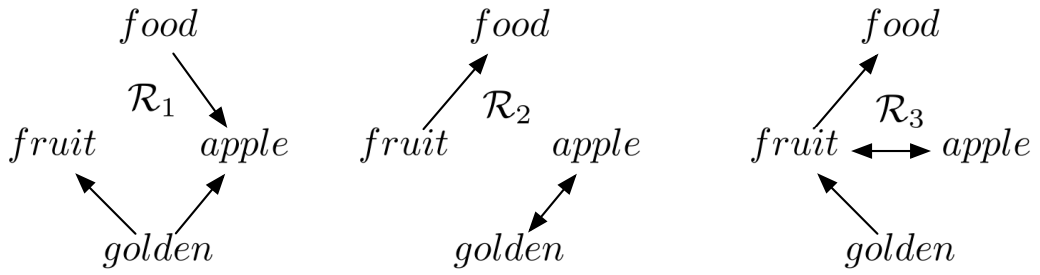


FIGURE 30 – Relations binaires définissant trois ensembles d'instances de subsomption sur E.

Chaque relation renferme une part importante de vérité mais aucune ne correspond exactement à la structure attendue, à savoir une chaîne permettant d'ordonner $food$, $fruit$,$apple$ puis $golden$ en les considérant ainsi du plus général au plus spécifique. Il en est de même, en considérant toutes les combinaisons possibles faisant intervenir tout ou partie de ces trois relations, par intersections, unions et fermetures.

Nous présentons finalement dans la Table 8 quelques unes des (18) fonctions de combinaisons possibles sur les trois relations précédentes, les fermés élémentaires obtenus sur les espaces prétopologiques engendrés et enfin les taxonomies obtenues. Nous observons alors qu'il existe une combinaison (deuxième ligne du tableau) permettant d'engendrer un espace prétopologique structurant parfaitement les termes.

On observera enfin qu'il existe une structure d'ordre sur les espaces prétopologiques induits par les fonctions logiques, si bien que pour deux fonctions logiques Q et Q' dans $\mathcal{Q}_K^+$, si Q est plus générale que Q' ($Q \succ Q'$) alors les fermés obtenus sur l'espace engendré par Q (notés $F_Q(A)$) contiennent ceux de l'espace engendré par Q' :

$$\forall Q,\ Q' \in \mathcal{Q}_K^+,\ \ Q \succ Q' \Rightarrow \forall A \subseteq E,\ F_Q(A) \supseteq F_{Q'}(A) \tag{46}$$

Il en résulte au niveau des structures de DAG, que l'ensemble des relations présentes (par fermeture transitive) dans la structure de DAG dérivée de (E, a_Q) (notée $T(E, a_Q)$) contient l'ensemble des relations observées sur $T(E, a_{Q'})$. Ce résultat s'illustre sur l'exemple de la Table 8 où

$$T(E, a_{q_1 \wedge q_2 \wedge q_3}) \subseteq T(E, a_{(q_1 \wedge q_2) \vee (q_1 \wedge q_3) \vee (q_2 \wedge q_3)}) \subseteq \left\{ \begin{array}{l} T(E, a_{q_1 \vee (q_2 \wedge q_3)}) \\ T(E, a_{q_2 \vee (q_1 \wedge q_3)}) \\ T(E, a_{q_3 \vee (q_1 \wedge q_2)}) \end{array} \right\} \subseteq T(E, a_{q_1 \vee q_2 \vee q_3})$$

Apprentissage d'espaces prétopologiques. La question essentielle qui se pose à présent est de savoir quel espace prétopologique choisir, autrement-dit comment construire la fonction logique Q, pour un problème donné ? Nous avons proposé dans

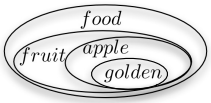
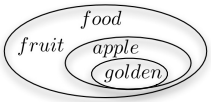
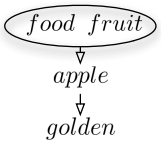
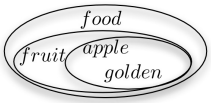
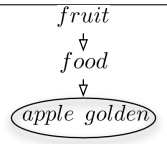
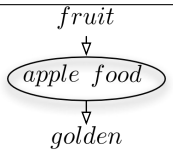
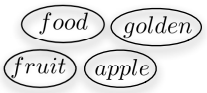
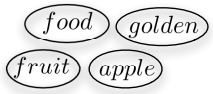
Q	$F_Q(\{x\})$	DAG $(T(E, a_Q))$
$q_1 \vee q_2 \vee q_3$		
$(q_1 \wedge q_2) \vee (q_1 \wedge q_3) \vee (q_2 \wedge q_3)$		
$q_1 \vee (q_2 \wedge q_3)$		
$q_2 \vee (q_1 \wedge q_3)$		
$q_3 \vee (q_1 \wedge q_2)$		
$q_1 \wedge q_2 \wedge q_3$		

TABLE 8 – Structurations obtenues sur l'exemple $\{food, fruit, apple, golden\}$, par les espaces prétopologiques engendrés de manière logique.

[Cleuziou and Dias, 2015] [29], un cadre général pour l'apprentissage semi-supervisé d'un espace prétopologique dans la classe $\mathcal{C}_{\mathcal{B}}$. Pour cela, l'approche LPS (*Learning Pretopological Spaces*), suppose qu'il est possible d'évaluer la qualité d'une structure $T(E, a_Q)$ par une fonction de score $f(Q)$ s'appuyant par exemple sur une connaissance partielle S de la structure attendue. À partir de cette fonction de score, l'objectif de la méthode LPS sera de rechercher un modèle (espace prétopologique) de structuration satisfaisant au mieux les informations contenues dans la connaissance partielle. La Figure 31 présente de manière schématique la stratégie d'apprentissage LPS : on y retrouve en entrée les différentes bases de filtres issues de divers critères et combinées dans une fonction logique Q ; la zone grisée matérialise le processus de structuration générateur du DAG, finalement évalué par rapport à une connaissance partielle de manière à corriger la fonction de combinaison Q ; le processus est itéré jusqu'à obtention d'une structure satisfaisante, c'est-à-dire maximisant la fonction de score $f()$.

Parcequ'il n'existe pas, *a priori*, de lien analytique directe entre la combinaison Q et son score $f(Q)$ du fait du processus complexe de propagation (directement lié à la

29. Dans cet article, la classe des espaces prétopologiques considérés est restreinte aux combinaisons Q linéaires à seuil. La formalisation proposée dans ce mémoire est plus générale puisqu'elle considère la classe des DNF positives.

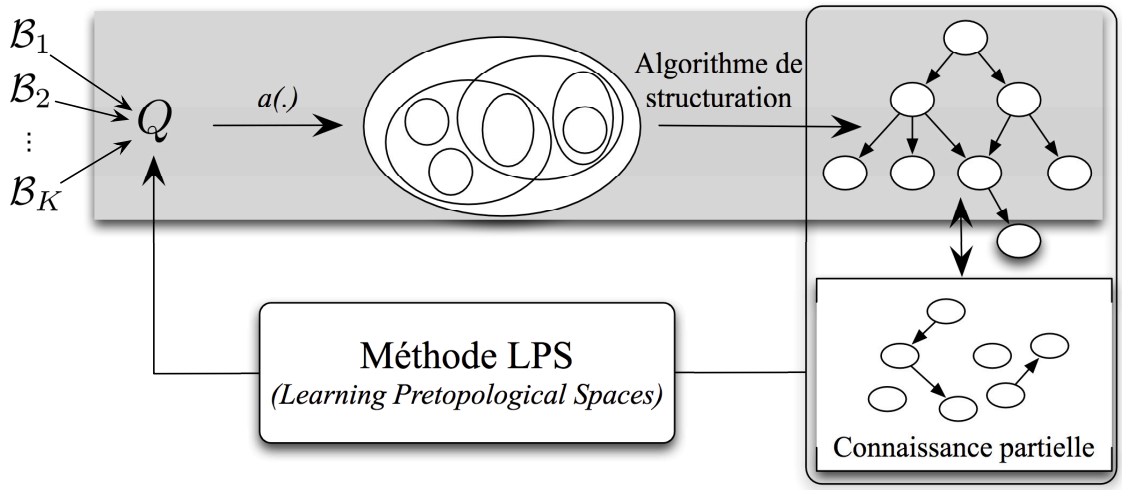


FIGURE 31 – Schéma de la stratégie d'apprentissage d'un espace prétopologique dans la classe $\mathcal{C}_{\mathcal{B}}$ par la méthode LPS.

non-idempotence de l'adhérence), il n'est pas envisageable de proposer un processus d'optimisation garantissant l'amélioration du score à mesure des itérations. L'approche LPS nécessite alors une stratégie d'optimisation stochastique *via* une heuristique ou une méta-heuristique consistant à explorer la classe d'espaces prétopologiques de manière efficace. Nous présentons dans la dernière section de ce chapitre une mise en œuvre de LPS pour la tâche d'acquisition de taxonomies lexicales.

5.6 Application de LPS à l'acquisition de taxonomies lexicales

Nous avons défini l'approche LPS de manière générique, de sorte qu'elle puisse être appliquée pour toute tâche de structuration d'un ensemble fini d'éléments en un DAG sur lequel on dispose d'une part de plusieurs critères (relations binaires) et d'autre part d'une connaissance partielle sur la structure attendue. Pour l'application qui nous intéresse particulièrement, la structuration de termes par des relations de subsomption, et pour laquelle nous avons déjà noté l'intérêt de combiner différents critères, il reste à définir un moyen d'obtenir une connaissance partielle.

Construction de la connaissance partielle. Comme évoqué au début de ce chapitre, les approches par patrons sont généralement assez précises mais ont une couverture très limitée. Ainsi ces approches sont susceptibles de fournir de manière automatique ce qui s'apparente à une connaissance partielle S même si sa fiabilité n'est pas totale. Ainsi, plutôt que de demander à un expert de proposer des exemples de relations afin de construire S (cadre semi-supervisé), nous avons suggéré d'utiliser les relations extraites automatiquement *via* les patrons en guise de connaissance partielle S, ce qui correspond à un cadre d'apprentissage que nous avons appelé "auto-supervisé" [30]. Notons que ce contexte d'utilisation de LPS est très spécifique au domaine d'application, qui permet une telle génération automatique d'exemples de relations avec une bonne fiabilité et ainsi d'envisager la structuration d'un ensemble de termes à partir uniquement de l'observation de leur

30. Apprentissage de type semi-supervisé sur des exemples extraits automatiquement.

usage dans un corpus de textes sans aucune autre information experte complémentaire. Pour d'autres domaines d'applications, une intervention experte serait *a priori* requise, mais elle resterait, de manière générale, beaucoup plus accessible que les structures complètes exigées par des approches supervisées [Snow et al., 2006, Yang and Callan, 2009].

La fonction de score $f()$ est à définir relativement au contexte applicatif. Pour le problème d'acquisition de taxonomies lexicales, nous avons construit la fonction de score (47) à partir d'un double objectif :

1. faire en sorte que la structure $T(E, a_Q)$ satisfasse au mieux la connaissance partielle S (objectif $f_1(T(E, a_Q), S)$)
2. obtenir une taxonomie $T(E, a_Q)$ proche des standards, c'est-à-dire proche d'une hiérarchie stricte (objectif $f_2(T(E, a_Q))$).

$$f(Q) = \eta.f_1(T(E, a_Q), S) + (1 - \eta).f_2(T(E, a_Q))) \quad (47)$$

Plus précisément, l'objectif f_1 est calculé par une F-Mesure évaluant les relations obtenues dans $T(E, a_Q)$ sur l'ensemble des relations et des non-relations [31] issues de la connaissance partielle S. L'objectif f_2 mesure quant à lui le nombre moyen de parents par noeuds dans $T(E, a_Q)$ et pénalise les structures dont le degré moyen d'ascendance s'écarte de 1. Le paramètre η permet d'influencer la recherche selon l'usage envisagé de la taxonomie :

— pour des usages de type raisonnement ontologique, on privilégiera des taxonomies précises même peu structurées (η proche de 1),
— si la taxonomie doit être utilisée afin de calculer par exemple des similarités entre les termes, une forte structuration est requise (e.g. η proche de 0.5).

(Meta)heuristiques d'optimisation. Différentes approches d'optimisation stochastiques peuvent être mises en œuvre et plusieurs d'entre elles ont été testées parmi lesquelles : les approches de type glouton, qui consistent à construire la fonction Q par ajout itératif de termes ; la recherche tabou qui procèdent par exploration avec mémoire des voisinages de solutions ; les approches évolutionnistes, qui génèrent itérativement des populations de solutions par croisements et mutations des meilleures fonctions Q observées à l'étape précédente.

Pour assurer une exploration plus large de l'espace des solutions, nous avons retenu dans nos expérimentations le dernier type d'approche. La recherche a donc été réalisée par un algorithme génétique dans lequel les fonctions Q sont encodées par des vecteurs binaires de longueur $p.K$ avec K le nombre de critères considérés en entrée, et p la taille maximale autorisée pour la DNF (disjonction d'au plus p monômes conjonctifs).

Expérimentations. Les expérimentations réalisées, notamment dans le cadre de la compétition internationale SEMEVAL [32] [Bordea et al., 2015, Cleuziou et al., 2015], ont consisté à reconstruire des taxonomies existantes à partir de la liste E des termes uniquement (entre 100 et 1,500 termes). Nous avons utilisé *en.wikipedia.org* comme corpus de textes afin de construire un ensemble de 10 critères (10 relations binaires) sur E :

31. La présence d'une relation $x \rightarrow y$ dans S (exemple positif de relation) donnant lieu à la génération de l'exemple négatif (ou non-relation) $y \rightarrow x$.
32. SEMEVAL-2015 task 17 :Taxonomy extraction evaluation.

TABLE 9 – Évaluation comparative de l'approche LPS pour la tâche de reconstruction des taxonomies *Vehicles* et *Plants* extraites de WordNet.

	Vehicles			**Plants**		
Approche	Prec.	Rec.	FMesure	Prec.	Rec.	FMesure
[Sanderson and Croft, 1999] associative	0.75	0.35	0.48	0.55	0.32	0.40
[Cleuziou et al., 2011] prétopologique	0.45	0.36	0.40	0.16	0.34	0.22
[Kozareva and Hovy, 2013] patrons	**0.79**	0.18	0.29	**0.62**	0.10	0.18
[Cleuziou and Dias, 2015] LPS	0.74	**0.48**	**0.58**	**0.62**	**0.40**	**0.49**

— $\mathcal{R}_1$ à $\mathcal{R}_3$, trois relations obtenues par la méthode associative de [Sanderson and Croft, 1999] avec trois seuils choisis tels que ces relations contiennent respectivement $|E|$, $2.|E|$ et $3.|E|$ relations exactement ;
— $\mathcal{R}_4$ à $\mathcal{R}_6$, trois relations obtenues en reliant chaque terme w_i de E avec ses 1, 2 et 3 plus proches ascendants respectivement (selon la probabilité conditionnelle $p(w_j|w_i)$),
— $\mathcal{R}_7$ à $\mathcal{R}_9$, trois relations obtenues en reliant chaque terme w_i de E avec ses 1, 2 et 3 plus proches descendants respectivement,
— $\mathcal{R}_{10}$ mettant en relation des termes observés dans un patron linguistique pré-défini ($\mathcal{R}_{10}$ est également utilisée en guise de connaissance partielle S).

Nous illustrons quelques résultats sur les deux taxonomies déjà évoquées : *Vehicles* et *Plants*. En complétant la précédente évaluation par les résultats obtenus à l'aide de l'approche LPS, nous observons (Table 9) un gain significatif dans la F-Mesure calculée sur les structures maximisant la fonction de score (47).

Nous proposons en Figure 32 une comparaison des différentes approches existantes (approche associative, approche par patron et LPS), sur un extrait de la taxonomie *Vehicle* composé de dix termes dont la sémantique est simple à appréhender.

Ces extraits illustrent bien d'une part la forte précision ainsi que la faible couverture des approches par patrons, d'autre part la meilleure couverture (mais au détriment de la précision) des approches associatives et finalement la manière dont l'approche LPS tire profit avantageusement de ces précédentes relations afin d'en déduire une taxonomie précise et à forte couverture.

La taxonomie obtenue par LPS sur *Vehicle* a été engendrée par la combinaison logique suivante :

$$Q(A,x) = q_{10}(A,x) \ \vee \ (q_3(A,x) \wedge q_4(A,x)) \ \vee \ (q_1(A,x) \wedge q_7(A,x))$$

qui révèle la sémantique de la fonction de propagation utilisée. Ainsi, pour qu'un ensemble de termes $A \subset E$ étende sa domination à un terme x, il faut que l'une des trois conditions suivantes soit vérifiée :

i) soit la connaissance partielle S ($=\mathcal{R}_{10}$) indique une relation de x vers l'un des termes de A ($q_{10}(A,x)$),

ii) soit les relations $\mathcal{R}_3$ et $\mathcal{R}_4$ indiquent toutes les deux une relation de x vers un des termes (éventuellement différent) de A ($q_3(A,x) \wedge q_4(A,x)$),

iii) soit les relations $\mathcal{R}_1$ et $\mathcal{R}_7$ indiquent toutes les deux une relation de x vers un des termes (éventuellement différent) de A ($q_1(A,x) \wedge q_7(A,x)$).

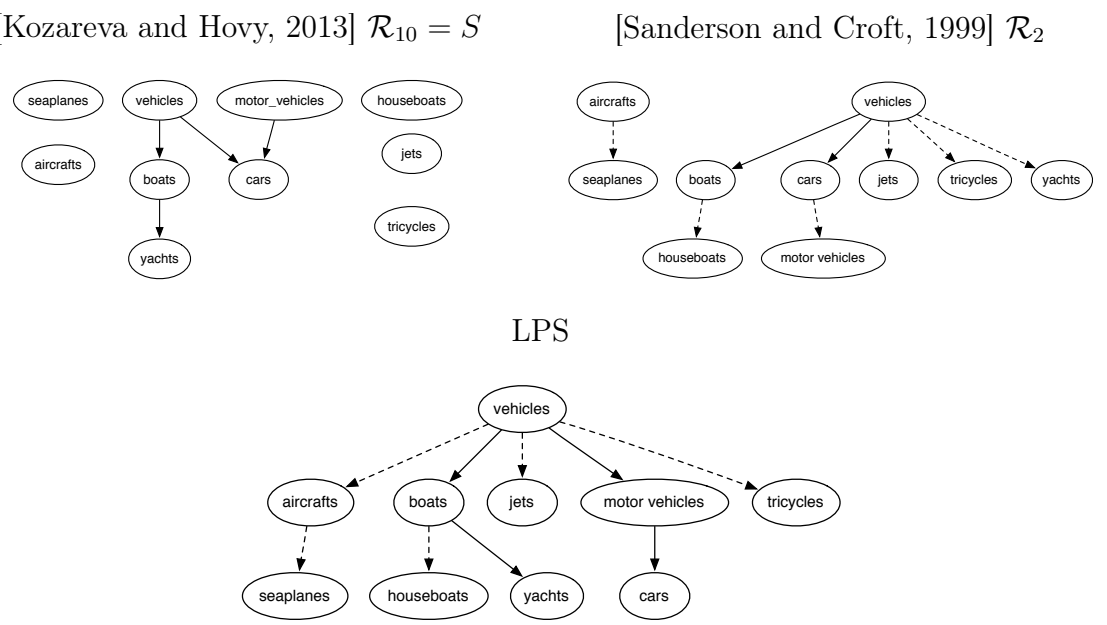


FIGURE 32 – Comparaison de méthodes sur un extrait de la taxonomie *Vehicles*.

Les différentes expérimentations menées ont pu montrer que, même si certains schémas sont récurrents, la combinaison logique Q n'est pas universelle et nécessite d'être réapprise pour chaque nouveau domaine (ou corpus) considéré. Cette dernière observation donne tout son sens à l'approche LPS qui pourra, dans ce contexte applicatif, être facilement mise en œuvre chaque fois qu'un utilisateur disposera d'un corpus et d'une liste de termes à structurer. Cependant, cela révèle également l'importance de proposer un processus d'apprentissage efficace : s'il n'est pas exigible de proposer un processus d'apprentissage en temps réel (un processus automatique, même long, sera toujours moins coûteux qu'une construction manuelle) il faut néanmoins que le processus soit en mesure de répondre en temps raisonnable à une requête de structuration sur plusieurs milliers de mots (voire plus), ce qui n'est pas possible actuellement. Une première réponse à cette objection, qui peut s'avérer déterminante à l'heure du big data, est apportée en observant que si Q nécessite effectivement d'être réapprise pour un nouveau domaine, il y a de forte raisons de penser qu'un apprentissage de Q sur un échantillon de taille raisonnable puisse suffire à faire émerger un espace prétopologique (E, a_Q) qui permettra aisément de structurer précisément *a posteriori* une collection étendue de termes du même domaine et observés sur le même corpus (par exemple dans un processus incrémental).

Dans ce dernier chapitre nous avons présenté une contribution significative pour la tâche d'acquisition automatique de taxonomies lexicales par une approche hybride (à la fois linguistique et associative). Les résultats obtenus sont très encourageants d'autant qu'ils restent largement perfectibles puisque les critères utilisés jusqu'à présent sont peu élaborés. En effet nous utilisons pour le moment des patrons simples et des mesures associatives de bas niveau quand d'autres mettent en œuvre par exemple des patrons linguistiques plus élaborés (permettant d'extraire des relations plus précises et plus nombreuses), ou des représentations sémantiques expressives telles que les espaces sémantiques continus [Mikolov et al., 2013]. L'utilisation du formalisme prétopologique

pour cette tâche nous permet d'envisager l'exploitation d'autres notions de cette théorie, ouvrant ainsi la voie à de nombreuses perspectives de recherche parmi lesquelles l'utilisation de relations valuées en utilisant les espaces préférenciés ou encore l'alignement de structures par combinaison d'espaces prétopologiques. Autant de perspectives susceptibles d'améliorer la qualité des structurations voire d'y introduire des relations sémantiques typées, multilingues voire temporelles.

Au delà même de l'application considérée, l'approche proposée et sa formalisation générique, nous incitent à explorer d'autres domaines pour lesquelles l'acquisition automatique d'une connaissance structurée par un graphe est au cœur des préoccupations. Il serait pertinent par exemple d'exporter l'approche LPS dans le domaine des réseaux sociaux ou des réseaux biologiques pour lesquels des structures en DAG peuvent être recherchées pour modéliser respectivement des structures de domination/influence sociale [Chen et al., 2010] et des sous-structures biologiques significatives [Mohamed Babou, 2012]. Dans cette démarche d'ouverture et de dépassement des limites actuelles, il s'agira d'explorer les possibilités d'acquérir dans le formalisme prétopologique des formes plus générales de graphes (avec cycles voire non-orientées) en considérant des espaces prétopologiques encore moins contraints (non-nécessairement ce type V).

Publications associées au chapitre 5

[Buscaldi et al., 2012] Buscaldi, D., Cleuziou, G., Dias, G., and Levorato, V. (2012). Exploitation de l'asymétrie entre termes pour l'extraction automatique de taxonomies à partir de textes. In *12èmes journées d'Extraction et de Gestion des Connaissances (EGC'2012)*.

[Cleuziou and Dias, 2015] Cleuziou, G. and Dias, G. (2015). Learning pretopological spaces for lexical taxonomy acquisition. In *Machine Learning and Knowledge Discovery in Databases - European Conference, ECML PKDD 2015*. Springer.

[Cleuziou et al., 2010] Cleuziou, G., Dias, G., and Levorato, V. (2010). Modélisation Prétopologique pour la Structuration Sémantico-Lexicale. In *17èmes rencontres de la Société Francophone de Classification (SFC'2010)*.

[Cleuziou et al., 2011a] Cleuziou, G., Dias, G., and Levorato, V. (2011). Acquisition de structures lexico-sémantiques à partir de textes : un nouveau cadre de travail fondé sur une structuration prétopologique. In *11èmes journées d'Extraction et de Gestion des Connaissances (EGC'2011)*.

[Cleuziou et al., 2011b] Cleuziou, G., Buscaldi, D., Levorato, V., and Dias, G. (2011). A pretopological framework for the automatic construction of lexical-semantic structures from texts. In *Proceedings of the 20th ACM international conference on Information and knowledge management*, pages 2453–2456. ACM.

[Cleuziou et al., 2015] Cleuziou, G., Buscaldi, D., Levorato, V., Dias, G., and Largeron, C. (2015). Qassit : A pretopological framework for the automatic construction of lexical taxonomies from raw texts. In *International Workshop on Semantic Evaluation (SEMEVAL 2015)*.

Conclusion

Pour terminer ce mémoire nous choisissons de revenir sur la citation introductive de Jean-Claude Clari qui tente, d'une certaine manière, de rassurer les plus sceptiques sur le sens de notre existence. "*Le monde n'a peut-être pas de sens, mais il a des structures, et tout est là.*". Nous donnerons deux interprétations, non opposables d'ailleurs, à l'affirmation finale que nous livre l'auteur : *tout est là* !

Une première explication, la plus immédiate, est de comprendre que le *sens* du monde doit être recherché dans les structures qu'il nous impose ; car les structures, elles, ont indiscutablement un sens. La seconde interprétation vient à considérer cette déclaration comme un encouragement à chercher ces structures car le monde nous offrirait toute l'information nécessaire pour les découvrir et les comprendre, *tout est là* !

Ces deux hypothèses illustrent assez bien les grandes orientations que nous choisissons comme perspectives aux études présentées dans ce mémoire. De nombreux sujets ont été identifiés comme autant de pistes à explorer, parmi lesquels le problème incontournable de la sélection de modèle dans les structurations recouvrantes (Chapitre 2), la nécessité d'adapter certaines méthodes proposées à la structuration de concepts de plus haut niveau *via* l'analyse de données symboliques (Section 2.4) ou encore le bénéfice que l'on peut attendre de l'intégration de connaissances extérieures dans la structuration thématiques des informations (Chapitre 4). Mais puisque la structuration de notre existence doit passer par l'ordonnancement de nos priorités, nous en choisirons deux.

Partant des hypothèses que les structures doivent avoir un sens et que toute l'information est disponible pour rechercher ces structures, nous nous intéresserons tout d'abord à exploiter la relation qui doit exister entre une structure générée et la réalité de la sémantique associée. Dans la continuité des travaux menés sur la construction de classifications conceptuelles (Section 2.6) d'une part, et la classification multi-vues (Annexe A) d'autre part, nous imaginerons des méthodologies originales visant à exploiter davantage les informations disponibles sur les objets du monde. Typiquement, nous proposerons de mixer des informations numériques et symboliques, non pas par uniformisation de leur traitement, mais dans un processus de fertilisation croisée dans lequel les "vues" numériques et symboliques se corrigent mutuellement de manière à converger vers une structuration valide, à la fois numériquement et sémantiquement.

Enfin nous avons apporté (Chapitre 5), les premiers gages de réussite quant à la possibilité d'apprendre automatiquement un modèle pour structurer la sémantique du langage, par le biais des termes qui l'instancient. Pour y parvenir nous avons fait l'hypothèse qu'il existait une sémantique à cette structure, dans une forme mathématique appropriée *via* la définition d'un espace prétopologique. Or nous avons dores et déjà observé que d'autres informations méritaient d'être intégrées, ce qui laisse entrevoir un potentiel d'amélioration significatif. Mais il s'agira également d'élargir le champ des modèles possibles afin de rechercher une sémantique encore plus précise et des modèles de structuration plus en adéquation avec les réalités du monde. Enfin, parce que *toute l'information est là*, mais souvent inaccessible pour des non-initiés, nous collaborerons avec des spécialistes d'autres domaines (biologie, réseaux sociaux) afin d'imaginer l'adaptation de ce méta-modèle pour apprendre à structurer l'objet de leurs études...

... tout est là !

Annexes

A Classification non-supervisée multi-vues

Cette étude a été réalisée dans le cadre des travaux de doctorat de Jacques-Henri Sublemontier [Sublemontier, 2012], co-encadrés avec Lionel Martin et Christel Vrain, respectivement maître de conférences et professeur au Laboratoire d'Informatique Fondamentale d'Orléans (LIFO).

Contexte de l'étude. La classification non-supervisée multi-vue (ou *multi-view clustering*) consiste à générer une unique partition Π consensuelle, d'un ensemble d'objets X défini non plus par un seul tableau de description $(N \times M)$ mais par plusieurs ensembles de descripteurs donnant lieu à plusieurs tableaux $X^{(1)}, \ldots, X^{(R)}$ de tailles $(N \times M_r)$. Le consensus recherché doit satisfaire les différentes représentations, de sorte que la partition Π soit de bonne qualité (à défaut d'être optimale) dans chaque vue $X^{(r)}$.

La résolution de cette problématique passe par la fusion des informations provenant des différentes vues. Trois stratégies sont possibles :

- la fusion *a priori* qui consiste à concaténer les descripteurs avant de réaliser le processus de clustering (construction d'un seul tableau ou combinaison de mesures de proximités),
- la fusion *a posteriori* qui nécessite de réaliser R classifications indépendantes avant de rechercher ou de reconstruire une classification consensuelle,
- la fusion centralisée pour laquelle les méthodes de clustering sont repensées de manière à combiner les différentes vues au sein-même du processus de construction des classes.

Les contributions réalisées dans cette thèse ont porté sur deux nouvelles méthodes de classification multi-vues par stratégie centralisée : la première méthode CoFKM est une adaptation de l'algorithme k-moyennes flou, modifié de manière à minimiser le "désaccord" entre les vues ; la seconde méthode CoBoC réalise un clustering semi-supervisé dans chaque vue avec échange de contraintes dynamiques entre les vues.

La méthode CoFKM [Cleuziou et al., 2009] procède par réallocations dynamiques (floues) des objets dans chaque vue, autour de centres mobiles également définis dans chaque vue. Le nouveau critère objectif proposé assure la réallocation et la mise à jour des centres de classe de manière concertée et intuitive entre les vues :

$$J_{CoFKM}(\{U^{(r)}\}, \{C^{(r)}\}) = \sum_{r=1}^{R} \sum_{k=1}^{K} \sum_{x_i \in X} {u_{i,k}^{(r)}}^{\beta} \|x_i^{(r)} - c_k^{(r)}\|^2 + \eta.\Delta(\{U^{(r)}\}, \{C^{(r)}\}) \quad (48)$$

On retrouve, dans le premier terme de la fonction objective, une somme (indépendante) des inerties floues pour chaque vue, à laquelle est ajouté un terme de pénalité Δ estimant le désaccord entre les partitions floues. La forme du terme de pénalité permet l'optimisation de la fonction objective à chaque étape avec une mise à jour intuitive des degrés d'appartenance et des centres de classes par le biais de moyennes pondérées mais dont les poids font intervenir l'ensemble des vues. La partition Π est finalement obtenue

par fusion (moyenne géométrique) des différentes partitions floues en faible désaccord, et affectation par la règle du plus fort degré d'appartenance.

La méthode CoFKM est générique car elle permet de réaliser un continuum entre les trois stratégies de fusions en fonction du paramètre η : fusion *a priori* ($\eta = \frac{R-1}{R}$), fusion *a posteriori* ($\eta = 0$) ou fusion centralisée ($\eta \in]0, \frac{R-1}{R}[$). En revanche CoFKM contraint chaque vue à une même forme de clustering (même nombre K de classes dans chaque vue et même critère objectif).

La méthode CoBoC [Sublemontier, 2013] offre un cadre d'apprentissage permettant de dépasser les contraintes précédentes. Dans cette méthode, les algorithmes de clustering $\mathcal{A}^{(1)}, \ldots, \mathcal{A}^{(R)}$ utilisés dans chaque vue sont considérés comme des "boîtes noires", ils peuvent donc être adaptés (type d'algorithme et/ou paramétrage) à chaque ensemble de descripteurs. À partir de l'hypothèse que chaque algorithme $\mathcal{A}^{(r)}$ tend à associer dans une même classe des objets similaires, la méthode CoBoC consiste à apprendre (par boosting) un espace de projection $X^{(r)'}$ propre à chaque vue permettant à chaque algorithme $\mathcal{A}^{(r)}$ de satisfaire au mieux des contraintes (*must-link* et *cannot-link*) favorisant le consensus entre les vues. La figure ci-dessous illustre schématiquement le processus d'apprentissage opéré par la méthode.

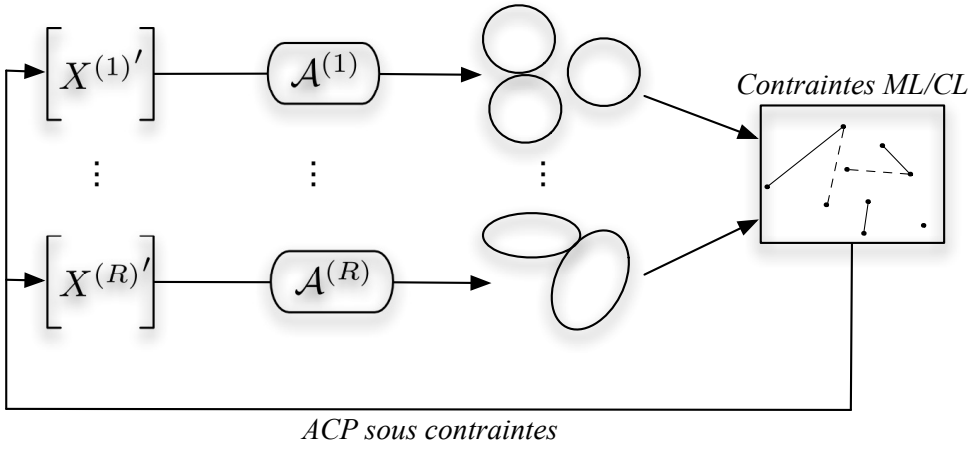


FIGURE 33 – Illustration de la méthode CoBoC : processus itératif d'apprentissage d'espaces de projection favorisant la convergence des classifications de chaque vue.

Les méthodes proposées sont particulièrement adaptées dans des situations pratiques où les vues sont physiquement distribuées et telles qu'il n'est pas possible (pour des raisons de volumétrie ou de confidentialité) de réunir toutes les données sur un même disque physique. En revanche une limite de ces approches est qu'elles se fondent sur l'hypothèse que toutes les vues sont pertinentes et qu'elles partagent des informations complémentaires. Une perspective de recherche intéressante consisterait à étudier la possibilité d'intégrer dans ces approches une notion de confiance (fournie, estimée voire apprise automatiquement) sur chaque vue.

B Apprentissage non-supervisé de structures de dépendances

Cette étude a été réalisée dans le cadre des travaux de doctorat de Marie Arcadias [Arcadias, 2015], co-encadrés avec Edmond Lassalle, Ingénieur expert à Orange Labs et Christel Vrain, professeur au Laboratoire d'Informatique Fondamentale d'Orléans (LIFO).

Contexte de l'étude. Une structure de dépendances modélise les relations de domination syntaxique entre les mots au sein de la phrase. L'arbre résultant est une représentation intermédiaire entre un arbre syntaxique, issu d'une analyse syntaxique complète, et un simple étiquetage grammatical des mots de la phrase. Cette représentation s'avère utile et suffisante pour de nombreuses tâches en recherche et en extraction d'information. Par exemple, la structure de dépendances de la phrase "*Jean-Claude Clari est un romancier québécois*" (figure ci-dessous) facilitera l'extraction d'informations prédicatives du type ORIGINE(*Jean-Claude Clari, Québec*) ou encore PROFESSION(*Jean-Claude Clari, romancier*) et par la même, l'enrichissement de bases de connaissances.

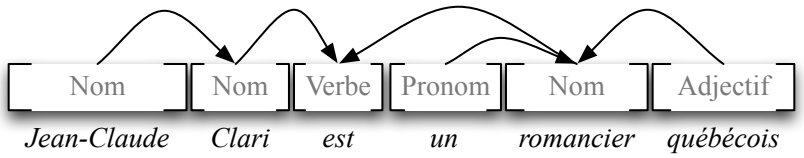


FIGURE 34 – Structure de dépendances pour la phrase "*Jean-Claude Clari est un romancier québécois*" révélant entre autres que le mot *québécois* est dominé par le mot *romancier*.

L'apprentissage non-supervisé de telles représentations consiste à apprendre un modèle de structuration à partir de corpus de phrases uniquement, de sorte que le modèle s'adapte aux séquences observées et qu'il soit capable de les analyser (rendre un arbre en sortie). L'absence d'exemples de bonnes structures, pour l'apprentissage, limite de fait la qualité des analyses que l'on peut en espérer. Mais cette légèreté rend ces méthodes presque universelles, du moment que l'on dispose d'un corpus de la langue (voire du domaine) dont on souhaite modéliser la structure syntaxique des phrases.

Les approches existantes reposent sur des modèles génératifs ou des grammaires formelles utilisant généralement le lexique et les étiquettes grammaticales des séquences de mots. Les paramètres ou règles de structuration sont alors très nombreux ; ils requièrent des temps d'apprentissage et d'analyse importants et rendent les modèles sensibles à l'initialisation et donc peu universels.

Nous avons proposé [Arcadias et al., 2014] une nouvelle famille de grammaires probabilistes hors contexte : *simples* du point de vue du nombre de règles et de l'initialisation et *efficaces* au regard des résultats obtenus.

Modèles proposés. La famille de grammaires DGdg repose sur des principes théorisés de *domination immédiate* et de *précédance linéaire* postulants qu'un même nœud de la structure peut dominer une fratrie, au sein de laquelle l'ordre peut être contraint. Les

grammaires proposées se définissent essentiellement à partir de quatre symboles non-terminaux (D, G, d et g) qui modélisent un découpage binaire récursif d'une séquence d'étiquettes grammaticales en une partie gauche (dominante G ou dominée g) et une partie droite (dominante D ou dominée d). La figure ci-dessous illustre une analyse possible de la phrase précédente avec une grammaire DGdg, et qui se traduit en l'arbre de dépendance initial attendu.

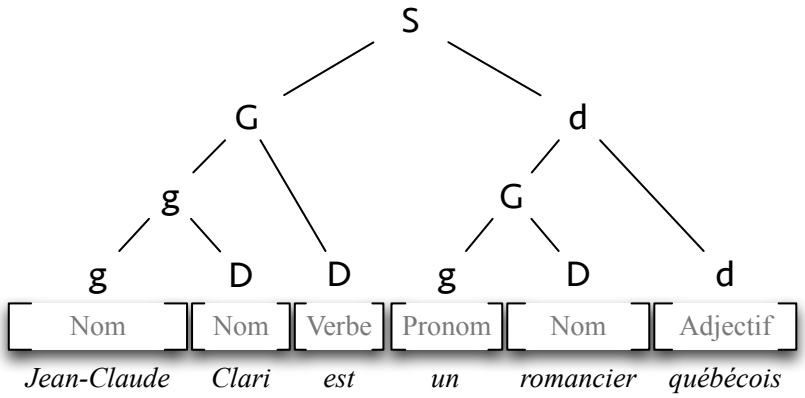


FIGURE 35 – Analyse de la phrase "*Jean-Claude Clari est un romancier québécois*" par une grammaire de type DGdg révélant entre autres que le mot *québécois* est dominé, par la gauche, par la séquence *un romancier*.

Apprentissage et analyse. Les grammaires proposées sont composées d'un ensemble réduit de règles de dérivation (sous forme normale de Chomsky) :

- d'un symbole non-terminal vers des symboles non-terminaux, du type $G \rightarrow gD$, $d \rightarrow Gd$, etc.
- d'un symbole non-terminal vers un symbole terminal, du type $D \rightarrow Verbe$, $g \rightarrow Pronom$, etc.

L'apprentissage de ces grammaires, réalisé par l'algorithme *Inside-Outside*, revient à apprendre les probabilités de chacune des règles de dérivation par maximisation de la vraisemblance du modèle étant données les observations d'entraînement (séquences d'étiquettes grammaticales). Enfin, étant donnée une grammaire probabiliste apprise sur un corpus de phrases d'entraînement, l'analyse peut être réalisée par l'algorithme CYK dans sa version probabiliste, qui recherche l'arbre d'analyse le plus probable pour une nouvelle séquence d'étiquettes donnée.

Les modèles proposés ont été testés sur une dizaine de langues (latines, germaniques ou orientales) et de corpus différents. Ils ont démontré des qualités de structuration meilleures que les approches de l'état de l'art pour la moitié de ces langues et presque systématiquement meilleures que des approches naïves d'attachement à gauche ou à droite (pour neuf des dix langues).

Ces travaux devront être poursuivis afin d'étudier, d'une part le choix optimal de grammaire dans la famille de modèles proposée, et d'autre part le potentiel des structures générées dans une application d'extraction d'information. Une expérience préliminaire encourageante d'extraction de relations prédicatives à l'aide de CRF sur les arbres de dépendances, a été réalisée.

Références

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

[Arcadias, 2015] Arcadias, M. (2015). *Apprentissage non-supervisé de dépendances à partir de textes*. PhD thesis, Université d'Orléans.

[Arcadias et al., 2014] Arcadias, M., Cleuziou, G., Lassalle, E., and Vrain, C. (2014). Apprentissage non supervisé de dépendances syntaxiques à partir de texte étiqueté, plusieurs variantes de pcfg légères. In *Extraction et Gestion des Connaissances (EGC)*, number E. 26, pages 155–160. Hermann.

[Bandyopadhyay and Saha, 2007] Bandyopadhyay, S. and Saha, S. (2007). Gaps : A clustering method using a new point symmetry-based distance measure. *Pattern Recognition*, 40(12) :3430–3451.

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

[Becker et al., 2012] Becker, E., Robisson, B., Chapple, C. E., Guénoche, A., and Brun, C. (2012). Multifunctional proteins revealed by overlapping clustering in protein interaction network. *Bioinformatics*, 28(1) :84–90.

[Belmandt, 1993] Belmandt, Z. (1993). Manuel de prétopologie et ses applications. *Hermes, Paris*, 472.

[Ben N'cir et al., 2014] Ben N'cir, C., Cleuziou, G., and Essoussi, N. (2014). Generalization of c-means for identifying non-disjoint clusters with overlap regulation. *Pattern Recognition Letters*, 45 :92–98.

[Ben N'Cir and Essoussi, 2013] Ben N'Cir, C. and Essoussi, N. (2013). Non-disjoint cluster analysis with non-uniform density. In Prasath, R. and Kathirvalavakumar, T., editors, *Mining Intelligence and Knowledge Exploration, First International Conference, MIKE 2013, Tamil Nadu, India, December 18-20, 2013*, volume 8284 of *Lecture Notes in Computer Science*, pages 100–111. Springer.

[Ben N'Cir et al., 2014] Ben N'Cir, C.-E., Cleuziou, G., and Essoussi, N. (2014). Generalization of c-means for identifying non-disjoint clusters with overlap regulation. *Pattern Recognition Letters*, 45 :92–98.

[Ben N'Cir and Essoussi, 2012] Ben N'Cir, C.-E. and Essoussi, N. (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 and Janowitz, 2003] Bertrand, P. and Janowitz, M. F. (2003). The k-weak hierarchical representations : An extension of the indexed closed weak hierarchies. *Discrete Applied Mathematics*, 127(2) :199–220.

[Billard and Diday, 2007] Billard, L. and Diday, E. (2007). *Symbolic Data Analysis : Conceptual Statistics and Data Mining (Wiley Series in Computational Statistics)*. John Wiley & Sons.

[Bock, 2003] Bock, H.-H. (2003). Clustering algorithms and kohonen maps for symbolic data (symbolic data analysis). *Journal of the Japanese Society of Computational Statistics*, 15(2) :217–229.

[Bordea et al., 2015] Bordea, G., Buitelaar, P., Faralli, S., and Navigli, R. (2015). Semeval-2015 task 17 : Taxonomy extraction evaluation (texeval). *Science*, 452(465) :429–441.

[Carpineto and Romano, 2010] Carpineto, C. and Romano, G. (2010). Optimal meta search results clustering. In *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval*, pages 170–177. ACM.

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

[Chavent et al., 2003] Chavent, M., de A.T. De Carvalho, F., Levhevallier, Y., and Verde, R. (2003). Trois nouvelles méthodes de classification automatique de données symboliques de type intervalle. *Revue de Statistique Appliquée*, 51(4) :5–29.

[Chen et al., 2010] Chen, W., Yuan, Y., and Zhang, L. (2010). Scalable influence maximization in social networks under the linear threshold model. In *Data Mining (ICDM), 2010 IEEE 10th International Conference on*, pages 88–97. IEEE.

[Chou et al., 2002] Chou, C.-H., Su, M.-C., and Lai, E. (2002). Symmetry as a new measure for cluster validity. In *2nd WSEAS Int. Conf. on Scientific Computation and Soft Computing*, pages 209–213.

[Chung and Lin, 2007] Chung, K.-L. and Lin, J.-S. (2007). Faster and more robust point symmetry-based k-means algorithm. *Pattern Recognition*, 40(2) :410–422.

[Cimiano et al., 2005] Cimiano, P., Hotho, A., and Staab, S. (2005). Learning concept hierarchies from text corpora using formal concept anaylsis. *Journal of Artificial Intelligence Research*, 24 :305–339.

[Cimiano et al., 2009] Cimiano, P., Mädche, A., Staab, S., and Völker, J. (2009). Ontology learning. In *Handbook of Ontologies*, pages 245–267. Springer Verlag.

[Cleuziou, 2006] Cleuziou, G. (2006). Classification avec recouvrement des classes : une extension des k-moyennes. In *13èmes rencontres de la Société Francophone de Classification*, pages 68–72, Paris.

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

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

[Cleuziou, 2009a] Cleuziou, G. (2009a). Okmed et wokm : deux variantes de okm pour la classification recouvrante. In *9èmes Journées Francophones d'Extraction et de Gestion des Connaissances*, volume 2. Revue des Nouvelles Technologies de l'Information, Cépaduès-Edition.

[Cleuziou, 2009b] Cleuziou, G. (2009b). Two variants of the OKM for overlapping clustering. In Guillet, F., Ritschard, G., Zighed, D. A., and Briand, H., editors, *EGC (best of volume)*, volume 292 of *Studies in Computational Intelligence*, pages 149–166. Springer.

[Cleuziou, 2013] Cleuziou, G. (2013). Osom : A method for building overlapping topological maps. *Pattern Recognition Letters*, 34(3) :239–246.

[Cleuziou, 2014] Cleuziou, G. (2014). Passage aux noyaux en classification recouvrante. In *14èmes Journées Francophones Extraction et Gestion des Connaissances, EGC 2014, Rennes, France, 28-32 Janvier, 2014*, pages 209–220.

[Cleuziou et al., 2011] Cleuziou, G., Buscaldi, D., Levorato, V., and Dias, G. (2011). A pretopological framework for the automatic construction of lexical-semantic structures from texts. In *Proceedings of the 20th ACM international conference on Information and knowledge management*, pages 2453–2456. ACM.

[Cleuziou et al., 2015] Cleuziou, G., Buscaldi, D., Levorato, V., Dias, G., and Largeron, C. (2015). Qassit : A pretopological framework for the automatic construction of lexical taxonomies from raw texts. In *International Workshop on Semantic Evaluation (SEMEVAL 2015)*.

[Cleuziou and Crémilleux, 2015] Cleuziou, G. and Crémilleux, B. (2015). Vers une classification conceptuelle recouvrante. In *22èmes rencontres de la Société Francophone de Classification*, Nantes.

[Cleuziou and de Carvalho, 2014] Cleuziou, G. and de Carvalho, F. (2014). Robustesse en classification recouvrante :
une approche par *trimming*. In *21èmes rencontres de la Société Francophone de Classification*, Rabbat, Maroc.

[Cleuziou et al., 2012] Cleuziou, G., de Carvalho, F., and Rousseau, L. (2012). OKM-L_1 et comparaison de recouvrements. In *19èmes rencontres de la Société Francophone de Classification*, Marseille.

[Cleuziou and Dias, 2008] Cleuziou, G. and Dias, G. (2008). Apprentissage de mesures de similarité sémantiques : étude d'une variante de la mesure infosimba. In *first joint meeting of the Société Francophone de Classification and the Classification and Data Analysis Group of the Italian Statistical Society*, pages 233–236, Caserta, Italy.

[Cleuziou and Dias, 2015] Cleuziou, G. and Dias, G. (2015). Learning pretopological spaces for lexical taxonomy acquisition. In *Machine Learning and Knowledge Discovery in Databases - European Conference, ECML PKDD 2015*. Springer.

[Cleuziou et al., 2009] Cleuziou, G., Exbrayat, M., Martin, L., and Sublemontier, J.-H. (2009). Cofkm : A centralized method for multiple-view clustering. In *Data Mining, 2009. ICDM'09. Ninth IEEE International Conference on*, pages 752–757. IEEE.

[Cleuziou et al., 2004] Cleuziou, G., Martin, L., and Vrain, C. (2004). PoBOC : an Overlapping Clustering Algorithm. Application to Rule-Based Classification and Textual Data. In R. López de Mántaras and L. Saitta, IOS Press, editor, *Proceedings of the 16th European Conf. on Artificial Intelligence*, pages 440–444, Valencia, Spain.

[Cleuziou and Moreno, 2015] Cleuziou, G. and Moreno, J. G. (2015). Kernel Methods for Point Symmetry-based Clustering. *Pattern Recognition*, 48 :2812–2830. (to appear).

[de Souto et al., 2012] de Souto, M. C., Coelho, A. L., Faceli, K., Sakata, T. C., Bonadia, V., and Costa, I. G. (2012). A comparison of external clustering evaluation indices in the context of imbalanced data sets. In *Neural Networks (SBRN), 2012 Brazilian Symposium on*, pages 49–54. IEEE.

[Dempster et al., 1977] Dempster, A., Laird, N., and Rubin, D. (1977). Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of Royal Statistical Society B*, 39 :1–38.

[Depril et al., 2008] Depril, D., Van Mechelen, I., and Mirkin, B. (2008). Algorithms for additive clustering of rectangular data tables. *Computational Statistics and Data Analysis*, 52(11) :4923–4938.

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

[Di Marco and Navigli, 2013] Di Marco, A. and Navigli, R. (2013). Clustering and diversifying web search results with graph-based word sense induction. *Computational Linguistics*, 39(3) :709–754.

[Dias et al., 2007] Dias, G., Alves, E., and Lopes, J. G. P. (2007). Topic segmentation algorithms for text summarization and passage retrieval : An exhaustive evaluation.

[Dias et al., 2011] Dias, G., Cleuziou, G., and Machado, D. (2011). Informative polythetic hierarchical ephemeral clustering. In *Web Intelligence and Intelligent Agent Technology (WI-IAT), 2011 IEEE/WIC/ACM International Conference on*, volume 1, pages 104–111.

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

[Diday, 1989] Diday, E. (1989). Introduction á l'analyse des données symboliques. Technical report, INRIA num.1074, Rocquencourt 78150, France.

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

[Dupuch et al., 2013] Dupuch, M., Engström, C., Silvestrov, S., Hamon, T., and Grabar, N. (2013). Comparison of clustering approaches through their application to pharmacovigilance terms. In *Artificial Intelligence in Medicine*, pages 58–67. Springer.

[D'Urso et al., 2014] D'Urso, P., Giovanni, L., and Massari, R. (2014). Trimmed fuzzy clustering for interval-valued data. *Advances in Data Analysis and Classification*, pages 1–20.

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

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

[Fisher, 1987] Fisher, D. H. (1987). Knowledge acquisition via incremental conceptual clustering. *Machine learning*, 2(2) :139–172.

[Forestier et al., 2010] Forestier, M., Velcin, J., and Zighed, D. (2010). Fouille de discussions pour l ?identification de rôles sociaux. *REiSO 2010*, page 59.

[Forgy, 1965] Forgy, E. (1965). Cluster analysis of multivariate data : Efficiency versus interpretability of classification. *Biometrics*, 21(3) :768–769.

[Fouchal et al., 2012] Fouchal, S., Monnet, Q., Mansouri, D., Mokdad, L., and Ioualalen, M. (2012). A clustering method for wireless sensors networks. In *Computers and Communications (ISCC), 2012 IEEE Symposium on*, pages 000888–000892. IEEE.

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

[Gale and Church, 1991] Gale, W. A. and Church, K. W. (1991). Identifying word correspondences in parallel texts. In *HLT*, volume 91, pages 152–157. Citeseer.

[García-Escudero and Gordaliza, 1999] García-Escudero, L. Á. and Gordaliza, A. (1999). Robustness properties of k means and trimmed k means. *Journal of the American Statistical Association*, 94(447) :956–969.

[García-Escudero et al., 2010] García-Escudero, L. A., Gordaliza, A., Matrán, C., and Mayo-Iscar, A. (2010). A review of robust clustering methods. *Adv. Data Analysis and Classification*, 4(2-3) :89–109.

[Gil-García and Pons-Porrata, 2010] Gil-García, R. and Pons-Porrata, A. (2010). Dynamic hierarchical algorithms for document clustering. *Pattern Recogn. Lett.*, 31(6) :469–477.

[Gregory, 2008] Gregory, S. (2008). A fast algorithm to find overlapping communities in networks. In *Machine Learning and Knowledge Discovery in Databases*, volume 5211 of *Lecture Notes in Computer Science*, pages 408–423. Springer Berlin Heidelberg.

[Halkidi et al., 2002] Halkidi, M., Batistakis, Y., and Vazirgiannis, M. (2002). Clustering Validity Checking Methods : Part II. *ACM SIGMOD*, 31(3) :19–27.

[Harris, 1954] Harris, Z. (1954). Distributional structure. *Word*, 10 :146–162.

[Hausdorff, 1962] Hausdorff, F. (1962). *Set theory*. Chelsea.

[Hearst and Pedersen, 1996] Hearst, M. A. and Pedersen, J. O. (1996). Reexamining the cluster hypothesis : Scatter/gather on retrieval results. pages 76–84.

[Heller and Ghahramani, 2007] Heller, K. and Ghahramani, Z. (2007). A nonparametric bayesian approach to modeling overlapping clusters. *Journal of Machine Learning Research*, 2 :187–194.

[Heskes, 1999] Heskes, T. (1999). Energy functions for self-organizing maps.

[Hubert and Arabie, 1985] Hubert, L. and Arabie, P. (1985). Comparing partitions. *Journal of Classification*, 2 :193–218.

[Karst, 1958] Karst, O. J. (1958). Linear curve fitting using least deviations. *Journal of the American Sstatistical Association*, 53(281) :118–132.

[Kohonen, 1984] Kohonen, T. (1984). Self-Organization and Associative Memory. *Springer*.

[Kozareva and Hovy, 2013] Kozareva, Z. and Hovy, E. (2013). Tailoring the automated construction of large-scale taxonomies using the web. *Language Resource Evaluation*, 47(3) :859–890.

[Kozareva et al., 2008] Kozareva, Z., Riloff, E., and Hovy, E. (2008). Semantic class learning from the web with hyponym pattern linkage graphs. In *46th Annual Meeting of the Association for Computational Linguistics : Human Language Technology (ACL-HLT)*, pages 1048–1056.

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

[Kulis and Jordan, 2011] Kulis, B. and Jordan, M. I. (2011). Revisiting k-means : New algorithms via bayesian nonparametrics. *arXiv preprint arXiv :1111.0352*.

[Largeron and Bonnevay, 2002] Largeron, C. and Bonnevay, S. (2002). A pretopological approach for structural analysis. *Information Sciences*, 144 :169–185.

[Lingras and West, 2004] Lingras, P. and West, C. (2004). Interval set clustering of web users with rough k-means. *Journal of Intelligent Information Systems*, 23(1) :5–16.

[Lu et al., 2012] Lu, H., Hong, Y., Street, W. N., Wang, F., and Tong, H. (2012). Overlapping clustering with sparseness constraints. In *Data Mining Workshops (ICDMW), 2012 IEEE 12th International Conference on*, pages 486–494. IEEE.

[Masson and Denœux, 2008] Masson, M.-H. and Denœux, T. (2008). Ecm : An evidential version of the fuzzy c-means algorithm. *Pattern Recognition*, 41(4) :1384 – 1397.

[Michalski and Stepp, 1983] Michalski, R. S. and Stepp, R. E. (1983). Learning from observation : Conceptual clustering. In *Machine learning*, pages 331–363. Springer.

[Mikolov et al., 2013] Mikolov, T., Yih, W.-t., and Zweig, G. (2013). Linguistic regularities in continuous space word representations. In *HLT-NAACL*, pages 746–751.

[Miller et al., 1990] Miller, G. A., Beckwith, R., Fellbaum, C., Gross, D., and Miller, K. J. (1990). Introduction to wordnet : An on-line lexical database. *International Journal of Lexicography*, 3(4) :235–244.

[Mirkin, 1987] Mirkin, B. (1987). The method of principal clusters. *Autom. Remote Control*, 10 :131–143.

[Mishra, 2012] Mishra, D. (2012). Discovery of overlapping pattern biclusters from gene expression data using hash based pso. *Procedia Technology*, 4 :390–394.

[Mohamed Babou, 2012] Mohamed Babou, H. (2012). *Comparaison de réseaux biologiques*. PhD thesis, Nantes.

[Moreno, 2014] Moreno, J. G. (2014). *Text-Based Ephemeral Clustering for Web Image Retrieval on Mobile Devices (version 1)*. PhD thesis, Université de Caen Basse-Normandie.

[Moreno et al., 2013] Moreno, J. G., Dias, G., and Cleuziou, G. (2013). Post-retrieval clustering using third-order similarity measures. In *ACL (2)*, pages 153–158.

[Moreno et al., 2014] Moreno, J. G., Dias, G., and Cleuziou, G. (2014). Query log driven web search results clustering. In *Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval*, pages 777–786. ACM.

[Navigli and Velardi, 2004] Navigli, R. and Velardi, P. (2004). Learning domain ontologies from document warehouses and dedicated websites. *Computational Linguistics*, 30(2) :151–179.

[Osiński et al., 2004] Osiński, S., Stefanowski, J., and Weiss, D. (2004). Lingo : Search results clustering algorithm based on singular value decomposition. pages 359–368. Springer.

[Paaß et al., 2004] Paaß, G., Kindermann, J., and Leopold, E. (2004). Learning prototype ontologies by hierachical latent semantic analysis. In *Workshop on Knowledge Discovery and Ontologies at the joint European Conferences on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML/PKDD)*.

[Page et al., 1999] Page, L., Brin, S., Motwani, R., and Winograd, T. (1999). The pagerank citation ranking : Bringing order to the web. Technical Report 1999-66, Stanford InfoLab. Previous number = SIDL-WP-1999-0120.

[Pantel and Lin, 2002] Pantel, P. and Lin, D. (2002). Discovering word senses from text. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 613–619, Edmonton, Alberta, Canada. ACM Press.

[Pedersen et al., 2004] Pedersen, T., Patwardhan, S., and Michelizzi, J. (2004). Wordnet : : Similarity : measuring the relatedness of concepts. In *Demonstration papers at hlt-naacl 2004*, pages 38–41. Association for Computational Linguistics.

[Pereira et al., 1993] Pereira, F., Tishby, N., and Lee, L. (1993). Distributional clustering of english words. In *31st Annual Meeting on Association for Computational Linguistics (ACL)*, pages 183–190.

[Pérez-Suárez et al., 2013] Pérez-Suárez, A., MartÌnez-Trinidad, J. F., Carrasco-Ochoa, J. A., and Medina-Pagola, J. E. (2013). An algorithm based on density and compactness for dynamic overlapping clustering. *Pattern Recognition*, 46(11) :3040–3055.

[Redmond and Heneghan, 2007] Redmond, S. J. and Heneghan, C. (2007). A method for initialising the< i> k</i>-means clustering algorithm using< i> kd</i>-trees. *Pattern Recognition Letters*, 28(8) :965–973.

[Régis et al., 2012] Régis, S., Doncescu, A., Takizawa, M., Cleuziou, G., et al. (2012). Initialization of masses by the okm for the belief function theory : Application to system biology. In *AINA Workshops*, pages 1167–1171.

[Rizoiu et al., 2011] Rizoiu, M.-A., Velcin, J., et al. (2011). Topic extraction for ontology learning. *Ontology Learning and Knowledge Discovery Using the Web : Challenges and Recent Advances*, pages 38–61.

[Sakai et al., 2013] Sakai, T., Dou, Z., Yamamoto, T., Liu, Y., Zhang, M., Kato, M. P., Song, R., and Iwata, M. (2013). Summary of the ntcir-10 intent-2 task : Subtopic mining and search result diversification. In *Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval*, pages 761–764. ACM.

[Sanderson and Croft, 1999] Sanderson, M. and Croft, B. (1999). Deriving concept hierarchies from text. In *22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*, pages 206–213.

[Scaiella et al., 2012] Scaiella, U., Ferragina, P., Marino, A., and Ciaramita, M. (2012). Topical clustering of search results. In *Proceedings of the fifth ACM international conference on Web search and data mining*, pages 223–232, New York, NY, USA.

[Schwarz et al., 1978] Schwarz, G. et al. (1978). Estimating the dimension of a model. *The annals of statistics*, 6(2) :461–464.

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

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

[Snoek et al., 2006] Snoek, C. G. M., Worring, M., van Gemert, J. C., Geusebroek, J.-M., and Smeulders, A. W. M. (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, pages 421–430, New York, USA. ACM.

[Snow et al., 2006] Snow, R., Jurafsky, D., and Ng, A. Y. (2006). Semantic taxonomy induction from heterogenous evidence. In *ACL '06 : Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the ACL*, pages 801–808, Morristown, NJ, USA. Association for Computational Linguistics.

[Su and Chou, 2001] Su, M.-C. and Chou, C.-H. (2001). A modified version of the k-means algorithm with a distance based on cluster symmetry. *IEEE Transactions on pattern analysis and machine intelligence*, 23(6) :674–680.

[Sublemontier, 2012] Sublemontier, J.-H. (2012). *Classification non supervisée : de la multiplicité des données à la multiplicité des analyses.* PhD thesis, Université d'Orléans.

[Sublemontier, 2013] Sublemontier, J.-H. (2013). Unsupervised collaborative boosting of clustering : an unifying framework for multi-view clustering, multiple consensus clusterings and alternative clustering. In *International Joint Conference on Neural Networks (IJCNN 2013).*

[Tang and Liu, 2009] Tang, L. and Liu, H. (2009). Scalable learning of collective behavior based on sparse social dimensions. In *Proceedings of the 18th ACM conference on Information and knowledge management*, pages 1107–1116.

[Tsoumakas et al., 2011] Tsoumakas, G., Spyromitros-Xioufis, E., Vilcek, J., and Vlahavas, I. (2011). Mulan : A java library for multi-label learning. *Journal of Machine Learning Research*, 12 :2411–2414.

[Velardi et al., 2013] Velardi, P., Faralli, S., and Navigli, R. (2013). OntoLearn Reloaded : A Graph-Based Algorithm for Taxonomy Induction. *Computational Linguistics*, 39(3) :665–707.

[Wang et al., 2010] Wang, X., Tang, L., Gao, H., and Liu, H. (2010). Discovering overlapping groups in social media. In *Proceedings of the 2010 IEEE International Conference on Data Mining*, pages 569–578.

[Whang et al., 2015] Whang, J. J., Dhillon, I. S., and Gleich, D. F. (2015). Non-exhaustive, overlapping k-means. In *proceedings of the SIAM International Conference on Data Mining (SDM) - to appear.*

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

[Wu et al., 2008] Wu, X., Kumar, V., Quinlan, J. R., Ghosh, J., Yang, Q., Motoda, H., McLachlan, G. J., Ng, A., Liu, B., Philip, S. Y., et al. (2008). Top 10 algorithms in data mining. *Knowledge and Information Systems*, 14(1) :1–37.

[Yang and Callan, 2009] Yang, H. and Callan, J. (2009). A metric-based framework for automatic taxonomy induction. In *Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 271–279.

[Youssef et al., 2009] Youssef, M., Youssef, A., and Younis, M. (2009). Overlapping multihop clustering for wireless sensor networks. *Parallel and Distributed Systems, IEEE Transactions on*, 20(12) :1844–1856.

[Zamir and Etzioni, 1998] Zamir, O. and Etzioni, O. (1998). Web document clustering : A feasibility demonstration. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '98, pages 46–54, New York, NY, USA. ACM.

[Zeng et al., 2004] Zeng, H.-J., He, Q.-C., Chen, Z., Ma, W.-Y., and Ma, J. (2004). Learning to cluster web search results. In *Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '04, pages 210–217, New York, NY, USA. ACM.