Clustering collaboratif : Principes et mise en œuvre*

Pierre Gançarski†
ICube - Université de Strasbourg

Antoine Cornuéjols
AgroParisTech - INRA/Université Paris-Saclay

Cédric Wemmert
ICube - Université de Strasbourg

Younès Bennani
LIPN - Université Paris 13

ABSTRACT

Pour tenter de faire sens des masses de données disponibles en quantité croissante, il est nécessaire de disposer d'outils performants limitant l'implication, souvent chronophage, de l'expert. Les méthodes non supervisées d'exploration de données telles que les méthodes de clustering sont une réponse à ce besoin. Cependant, leur mise en œuvre effective demande que l'utilisateur présuppose un certain nombre de propriétés des structures internes des données à mettre en évidence telles que, par exemple, leur type ou simplement leur nombre. Il doit aussi être capable de traduire cet a priori sous forme d'un ou plusieurs objectifs d'analyse et de choisir les méthodes adéquates (et leurs paramètres) de façon optimale. Malheureusement, ceci est une tâche ardue pour laquelle il n'existe pas actuellement de « recette miracle ». Par ailleurs, l'utilisation de bases de données distantes pour lesquelles il peut être intéressant de partager des informations nécessite souvent des analyses locales afin de préserver la confidentialité, .

Dans ce contexte, les approches collaboratives, dans lesquelles différents algorithmes de clustering bien que travaillant sur des données éventuellement différentes, échangent et partagent des informations apparaissent comme une voie prometteuse. En effet, les expériences montrent que l'influence des choix initiaux des méthodes et de leurs paramètres est fortement atténué et qu'un tel schéma permet de découvrir de meilleures structures que si chacune de ces méthodes travaillait isolément.

Cet article présente les questions que posent le clustering en général, et le clustering distribué en particulier ainsi que les défis à relever. Il propose ensuite un cadre dans lequel il est possible d'organiser et d'inscrire les approches de clustering collaboratif. Les différentes possibilités sont illustrées par des exemples de travaux existants.

1 INTRODUCTION

Contrairement à l'apprentissage supervisé, l'objectif de *apprentissage non supervisé* n'est pas de faire des prédictions à partir des valeurs d'entrée non encore vues, mais de révéler des structures ou des régularités potentielles cachées dans un ensemble de données. Si ces structures ou régularités putatives peuvent parfois être extrapolées pour faire des prédictions sur les données futures, ce n'est pas l'objectif principal de ce type d'apprentissage. Une autre distinction fondamentale avec l'apprentissage supervisé est qu'*il n'y a pas de moyen absolu de mesurer la pertinence des structures ou régularités mises en évidence* quelle que soit leur forme [26]. Dans l'apprentissage supervisé, on peut utiliser des ensembles de validation et via une validation croisée estimer la valeur prédictive du modèle appris. Ce n'est pas le cas en apprentissage non supervisé : un algorithme ne peut trouver que le type de structures sous-jacentes pour lequel il a été implicitement ou explicitement codé. Il n'y a donc pas de moyen objectif de mesurer la valeur des résultats d'un tel apprentissage, c'est-à-dire d'évaluer si les structures trouvées correspondent réellement à des structures sous-jacentes des données ou si ce sont simplement des artefacts issus de l'« imagination » de l'utilisateur ou de l'algorithme choisi. De ce fait, l'apprentissage non supervisé peut être considéré comme *un problème mal défini*. Cette propriété rend l'apprentissage non supervisé très difficile. On souhaite d'une part, concevoir des méthodes basées sur des théories fortes, or des biais a priori, en partie arbitraires, sont nécessaires, et d'autre part, pouvoir les appliquer aux données à traiter avec un niveau de confiance évaluable, ce qui est difficile dans l'absolu.

Le *clustering* est un type d'apprentissage non supervisé où l'objectif est de partitionner l'ensemble des données [1] en groupes appelés *clusters*. Selon l'approche utilisée, les groupes peuvent être mutuellement exclusifs (hard clustering), se chevaucher (soft clustering) ou être définis de façon floue (fuzzy clustering).

Deux grandes familles d'approches existent. L'approche *générative* suppose qu'il existe un modèle générateur des données sous-jacent, de nature souvent statistique, et l'objectif du clustering est alors de trouver les paramètres du modèle pressenti maximisant la probabilité que les données aient été générées par celui-ci. L'approche *discriminative* s'appuie sur des mesures de similarité et sur des critères d'optimisation pour trouver des groupes dans les données. On attend du résultat d'un tel clustering que les objets à l'intérieur d'un cluster sont plus similaires (ou proches) les uns des autres que les objets d'autres clusters. La mesure de similarité (ou de distance) est alors d'une importance primordiale pour définir le type de structures ou de clusters qui peuvent être découverts dans les données. Ainsi de très nombreuses mesures de similarité ou distances ont de fait été proposées dans la littérature en fonction du problème et du contexte.

Néanmoins, indépendemment du type d'approche utilisé, avant qu'un algorithme ne puisse être correctement défini, de nombreuses questions doivent trouver une réponse.

Dans la suite de cet article, nous présentons en Section 2 un bref aperçu du clustering et des questions liées et introduisons le paradigme collaboratif. Après avoir donné des exemples de scénario de clustering collaboratif (Section 3), nous présentons les caractéristiques clé des schémas collaboratifs proposés dans la littérature

*. Article extrait de [8]

†. Corresponding author : gancarski@unistra.fr

1. Dans la suite, nous utiliserons indifféremment les termes *donnée*, *objet* voire *instance* pour désigner un élément de cet ensemble

ainsi que les principaux défis liés à l'organisation et au contrôle du processus collaboratif (Section 4). La section 5 est dédiée aux questions posées par la mise en œuvre des algorithmes collaboratifs. La section 6 présente succinctement des réalisations issues de l'état de l'art. Enfin, la section 7 présente une conclusion et propose une liste de problèmes ouverts liés au clustering collaboratif.

Par souci de simplification de la lecture, nous utiliserons dans la suite de cette article, le terme de ***classifieur*** [2] *pour désigner une entité dynamique instanciant un algorithme de clustering.*

2 COMBINAISON DE CLASSIFIEURS

2.1 Pourquoi ?

En première approche, le clustering cherche à mettre les objets similaires dans un même cluster et les objets différents dans des clusters différents. Malgré cette définition apparemment simple, le clustering est en fait *un problème mal défini*. En effet, d'un point de vue mathématique, la similarité n'est pas une relation transitive : deux objets similaires à un même troisième peuvent ne pas être similaires entre eux, alors que la relation d'appartenance à un même groupe est une relation transitive.

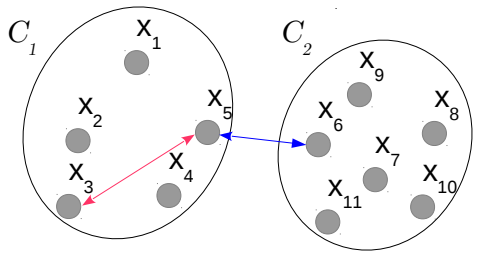


FIGURE 1: Le clustering, un problème mal défini

Ainsi, dans la figure 1 représentant des points dans un espace euclidien à deux dimensions, le clustering proposé avec les groupes C_1 et C_2 semble raisonnable. Pourtant, puisque $\mathbf{x}_5$ est similaire à $\mathbf{x}_6$, ils devraient être placés dans le même cluster, tandis que $\mathbf{x}_3$ et $\mathbf{x}_5$ qui sont dissimilaires devraient appartenir à des groupes différents.

La définition du clustering s'appuyant sur les relations de similarité est donc intrinsèquement ambiguë. Et cette ambiguïté ne peut être supprimée que par l'ajout d'un biais. Ainsi, en définissant différemment la distance utilisée pour mesurer la dissemblance entre les objets, ou entre les groupes en clustering hiérarchique (par ex. en choisissant le « single linkage », le « average linkage », ou le « complete linkage »), on favorisera un type de structures par rapport à d'autres. Cependant, ce biais affecte le processus de regroupement et non le critère d'optimalité qui reste intrinsèquement ambigu.

Par ailleurs, il n'existe pas à ce jour de critère d'optimalité tel que le problème d'optimisation devienne convexe dans l'espace de recherche. Plusieurs solutions sont possibles et le problème devient *mal posé*. Il est donc indispensable de concevoir des heuristiques, qui pour un critère d'optimalité donné, soient capables de rechercher efficacement dans l'espace des solutions malgré la taille de celui-ci.

2. Nous prenons ce terme souvent associé à la classification supervisé car malheureusement, en dehors du peu élégant anglicisme « clusterer », il n'existe pas de terme équivalent pour l'apprentissage non supervisé.

En effet, considérer toutes les partitions possibles et choisir celle qui optimise le critère devient très rapidement irréalisable, le nombre de partitions possibles en K groupes, pour N objets, étant donné par :

$$S_{N,K} = \frac{1}{K!}\sum_{k=0}^{K}(-1)^k\,(K-k)^N\binom{K}{k} \simeq \frac{K^N}{K!} \text{ quand } N \to \infty \quad (1)$$

Le nombre total de partitions à examiner est alors :

$$B_N = \sum_{k=1}^{N} S_{N,k} \quad (2)$$

Ainsi, en utilisant une machine pouvant tester 1 000 000 partitions par seconde, étudier toutes les partitions d'un ensemble de 25 éléments seulement nécessitera 147 000 années, car il y a 4 638 590 332 229 999 353 partitions possibles !

Pour résumer, le clustering est à la fois un problème *mal défini* [1, 49] et *mal posé* qui requiert l'ajout d'un biais et l'emploi d'heuristiques pour être résolu en pratique. En fonction de ceux-ci, un même ensemble d'objets peut être décomposé en clusters de différentes façons. De plus, ces heuristiques impliquent souvent le choix de valeurs de paramètres et de procédures d'initialisation susceptibles de modifier les résultats. Ainsi, il est bien souvent nécessaire de fixer a priori le nombre attendu de clusters alors que le nombre de groupes « réels » sous-jacents dans les données est rarement connu d'avance.

Par conséquent, parce que les résultats obtenus par un classifieur sont très dépendants des choix initiaux faits par l'utilisateur (algorithme et similarité utilisés, nombre de clusters attendus, etc.), ces choix doivent être effectués avec précaution afin de limiter leur influence. Cependant, il est difficile (voire impossible) d'identifier une recette idéale et générique pour cela. En l'absence d'une telle recette, une solution tentante est de ne pas avoir à choisir et d'utiliser plusieurs algorithmes de clustering avec plusieurs paramétrages et, s'inspirant du succès des méthodes d'ensembles en apprentissage supervisé, de recourir à des méthodes de collaboration pour construire soit un clustering final de meilleure qualité, soit un ensemble de clusterings décrivant des aspects alternatifs d'un même ensemble de données (*Multiple clustering*). Dans ce dernier cas, l'objectif est de favoriser des vues différentes non redondantes.

Bien qu'il soit loin d'être évident que le même succès puisse être assuré dans le contexte de l'apprentissage non supervisé, de nombreuses recherches ont été inspirées par cette idée.

2.2 Comment combiner des classifieurs

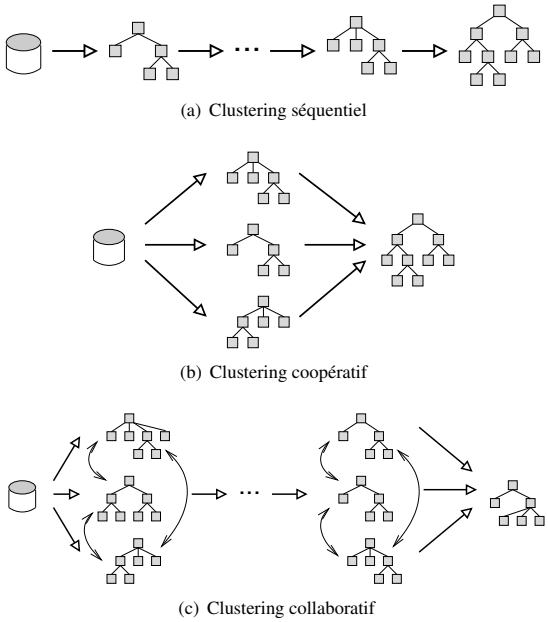
De nombreuses approches *combinant des clusterings* ont émergé [30]. L'espoir est que, en combinant plusieurs classifieurs, chacun avec son propre biais et ses imperfections, on obtiendra une meilleure solution globale [54]. L'idée principale est que les solutions fortuites et non significatives s'annuleront et que la structure réelle des données apparaîtra comme la plus partagée par les classifieurs car elle sera plus robuste aux variations des paramètres de l'algorithme. Différentes caractéristiques peuvent alors distinguer les grandes familles d'approches communément rencontrées :

• Soit les classifieurs travaillent *séquentiellement* (Figure 2-a), comme dans les approche de type cascading [2], chaque classifieur utilisant des informations fournies par les classifieurs de l'étape précédente,

voire des données supplémentaires. Soit les classifieurs travaillent en *parallèle* :

— En clustering *coopératif* (Figure 2-b), chaque classifieur produit ses résultats *indépendemment*. Le clustering final est calculé dans une étape de post-traitement.

— En clustering *collaboratif* (Figure 2-c), les classifieurs collaborent *en interaction*. L'action de chaque classifieur est adaptée aux performances du groupe et vice versa. Une question essentielle est alors de déterminer si le système converge bien vers un point fixe.



(a) Clustering séquentiel

(b) Clustering coopératif

(c) Clustering collaboratif

FIGURE 2: Trois schémas pour la combinaison de classifieurs

• Chaque classifieur peut être appliqué soit sur l'*ensemble des données* ou soit uniquement sur une *sous-partie*. En effet, pour des raisons de confidentialité, de propriété ou de stockage, il se peut que les données ne puissent pas être réunies sur un site unique pour leur traitement [45]. On distingue alors :

— le *clustering vertical*. : un classifieur accède uniquement à une vue partielle des caractéristiques (attributs) des données.

— le *clustering horizontal* : un classifieur accède à toutes les caractéristiques mais uniquement sur un sous-ensemble de données.

2.3 Coopération vs. collaboration

Dans les *méthodes d'ensembles* développées en apprentissage supervisé chaque classifieur apprend indépendemment des autres [3]. En revanche, le classement d'une nouvelle donnée est obtenu par un vote (pondéré) de leurs prédictions [30]. De nombreuses méthodes ont été développées, incluant le Bagging, le Boosting et les Random Forets [11, 32, 47, 60] et sont devenues des standards obtenant souvent de bons résultats. Les méthodes d'ensembles sont en effet faciles à mettre en œuvre. D'une part, il est simple de mesurer à la fois la performance et la diversité des fonctions de prédiction individuelles h_i candidates à faire partie de la fonction de prédiction agrégée H. D'autre part, il existe de nombreuses possibilités pour agréger ces fonctions prédictives en une telle fonction globale H, comme par exemple une combinaison linéaire. Or, en clustering, il n'existe pas, en général, de correspondance directe entre les clusters trouvés par les différents classifieurs. La mise en œuvre d'algorithmes de vote n'est plus du tout triviale. Diverses propositions heuristiques ont néanmoins été faites pour tenter de remédier à ce problème :

— [58] introduit une nouvelle méthode de vote dans laquelle une « bonne » correspondance suffit

— [41] présente trois méthodes de consensus par vote et par paires

— [3] propose une méthode de vote cumulatif à partir de partitions à nombre variable de clusters

— [18] utilise une matrice $N \times N$ de co-appartenance aux clusters comme mesure de similarité entre les données.

— [29] présente un modèle coopératif basé sur une représentation histogrammique des similarités des objets dans les clusters.

Toutefois, devant la complexité algorithmique et les coûts en temps de calcul qui en découlent dès que le volume des données augmente, les travaux sur les approches coopératives [12, 22, 33, 51, 55, 57, 60, 61] se sont fait rares au bénéfice du schéma collaboratif dans lequel l'objectif est que chaque classifieur, travaillant éventuellement sur un ensemble de données distinct, bénéficie du travail effectué par les autres « collaborateurs ». Cela peut se faire via l'échange d'informations locales sur les données ou sur le clustering local ou encore sur la valeur des paramètres ou d'initialisation d'un algorithme. La validité de l'approche repose sur l'hypothèse que des informations utiles peuvent être partagées entre des processus indépendants.

Ce schéma conduit naturellement à des mises en œuvre distribuées. Mais contrairement au modèle coopératif, il implique généralement plusieurs itérations avant la convergence vers une solution globale satisfaisante. De fait, en plus du problème de l'information à échanger entre les classifieurs, un mécanisme de suivi de l'évolution de la solution globale capable d'arrêter le processus, doit être défini.

La suite du document est consacrée au clustering collaboratif.

3 EXEMPLES DE COLLABORATIONS

Dans la suite nous considérons quatre scénarios illustratifs :

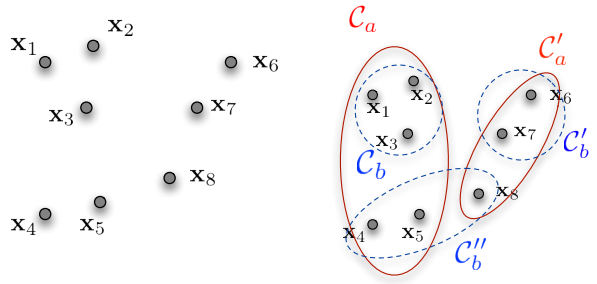
(1) Scénario 1 : Les classifieurs ont accès au *même ensemble de données* avec les mêmes attributs.

(2) Scénario 2 : Les classifieurs ont accès au *même ensemble de données*, mais ils n'en ont qu'une *vue partielle* (sous ensemble des attributs) : clustering vertical.

(3) Scénario 3 : Les classifieurs n'ont accès qu'à des *sous-ensembles d'objets* supposés tirés de la même distribution et décrits avec le *même ensemble d'attributs* : clustering horizontal.

(4) Scénario 4 : Les classifieurs ont accès à *différents objets* supposés tirés de la même distribution et décrits avec *différents ensembles d'attributs indépendants*.

3. Mais éventuellement en étant influencé par leurs résultats, comme dans le boosting.



(a) Données partagées (b) Résultats des deux classifieurs

FIGURE 3: Données et résultats à unifier

Pour chaque scénario ci-dessus, nous examinerons quelle communication pourrait être mise en place entre les classifieurs dans le cadre d'un processus collaboratif. En effet, il existe deux types d'informations que les classifieurs peuvent partager, utiliser et mettre à jour au cours de leurs calculs : des informations sur l'*appartenance des données aux clusters* (liste des objets dans les clusters, fonction d'appartenance, ...) d'une part, et des informations sur des *paramètres internes* (nombre K de clusters recherchés, proportion d'objets affectés à chaque cluster, coordonnées du centre de chaque cluster, ...) d'autre part (cf. Section 5.1).

3.1 Les différents scénarios

• *Scénario 1 : Mêmes données, mêmes attributs*

Lorsque les algorithmes peuvent partager les identifiants des objets, il est facile de comparer les clusters. Par exemple, la figure 3 présente 8 objets et deux résultats de clustering A et B. Les appartenances peuvent être comparées comme suit :

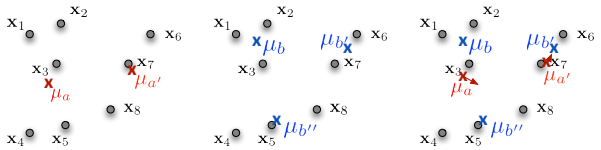
Algorithme	$\mathbf{x}_1$	$\mathbf{x}_2$	$\mathbf{x}_3$	$\mathbf{x}_4$	$\mathbf{x}_5$	$\mathbf{x}_6$	$\mathbf{x}_7$	$\mathbf{x}_8$
A	C_a	C_a	C_a	C_a	C_a	C'_a	C'_a	C'_a
B	C_b	C_b	C_b	C'_b	C'_b	C''_b	C''_b	C''_b

Dans ce cas, il est évident que le cluster C_b est compatible avec le cluster C_a : $C_b \subseteq C_a$. De même, $C'_b \subseteq C'_a$. Par contre, il existe un conflit concernant le cluster C''_b : il recouvre en partie les clusters C_a et C'_a. Chaque classifieur doit prendre des mesures pour le réduire.

Une première approche consiste à ce que les classifieurs agissent sur les objets en conflit. Ainsi, le classifieur A déplacera $\mathbf{x}_8$ vers le cluster C_a avant de mettre à jour le prototype μ_a afin de contraindre sa solution à se rapprocher de celle de B. Simultanément, le classifieur B déplacera $\mathbf{x}_8$ vers C'_b. Bien sûr, si cela se fait simultanément, le conflit persistera mettant en évidence l'effet « Red Queen » par lequel deux opérations a priori positives s'annulent [4].

Une deuxième approche consiste à ce que les classifieurs échangent les coordonnées de leurs centres de clusters ainsi que leurs cardinalités (Figure 4-a et b). Ainsi, A informera B qu'il a deux centres (ou prototypes) de coordonnées $\mu_a = [\mu_a^{(1)}, \mu_a^{(2)}]^\top$ et $\mu_{a'} = [\mu_{a'}^{(1)}, \mu_{a'}^{(2)}]^\top$

4. Cette hypothèse tire son nom d'un épisode du livre de Lewis Carroll *De l'autre côté du miroir* (deuxième volet d'Alice au pays des merveilles) au cours duquel Alice et la Reine Rouge se lancent dans une course effrénée alors que parallèlement les gardes reforment la haie d'honneur. Alice demande alors : « Mais, Reine Rouge, c'est étrange, nous courons vite et le paysage autour de nous ne change pas ? » Et la reine de répondre : « Nous courons pour rester à la même place »



(a) Centres des clusters A (b) Centres des clusters B (c) Nouveaux centres A

FIGURE 4: Échanges d'information sur les centres des clusters

de cardinalités relatives 5/8 et 3/8 (Figure 4 - gauche). Réciproquement, B communiquera vers A, $\mu_b = [\mu_b^{(1)}, \mu_b^{(2)}]^\top, \mu_{b'} = [\mu_{b'}^{(1)}, \mu_{b'}^{(2)}]^\top$ et $\mu_{b''} = [\mu_{b''}^{(1)}, \mu_{b''}^{(2)}]^\top$ avec les cardinalités relatives 3/8, 2/8 et 3/8. Le classifieur A pourra alors affecter chaque centre issu de B au centre de A le plus proche et recalculer alors ses propres centres (Figure 4-c). Enfin, il réaffectera tous les objets au centre le plus proche. Ici, $\mathbf{x}_5$ se déplacera vers C_a mettant fin au conflit avec B.

Il est évident que ces approches collaboratives peuvent être aisément implémentées et étendues à plus de deux algorithmes. Néanmoins, le nombre de messages échangés augmentera en $O(P^2)$ où P est le nombre de classifieurs. De plus, le calcul des objets en conflit sera plus compliqué que dans l'exemple simple décrit ici. Enfin, il n'est pas évident de définir un processus de collaboration garantissant que tous les processus convergent vers des solutions locales stables. En effet, une modification d'une solution peut déclencher, par effet domino, des changements dans toutes les autres solutions. De fait, une amélioration itérative de la solution globale par des boucles `Détection de conflits - Résolution des conflits` ne peut être garantie que si (1) une fonction de qualité globale est correctement définie et (2) le schéma de collaboration permet une amélioration de celle-ci à chaque boucle (ou a minima en tendance).

• *Scénario 2 : Mêmes données, sous-ens. d'attributs différents*

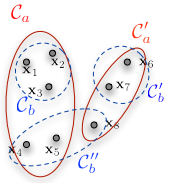
Lorsque les classifieurs peuvent partager les identifiants des objets, mais que les attributs sont différents, les classifieurs ne peuvent plus utiliser d'informations sur les centres des clusters issus d'autres classifieurs. Cependant, ils peuvent encore comparer les affectations des objets à leurs clusters par exemple, en utilisant des matrices de consensus [39]. Une *matrice de consensus* M est une matrice $N \times N$ où $M(i,j) = p_{i,j}/P$ où $p_{i,j}$ est le nombre de fois où les objets i et j sont mis ensemble dans le même cluster par un classifieur et P est le nombre de classifieurs. Idéalement, en cas d'accord sur le clustering par tous les classifieurs $M(i,j) \in \{0,1\}$ et il est alors évident de construire les clusters.

La figure 5 présente la matrice de consensus pour l'ensemble de données de la figure 3. Il est évident que $\mathbf{x}_8$ est source de conflits.

• *Scénario 3 : Données différentes, mêmes attributs*

Lorsque les ensembles de données considérés par les classifieurs sont différents, une collaboration n'a de sens que si l'on suppose que ces ensembles proviennent d'une même distribution. En général, il existe un intersection non vide entre les ensembles. La solution des matrices de consensus peut alors être utilisée en les normalisant. Dans le cas contraire, une autre solution consiste à utiliser le transfert des centres de clusters puisque l'espace des données est le même.

• *Scénario 4 : Données différentes, attributs différents*



	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8
X_1	-	1	1	½	½	0	0	0
X_2	1	-	1	½	½	0	0	0
X_3	1	1	-	½	½	0	0	0
X_4	½	½	½	-	1	0	0	½
X_5	½	½	½	1	-	0	0	½
X_6	0	0	0	0	0	-	1	½
X_7	0	0	0	0	0	1	-	½
X_8	0	0	0	½	½	½	½	-

Les rectangles grisés montrent les clusters potentiels

FIGURE 5: Deux clusterings et la matrice de consensus associée.

Même si les attributs considérés par les classifieurs ne sont pas identiques, l'échange des coordonnées des centres peut souvent être fait car, en général, il existe une intersection non vide entre les sous-ensembles d'attributs utilisés. Bien évidement, plus petite est l'intersection, plus le risque que les attributs communs ne soient plus assez informatifs est grand.

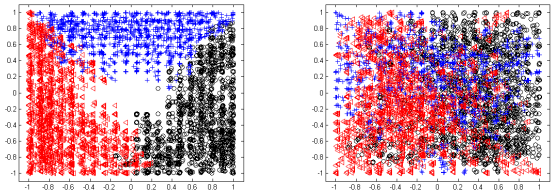
Dans le cas "extrême" où les sous-ensembles de données et les sous-ensembles d'attributs sont (quasi) disjoints, la seule information partageable porte alors sur les structures des clusters comme par exemple les cardinalités relatives des clusters pour chacun des classifieurs. Ainsi, si pour A, les cardinalités des deux clusters trouvés sont $\{0.25, 0.75\}$ et que B trouve $\{0.10, 0.13, 0.77\}$ pour trois clusters, alors il n'est pas absurde de supposer que les clusters trouvés par les deux classifieurs peuvent être mis en relation étroite ($0.25 \approx 0.10 + 0.13$). Cependant, ce genre de situation est trop simpliste pour être réaliste et une analyse plus sophistiquée devra être menée.

3.2 Un exemple concret

En clustering collaboratif, l'espoir est que la combinaison de plusieurs classifieurs débouchera sur de meilleurs résultats qu'en utilisant chaque classifieur indépendemment. Mais cela est-il toujours vrai ? Par exemple, que se passe-t-il si deux classifieurs A et B utilisent des attributs différents tels les attributs utilisés par A sont pertinents pour détecter la structure existante sous-jacente alors que ceux utilisés par B sont totalement aléatoires, ne contenant a priori aucun information utile.

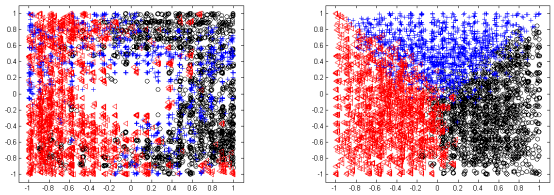
Pour illustrer ce cas, l'algorithme F-VBGTM [20] a été utilisé sur les données UCI *Waveform*, soit 5 000 observations décrites par 21 variables pertinentes et 19 variables non informatives. Deux ensembles de données D_A et D_B ont été créés : D_A contient toutes données initiales décrites avec uniquement les variables pertinentes alors que dans D_B ces mêmes données sont décrites par les variables non informatives. Les clusterings obtenus indépendemment par les deux classifieurs sont présentés dans la figure 6. Il y apparaît que le classifieur A utilisant D_A a pu capturer la structure sous-jacente de l'ensemble de données, tandis que le classifieur B utilisant D_B a produit une solution manisfestement aléatoire.

Dans un deuxième temps, nous avons mis en œuvre un schéma de collaboration dans lequel les résultats de B ont été envoyés à A avant son propre traitement de l'ensemble de données. Comme on pouvait s'y attendre, lorsque le classifieur B influence le classifieur A, la structure découverte est gravement altérée (Figure 7-a). À l'inverse, le classifieur B bénéficie des informations transmises par A (Figure 7-b). Ces résultats ne sont que qualitatifs, mais ils montrent le potentiel bénéfique mais aussi « nocif » de la collaboration.



(a) Classifieur A (b) Classifieur B

FIGURE 6: Deux classifieurs (Même données, sous-ensembles d'attributs différents)



(a) Classifieur A (b) Classifieur B

FIGURE 7: Influence de la collaboration sur les résultats

4 CONDITIONS POUR UNE COLLABORATION BÉNÉFIQUE

En clustering *coopératif*, l'objectif est de trouver la structure sous-jacente dans un ensemble de données en utilisant plusieurs classifieurs. La notion de *consensus* entre les solutions locales est centrale pour contrôler le processus coopératif. Et bien que la mesure du consensus puisse dépendre des spécificités du domaine d'application, il semble relativement facile de définir une telle mesure.

La situation est très différente dans le clustering *collaboratif* où l'objectif est d'identifier les structures propres à chacun des ensembles de données locales considérés par les différents algorithmes. Dans ce cadre, il est plus difficile de garantir que la collaboration apporte des améliorations locales, d'une part, et de contrôler le processus, d'autre part, puisque la notion de consensus global n'est plus opérationnelle. Néanmoins, on peut juger que la collaboration est utile si les clusterings locaux finaux ont une qualité supérieure à ceux obtenus sans collaboration. Les avantages potentiels de la collaboration peuvent avoir trois origines :

(1) Une *augmentation du nombre des données disponibles* par l'utilisation de données locales implique en général qu'une meilleure solution sera atteinte par une procédure inductive qu'elle soit supervisée (données d'apprentissage) ou non (données non labélisées). En effet, en général, un échantillon de données plus important réduit la variance, et donc l'erreur d'approximation tout en laissant l'erreur due au biais intacte. Néanmoins, un classifieur ne pourra bénéficier d'informations provenant d'un autre classifieur que si l'ensemble de données de ce dernier partage suffisamment de régularités avec son

propre ensemble de données (voir l'exemple concret précédent - section 3.2). Un problème est alors de s'assurer que les données entrantes obéissent effectivement à la même distribution. Les mesures de la divergence des distributions de probabilité telles que la divergence de Kullback-Leibler, l'entropie relative, le gain d'informations ... peuvent être utilisées pour cela [37].

(2) Le *rôle perturbatif* des informations venant d'autres classifieurs. En apprentissage non supervisé, les critères inductifs sont souvent liés, dans une certaine mesure, à la compacité et dissemblance des sous-structures que l'on s'intéresse à découvrir dans les données. Par exemple, dans le cas de l'algorithme K-MEANS, le critère à minimiser n'étant pas convexe, l'algorithme peut tomber dans des minimums locaux. Toute source de perturbation peut potentiellement aider l'algorithme à s'en échapper, et les informations provenant d'autres algorithmes collaboratifs peuvent jouer ce rôle. La difficulté reste néanmoins de déterminer si les informations entrantes sont bénéfiques, inutiles ou nuisibles. Ceci reste un problème ouvert bien que de nombreux travaux s'y soient intéressés [21, 25, 40, 43, 50].

(3) L'apport d'informations externes peut *atténuer le biais propre à chaque algorithme*. Par exemple, si classifieur *A* utilise K-MEANS, il sera biaisé vers la découverte de clusters convexes. Un classifieur *B* basé sur COBWEB produira quant à lui des hiérarchies. Utiliser des informations sur les clusters produits par *B* pour fusionner des clusters de *A* peut permettre d'obtenir des clusters non convexes. Réciproquement, fusionner des noeuds de la hiérarchie ne partageant pas un père commun, amène à des structure en graphe et non plus en hiérarchie stricte. Néanmoins, la question demeure : en quoi ces modifications de biais, correspondant dans ce cas à des pertes de propriétés, sont-elles bénéfiques plutôt que nuisible ? Et comment pourrait-on évaluer le bénéfice, le cas échéant, d'une telle modification ? Encore une fois, nous sommes ramenés à la question d'évaluer la pertinence d'une méthode de clustering, une question pour laquelle il n'existe pas de réponse définitive.

Malheureusement l'identification de ces trois origines ne se traduit pas trivialement dans une procédure de contrôle globale qui conduirait à de meilleures performances. En outre, le fait que chaque algorithme local est influencé par les résultats d'autres algorithmes et les affecte en retour débouche sur un processus global difficile à caractériser et ne présentant pas nécessairement un point fixe terminal. L'analyse des schémas collaboratifs est donc délicate à réaliser. Premièrement, déterminer si une de ces situations est valide dans un contexte donné n'est pas évident. Deuxièmement, même s'il est établi qu'un ou plusieurs de ces trois cas sont présents, aucune de ces trois causes potentielles ne garantit des améliorations. En effet, un algorithme local qui fonctionne sur son propre ensemble de données ne peut bénéficier d'informations provenant d'un autre algorithme que si l'ensemble de données de ce dernier partage suffisamment de régularités avec son propre ensemble de données. De même, les perturbations peuvent aider le processus d'exploration d'un classifieur mais peuvent aussi l'entraver si elles l'orientent vers des régions peu intéressantes de l'espace de recherche. Enfin, il n'y a aucune raison a priori que les informations externes qui modifient de fait le critère d'optimisation d'un algorithme local, le modifient de façon bénéfique.

Par conséquent, une attention particulière doit être portée afin de s'assurer que le processus collaboratif peut améliorer les performances de chaque algorithme local et que la stratégie de contrôle a été soigneusement conçue. L'absence d'une notion naturelle de consensus global rend le problème plus compliqué.

5 CONDITIONS DE MISE EN ŒUVRE

Dans le clustering collaboratif, un ensemble de classifieurs (ou *experts*) échangent des informations pendant qu'ils exécutent leur tâche locale. Pour réaliser cela, il faut répondre aux questions suivantes :

(1) Quelles sont les *informations* qui devraient être échangées entre les classifieurs ?

(2) Comment mesurer la *diversité* des classifieurs ? et celle des résultats locaux ?

(3) Comment mesurer la *performance* de chaque classifieur ?

(4) Comment mesurer la qualité de l'*objectif commun* s'il en existe un ? Et estimer le degré d'avancement vers cet objectif ?

(5) Comment contrôler le *processus* collaboratif ?

5.1 Échange d'informations entre classifieurs

Dans le paradigme collaboratif, les classifieurs peuvent échanger des informations sur :

(1) l'assignation des données aux clusters.

(2) les structures candidates

Information sur les données. La communication d'informations sur l'*appartenance hypothétique aux données* aux clusters consiste en général à transférer des identifiants de clusters et les données associées. Il est évident que cela exige qu'il existe un accord sur ces identifiants. Une difficulté est que chaque classifieur utilise son propre étiquetage des clusters et, par conséquent, une traduction doit être faite localement afin de comparer les assignations et clusters locaux avec ceux qui sont communiqués. Ce type d'échange est plus onéreux dans la mesure où il est proportionnel à la taille de l'ensemble de données. En outre, il nécessite généralement plus de calculs de la part de chacun des classifieurs.

Informations sur la structure. Typiquement, ce sont souvent les attributs détectés comme pertinents (ou non), les distances à utiliser, ou encore sur le nombre de clusters à envisager qui sont transmis entre les classifieurs. Cet échange est particulièrement efficace en clustering horizontal où les classifieurs travaillent sur des données, voire des espaces de représentation des données différents et donc ne peuvent échanger sur les données elles-mêmes. L'avantage de ce type de communication est qu'il préserve naturellement la confidentialité des données et qu'il est également parcimonieux dans la mesure où seules quelques valeurs numériques sont transmises.

Outre le type d'information échangée entre experts, une question concerne la directionnalité et l'intensité de la communication. Si, par exemple, un classifieur semble supérieur à un autre selon une mesure de performance donnée, doivent-ils quand même échanger des informations de manière symétrique ? La question, peu traitée

dans littérature, reste ouverte et nous ramène une nouvelle fois à la question du calcul de la performance d'un classifieur.

5.2 Diversité et choix des classifieurs

Dans la plupart des approches existantes en clustering coopératif, toutes les méthodes disponibles (et leur paramétrages propres) sont potentiellement utilisables dans la mesure où toutes les méthodes sont jugées pertinentes et peuvent contribuer à la qualité du résultat final.

De plus, comme dans toute « assemblée d'experts », s'il n'y a pas de variété dans les opinions, la collaboration ne peut guère être fructueuse. Il est donc communément admis qu'en clustering collaboratif, une certaine diversité dans les méthodes, leurs paramètrisations et leurs initialisations est nécessaire, voire indispensable. Mais dans quelle mesure ? Ceci n'est pas une question simple, et si des travaux ont tenté d'y répondre (voir section 6), leurs résultats ne sont pas assez concluants car de nombreux facteurs entrent en jeu pour expliquer les résultats. Ainsi, par exemple, [7] propose une méthode pour mesurer la diversité des classifieurs en examinant leur accord sur un jeu de données spécifique par rapport à un jeu de données créé aléatoirement. Cette mesure est alors utilisée pour estimer la contribution de chaque expert et donc l'apport de la diversité.

5.3 Qualité des résultats locaux et performance des classifieurs

La mesure des performances dans le clustering est l'une des questions les plus difficiles en apprentissage non supervisé, sans réponse parfaite ni même une réponse meilleure que les autres. Intrinsèquement, le clustering est une technique exploratoire qui vise à découvrir des structures insoupçonnées dans les données. Or il semble impossible de définir un critère d'optimisation universel car il sera toujours biaisé vers le type de régularités recherchées qui peut être différent des régularités réellement présentes dans les données. De ce fait, si la structure sous-jacente des données ne présente pas ce type de régularités, l'algorithme ne pourra être évalué via ce critère. Par exemple, si le critère de qualité est basé sur la compacité des clusters, les algorithmes utilisant le lien minimal seront défavorisés car ne tenant compte que de l'écart entre les groupes à fusionner. De fait, pour qu'une mesure puisse être utilisée comme évaluation de la performance du classifieur, elle devra être en adéquation avec le type de structures recherchés. Sinon, elle sera trompeuse.

Supposons alors que l'on dispose d'un critère adéquat (e.g. la compacité) et partagé par tous les classifieurs (e.g. basés sur K-means), alors l'utilisation d'une approche d'ensembles ou collaborative peut être une bonne idée. Néanmoins on voit bien que la condition d'avoir un critère commun limite les possibilités d'avoir une grande diversité des classifieurs.

La question est donc, en cas d'absence d'un tel critère commun, peut-on espérer qu'en utilisant des algorithmes différents, chacun classifieur cherchant donc à optimiser une mesure de performance différente, on obtiendra des résultats locaux améliorés ? Il n'y malheureusement pas actuellement de moyen de trancher cette question.

Un dernière possibilité est d'évaluer les résultats proposés suivant plusieurs mesures en espérant ainsi aider les algorithmes à échapper aux minima locaux en les soumettant à des évaluations « étrangères » à leur type de structures.

5.4 Évaluation d'un résultat global

L'agrégation (ou unification) de différents classifieurs dépend de l'objectif final. Il peut s'agir :

- Obj_1 de construire à partir des résultats des classifieurs locaux, un clustering consensus meilleur que le meilleur des clusterings proposés selon un critère de qualité de clustering donné ;
- Obj_2 de choisir parmi tous les résultats de ces classifieurs, un clustering consensus *moyen*, c'est à dire par exemple, le clustering le plus cohérent avec tous les autres ;
- Obj_3 de construire une solution dans laquelle (1) chaque clustering est de meilleure qualité (selon son propre critère de qualité) que s'il avait été produit indépendemment et (2) ces clusterings locaux sont les plus cohérents possible entre eux ;
- Obj_4 de combiner ces deux aspects via un procédure en deux étapes consistant d'abord à améliorer les clusterings locaux (Obj_3) avant de les réunir en un seul clustering (Obj_1 ou Obj_2) .

Si la qualité d'un consensus construit (Obj_1) ou extrait (Obj_2) ne dépend que du consensus lui même, le choix du consensus moyen et la qualité du résultat global dans les autres cas dépendent de la cohérence entre les clusterings locaux. Ceci se traduit souvent en terme d'*agrément* ou de *similarité* ou, inversement, de *discordance* ou *distance* inter-résultats. Cette notion doit donc être formalisée pour pouvoir être utilisée comme une mesure. La qualité globale d'un ensemble de P classifieurs $C = \{C_1, \ldots, C_p, \ldots, C_P\}$ peut se définir par :

$$Q_{global}(C) = \sum_{i=p}^{P} G(C_p, Q_p)$$

où $G(C_p, Q_p)$ est la qualité du clustering C_p, muni d'une mesure de qualité propre Q_p, au sein de l'ensemble des P classifieurs avec :

$$G(C_p, Q_p) \;=\; \alpha Q_p(C_p) \;+\; \beta(\text{Agrement}(C_p, \{C_{p' \neq p}\})\Big)$$

où *Agrement* est la mesure de cohérence et α, β les coefficients permettant de quantifier l'importance relative des deux critères.

Il est important de noter que la fonction *Agrement* ne tient généralement pas compte des descriptions originales des données, mais seulement de leurs attributions aux clusters dans les différents classifieurs. Ainsi, cet agrément est souvent basé sur la *cohérence* des clusterings deux à deux. Celle-ci est généralement mesurée via l'*information mutuelle* portée par les affectations aux clusters dans les deux résultats [51].

Ainsi, pour deux clusterings $C_p = \{C_{(p,1)}, \ldots, C_{(p,i)}, \ldots, C_{(p,K_p)}\}$ et $C_{p'} = \{C_{(p',1)}, \ldots, C_{(p',i)}, \ldots, C_{(p',K_{p'})}\}$ avec K_p et $K_{p'}$ le nombre respectif de clusters, elle est définie par :

$$I(C_p, C_{p'}) \;=\; \sum_{i=1}^{K_p} \sum_{j=1}^{K_{p'}} p(C_{(p,i)}, C_{(p',j)}) \log\left(\frac{p(C_{(p,i)}, C_{(p',j)})}{p(C_{(p,i)})\, p(C_{(p',j)})}\right)$$

Une stratégie alternative pour mesurer cet agrément est basée sur l'utilisation d'une *matrice de consensus* donnant la proportion de fois où deux données x_a et x_b sont mises dans le même cluster pour

chacun des classifieurs. Cette matrice $N \times N$ est définie par :

$$M(a,b) = \frac{1}{P}\sum_{p=1}^{P} M_p(x_a, x_b)$$

où

$$M_p(x_a, x_b) = \begin{cases} 1 & \text{si } x_a \in C_p \text{ et } x_b \in C_p \\ 0 & \text{sinon} \end{cases}$$

L'agrément est alors calculé à partir de la similarité entre les P classifieurs qui peut être évaluée par :

$$S(C_1, \ldots, C_P) = \sum_{a=1}^{N} \sum_{b=1, b\neq a}^{N} M(a,b)$$

D'autres techniques ont été proposées basées sur des structures de graphes [34] ou sur les recouvrements mutuels des clusters des différents classifieurs [19].

Néanmoins, trouver le clustering consensuel C_{opt} à partir de ces fonctions et tel que $Q_{global}(C)$ soit maximale sur l'ensemble des résultats possibles est malheureusement un problème d'optimisation combinatoire qui s'est révélé NP-complet. Des heuristiques existent qui aboutissent à des solutions approximatives, mais elles dépendent des caractéristiques des données et du nombre de clusters recherchés.

Un exemple de fonction de consensus utilisée pour le clustering distribué de pixels issus d'images satellitaires peut être trouvé dans [53]. Il est à noter que dans le cas d'un clustering vertical, il n'est plus possible de faire des comparaisons directes au niveau des données car les classifieurs travaillent sur des ensembles de données différents. Seules les descriptions des clusters produits par chaque classifieur peuvent être échangées. Très peu de travaux se sont intéressés à ce cas, une exception étant [14].

5.5 Contrôle du processus de collaboration

Plusieurs stratégies de contrôle du processus de collaboration sont possibles et peuvent être caractérisées selon plusieurs dimensions, fortement dépendantes, telles que :

• *Synchronisme* versus *asynchronisme* : En mode asynchrone, chaque classifieur fonctionne à son rythme et n'échange de l'information que sur demande ou lorsqu'il a besoin lui-même d'une information supplémentaire. Cette stratégie est la meilleure lorsque chaque classifieur a son propre objectif et n'échange de l'information que dans la mesure où cela peut l'aider à atteindre de celui-ci. Lorsque tous les experts sont impliqués dans un même objectif global, le synchronisme est généralement requis car le résultat final dépend de toutes les réalisations locales. Le choix est évidemment tributaire du type de processus : itératif ou en une-passe.

• Processus *itératif* versus *en une-passe*. Dans un processus en une-passe, les classifieurs calculent leurs solutions locales lors d'une première phase, et la solution globale est construite directement à partir de ces résultats. Si, en première phase, les classifieurs prennent des temps différents pour produire leur solution, le processus global de calcul doit soit attendre la dernier résultat, soit être capable de construire des solutions globales de façon incrémentale. La plupart des techniques proposées dans la littérature se synchronisent sur le classifieur le plus lent car les techniques incrémentales d'apprentissage et d'unification restent pour l'instant peu développées et validées. Lorsque, en revanche, les calculs locaux peuvent prendre en compte des solutions partielles communiquées par d'autres experts, soit le processus est cadencé avec une horloge principale - chaque classifieur doit produire et communiquer sa solution temporaire avant qu'une autre phase de collaboration commence - soit les classifieurs communiquent librement, de manière asynchrone, de leur propre initiative. La plupart des propositions actuelles utilisent une stratégie de contrôle synchrone.

• Contrôle *local* versus *global*. Le choix entre un contrôle local ou global se fait en général en fonction de la volonté ou non de produire un consensus final.

• Garantie de *terminaison*. En mode asynchrone, il peut être difficile de s'assurer que tous les classifieurs s'arrêtent même si chacun exécute des fonctions en temps fini. En effet, ce problème bien connu en programmation de systèmes répartis, provient du fait que les échanges d'information entre les classifieurs peuvent produire des réactivations de ceux-ci. La mise en œuvre de solutions proposées dans le cadre des systèmes répartis est malheureusement à la fois difficile et trop pénalisant en général. Néanmoins, une méthode générale qui ne tient pas compte de l'avancement de la solution, consiste à définir une fonction d'énergie qui diminue à chaque itération. C'est, par exemple, l'approche adoptée dans [53].

• Garantie de *convergence* vers la solution optimale. Comme tout système d'optimisation heuristique, une telle garantie est malheureusement impossible dans le majorité des cas. Néanmoins, les expériences montrent que les perturbations apportées par la collaboration sur les différents classifieurs favorisent la sortie des optimaux locaux et donc la convergence vers des solutions de meilleures qualités.

6 ÉTAT DE L'ART

L'apprentissage collaboratif désigne les algorithmes distribués qui s'influencent mutuellement lors de leurs propres calculs en échangeant des informations. Le Pandemonium conçu par Oliver Selfridge en 1958 [42, 48] et le système sophistiqué Hearsay II développé pour la compréhension de la parole dans les années 1970 [4] sont parmi les premiers exemples de tels systèmes. Le schéma de co-apprentissage développé pour la classification supervisée à partir de vues partielles et disjointes [5] a également une place particulière dans l'histoire de l'apprentissage collaboratif puisqu'il était basé sur une analyse formelle menant à la définition de l'algorithme.

En apprentissage non supervisé, les travaux sur le co-clustering et le bi-clustering, initiés par [24], peuvent être considérés comme des précurseurs de ce qui n'était pas encore appelé clustering collaboratif. Le co-clustering est en effet un cas limite où l'algorithme peut être considéré comme un processus impliquant deux classifieurs qui échangent des informations sur le même ensemble de données, chacun ne voyant qu'une seule dimension (p. ex. objets vs. attributs) et essayant de trouver des clusters sur cette dimension en tenant compte de ce qui est découvert par l'autre classifieur. Des travaux plus récents sur le co-clustering [5] se trouvent dans [6, 27, 59].

Cependant, ce n'est qu'à la fin des années 1990 que les approches collaboratives émergent réellement comme un concept à part entière digne d'intérêt [28]. Depuis, de nombreuses méthodes ont été proposées, que l'on peut organiser selon différents aspects tels que : méthodes proposant des clusterings durs vs. floues, méthodes s'appuyant sur des cartes topologiques ou autres algorithmes existants

5. Domaine qui n'entre pas dans le périmètre de cet article

tels que K-MEANS ou encore méthodes dédiées aux données distribuées .

6.1 Un tour d'horizon

• *Fuzzy collaborative clustering.* Un schéma de clustering collaboratif à base de classifieurs flous a été proposé dans [44]. Dans ce schéma différents sous-ensembles des données sont traités de manière indépendante. Puis, chacun de ces clusterings est modifié en fonction des résultats trouvés sur les autres sous-ensembles. Des expériences approfondies de la méthode sont également proposées dans [46] avec des détails algorithmiques. Une application de cette méthode collaborative à l'analyse de contenus Web a été proposée dans [36] pour découvrir les structures cachées à la fois dans l'espace sémantique et dans celui des données.

Une autre approche est proposée par [38], où des ensembles flous sont utilisés dans un paradigme collaboratif dans lequel plusieurs sous-ensembles de motifs sont traités ensemble pour trouver une structure commune. Un algorithme de clustering est développé en intégrant les avantages des clusters flous et des clusters durs. Une analyse quantitative des résultats expérimentaux est également fournie pour des données synthétiques et réelles.

• *Cartes topologiques.* Dans [23], les auteurs présentent un formalisme basé sur un clustering collaboratif topologique utilisant des techniques de regroupement fondées sur des prototypes. Les cartes topologiques représentant différents sites sont gérées par des classifieurs distribués collaborant sans avoir à accéder aux données originales, préservant ainsi leur privacité. Deux approches différentes de clustering collaboratif sont présentées : avec collaboration horizontale ou verticale. La force de la collaboration (échange de niveaux de confiance) entre chaque paire d'ensembles de données est déterminée par un paramètre, appelé *coefficient de collaboration*, à estimer itérativement pendant la phase de collaboration en utilisant une optimisation basée sur une descente de gradient. Le procédé comprend deux étapes. Dans la première, l'algorithme SOM (Self-Organizing Map) [31] est appliqué à chaque jeu de données indépendamment. Dans l'étape de collaboration, chacune des cartes est améliorée par échange d'informations sur leur structure.

Dans [20], une autre approche basée sur des cartes topographiques est proposée. L'idée est de combiner un algorithme à base de cartes topographiques génératives bayésiennes (VBGTM) avec un algorithme K-MEANS flou (FCM) pour déterminer les centres ainsi que les clusters en fonction des variables latentes obtenues à partir de VBGTM.

• *Collaboration à partir de méthodes de clustering existantes* Bien qu'efficaces dans de nombreux domaines, les méthodes précédentes ne permettent pas de tirer profil de la variété, de la complémentarité et de la profusion des méthodes de clustering existantes. De nombreux travaux se sont intéressés au développement de *méta-classifieurs* capables de faire collaborer ces classifieurs, dits de base, de façon efficace.

Par exemple, un des pionners de ces approches [19] présente une nouvelle architecture générique de collaboration entre des classifieurs de base pouvant effectuer trois types d'opérations sur leur propre résultat : scission, fusion et suppression de clusters. Ainsi, toute méthode de clustering pourra être à la base d'un classifieur pour peu que ces trois opérations soient définies. Les auteurs montrent d'une part, que définir celles-ci est réalisable pour la majorité des méthodes et, d'autre part, que le schéma proposé en trois phases (détection des conflits, résolution locale des conflits, prise en compte globale) débouche sur des améliorations notables des résultats locaux et globaux. Dans [16], une extension de la méthode permet d'introduire des données exogènes sous forme de contraintes (données labellisées, nombre de clusters ...) pour améliorer le contrôle de la collaboration.

Un autre cadre collaboratif est donné dans [52] dans lequel l'objectif est de résoudre un problème de visualisation multi-point de vue et de clustering alternatif. Des approches d'ensembles et de clustering semi-supervisé sont utilisées pour contrôler les différents classifieurs qui partagent un modèle commun. Le but de la méthode est de parvenir à un consensus ou, en variante, d'améliorer les solutions locales.

Une approche similaire a été proposée par [29]. Les auteurs ont défini un nouveau modèle de clustering coopératif qui implique la coopération entre plusieurs techniques de regroupement pour augmenter l'homogénéité des clusters. Le modèle est capable de gérer des ensembles de données avec différentes propriétés en développant deux structures de données : une représentation par histogrammes des similarités deux-à-deux des données et un graphe de contingence coopératif. Les deux structures de données sont conçues pour trouver les sous-clusters correspondants entre différents clusters recherchés puis d'obtenir une hiérarchie de clusters par un algorithme hiérarchique ascendant.

Plus récemment, [53] a proposé un nouveau cadre collaboratif qui fonctionne avec la plupart des algorithmes de clustering. Le schéma de collaboration est basé sur une collaboration horizontale : tous les classifieurs de base travaillent soit sur des sous-ensembles représentant les mêmes données avec des ensembles d'attributs différents, soit sur les mêmes données mais en cherchant un nombre de clusters différent, soit enfin, un mélange des deux.

• *Données distribuées* En raison de l'augmentation récente du nombre de données collectées automatiquement, un besoin croissant de méthodes efficaces traitant des données distribuées se fait sentir de façon cruciale. Ainsi, [10] définit une approche collaborative qui se concentre sur la découverte de connaissances lorsque seul l'accès à l'ensemble de données locales est possible ou autorisé. Les auteurs présentent deux versions différentes de la méthode : la première pour des données identiques, mais décrites par différents attributs, et une deuxième où les attributs sont identiques, mais les ensembles de données locaux sont différents. L'approche consiste en deux étapes : une étape collaborative basée sur un clustering flou puis une étape d'optimisation par essaims de particules. Pour aborder le problème des données distribuées, [56] a proposé un cadre pour regrouper des classifieurs distribués. Il y est montré que le regroupement de telles classifications distribuées améliore les performances de l'ensemble.

Plus récemment, [9] propose une approche dédiée aux grandes bases de données distribuées. La méthode est basée sur un algorithme collaboratif `diviser-et-conquerir` en utilisant K-MEANS comme classifieur de base. La collaboration consiste à échanger des centres des clusters pour accélérer la convergence dans chaque partition.

6.2 Dimensions du clustering collaboratif

Le tour d'horizon précédent, bien que loin d'être exhaustif, et l'analyse des méthodes décrites permettent néanmoins de mettre en évidence différentes dimensions pouvant être utilisées pour caractériser les méthodes collaboratives :

• *Mono-objectif versus multi-objectif* : La notion d'objectif se rapporte au(x) critère(s) (compacité, nombre de clusters, séparabilité, ...) à optimiser par le processus collaboratif. De nombreuses méthodes imposent de définir un critère d'optimalité Cr_{opt} unique à partir de plusieurs objectifs. Ainsi, par exemple, dans la méthode SAMARAH [19], le critère à optimiser est une combinaison entre la qualité interne moyenne des clusterings Q (e.g. basée sur la compacité moyenne des clusters), la similarité des résultats S et le respect des contraintes globales imposées G (e.g. le nombre de clusters) :

$$Cr_{opt} = p_c.(q.Q + s.S) + p_g.G$$

où p_c et p_g permet de fixer l'importance relative entre la qualité/similarité des clusterings et le respect de contraintes externes, q et s de régler l'importance relative de la qualité des clusterings et leur similarité.
En revanche, dans le cas d'un objectif multiple il s'agit de trouver des clusters dans un ensemble de données en appliquant un ou plusieurs algorithmes correspondant à différents critères d'optimalité. Il n'y a donc plus nécessairement de solution unique meilleure sur tous ces critères simultanément. Le clustering final qui sera alors un compromis entre ces critères de base pourra être choisi en utilisant un front de Pareto.

• *Clustering complet versus partiel* : L'objectif de la collaboration peut être de produire une partition unique de l'ensemble de données (clustering complet) ou il peut aussi s'agir de trouver uniquement des partitions de sous-ensembles de données, qu'ils soient décrits par les mêmes attributs ou non (*clustering partiel*).

• *Mono versus multi-domaine* : Dans le cas, mono-domaine tous les objets appartiennent à la même super-classe d'objets (par exemple, les meubles ou les êtres vivants). Dans le cas multi-domaine, les objets classés par les différentes méthodes locales peuvent appartenir à des superclasses différentes. Par exemple, il pourrait être intéressant de voir de quelle manière et dans quelle mesure les insectes dans une colonie et les citoyens dans un système économique partagent des critères ou des paramètres d'un même modèle.

• *Mono versus multi-échelle* : Il est possible que dans certains cas, les données correspondent aux mêmes éléments, mais sont décrites à différentes échelles. C'est notamment le cas pour les images de télédétection où, en fonction de la résolution, les objets peuvent être vus à différentes échelles mais peuvent également impliquer des objets décrits à différents niveaux d'abstraction. Dans un tel cas, il peut être avantageux de faire collaborer des classifieurs travaillant à ces différents niveaux de description afin de produire des résultats plus structurés rendant éventuellement compte de différents niveaux d'analyse sémantique.

• *Mono versus multi-stratégie* : Dans une approche à stratégie unique, tous les classifieurs utilisent le même algorithme, mais éventuellement paramétré ou initialisé différemment. L'échange d'informations entre eux est de fait facilité. Dans une approche multi-stratégie, les classifieurs n'utilisent pas tous le même algorithme de base. Il est plus difficile de définir (et bénéficier de) la collaboration entre, par exemple, l'algorithme K-means et l'algorithme hiérarchique COBWEB en raison du manque de correspondance directe entre les solutions.

Référence	Objectif	Partition	Domaine	Échelle	Stratégie	Échange
[44]	mono	complet	mono	mono	mono	appartenance
[56]	mono	partiel	multi	mono	multi	résultat
[23]	mono	partiel	multi	mono	mono	résultat
[20]	multi	partiel	multi	mono	mono	résultat
[16]	mono	complet	mono	mono	multi	appartenance
[52]	multi	partiel	multi	multi	multi	résultat
[29]	mono	complet	mono	mono	mono	appartenance
[53]	mono	complet	mono	mono	multi	appartenance
[10]	mono	partiel	multi	mono	multi	résultat
[9]	mono	complet	multi	mono	mono	appartenance

TABLE 1: Caractéristiques des principales méthodes.

• *Échange d'information basé sur l'appartenance des données aux clusters versus d'information sur les résultats* : Une façon d'échanger de l'information entre des classifieurs qui diffèrent dans leur fonctionnement est de communiquer au niveau de l'appartenance de chaque objet aux clusters dans chaque solution locale. Cela n'exige pas que les algorithmes partagent des informations sur leur état interne. Cependant, l'échange d'informations sur ces états internes peut enrichir la communication et accélérer le processus collaboratif. Par exemple, les algorithmes basés sur des distributions de mélange de probabilités pourraient être utiles pour échanger des informations sur les formes de ces distributions.

La table 1 présente les principales méthodes de clustering collaboratif en fonction des caractéristiques présentées.

6.3 Applications

En général, dans la littérature, les méthodes présentées sont majoritairement comparées sur les jeux de données UCI classiques. Une des raisons de ce choix est le coût d'exécution élevé de plusieurs de ces algorithmes, ce qui les rendait jusqu'à récemment difficiles à appliquer à des ensembles de données trop volumineux. Cependant, certaines de ces méthodes ont été utilisées sur des données réelles pour résoudre des problèmes concrets. Par exemple, [35] décrit une méthode de segmentation de mouvements 3-D, basée sur un clustering collaboratif. La segmentation est calculée à partir de deux vues en perspective. Une extension multi-vue de l'algorithme SPARSE SUBSPACE CLUSTERING [13] est proposée pour combiner l'information sur plusieurs images. Dans [15], une étude en recherche marketing est proposée. Les auteurs ont utilisé une optimisation multi-objectif par essaim de particules (MOPSO) dans un cadre collaboratif flou. Les principales contributions de ce travail consistent à donner une méthode pour calculer la matrice de collaboration entre les différents référentiels de données et à proposer des critères d'optimalité au niveau des données et à celui de l'information sur les résultats ce qui permet de construire un consensus entre tous les sites de données. Dans [17], les auteurs présentent une application de

l'approche collaborative SAMARAH à la classification d'images de télédétection d'une zone urbaine via une approche dite orientée-objet . Cette méthode multi-stratégie intègre différents types d'algorithmes de clustering qui collaborent pour produire un résultat consensuel unique. Le document souligne comment de tels classifieurs peuvent collaborer et présente des résultats très intéressants.

7 CONCLUSION ET PERSPECTIVES

Malgré le nombre croissant de méthodes et d'outils dédiés à la classification collaborative non supervisée, ce paradigme est encore étonnamment peu utilisé. Cependant, l'émergence du phénomène « Big Data », avec pour corollaire la nécessité de calculs distribués mais aussi collaboratifs, devrait changer radicalement cette situation. Jusqu'à récemment, les méthodes d'ensembles concernaient essentiellement l'apprentissage supervisé, mais un changement est en cours qui est du en partie au besoin croissant d'explorer des masses de données non étiquetées sans préconceptions claires sur les structures qui peuvent s'y trouver.

Dans cet article, nous avons souligné les différences entre les approches visant à proposer des clusterings de meilleures qualités que ceux potentiellement fournis par une méthode unique. Nous avons donc introduit différents aspects de la combinaison de classifieurs non supervisés : approches multi-vues et de consensus via des schémas coopératifs ou collaboratifs. Dans ce dernier cas, les classifieurs traitent habituellement des ensembles de données différents car soit les instances, soit les descripteurs, soit les deux, diffèrent. En outre, et ceci est essentiel, les algorithmes utilisés par ces classifieurs cherchent à trouver des structures dans les données qui peuvent différer entre elles. Cependant, la force de la collaboration est justement que les classifieurs sont rendus aptes à utiliser des informations provenant d'autres classifieurs afin de découvrir de meilleures structures. Une autre caractéristique clé du schéma collaboratif est que les classifieurs peuvent échanger des informations de façon itérative, et non en prélude à une phase finale de consensus.

Malgré le nombre croissant de travaux sur le clustering collaboratif, il reste beaucoup de pistes à explorer et de nombreuses questions à régler. Dans cet article, nous avons essayé d'exposer un ensemble d'éléments pour caractériser les méthodes et pour concevoir de nouveaux schémas. Mais surtout, nous avons souligné un ensemble de problèmes et de questions auxquels toute méthode collaborative devra faire face. Les plus cruciales de ces dernières porteront sur le type d'informations échangées, le protocole de contrôle des communications et du processus, la façon dont les informations (et les perturbations potentielles) sont prises en compte par chaque classifieur dans un intérêt partagé ou encore, pour ne nommer que les plus importantes, comment définir des critères d'arrêt. Ainsi, bien qu'il existe maintenant un grand nombre d'approches appliquées avec succès à des cas concrets, ce domaine souffre encore d'un manque de formalisation et de compréhension théoriques des conditions nécessaires à une collaboration fructueuse : l'expérience prouve que toutes les collaborations ne le sont pas. La question est donc de définir de façon formelle et non-uniquement empiriquement les conditions pouvant favoriser une collaboration fructueuse : quelles doivent être les liens entre les classifieurs ? Y a-t-il une diversité « minimale » à garantir ? Dans quelle mesure un classifieur doit-il partager son a priori sur les structures de données intéressantes avec les autres classifieurs ? Quelles sont les informations à échanger ? Quand et avec qui ces échanges doivent-ils prendre place ? Existe-t-il un moyen de détecter ou de prévenir les collaborations négatives ? Enfin, dans un cadre distribué et collaboratif, comment définir un critère d'arrêt ? Toutes ces questions, et bien d'autres, sont encore largement ouvertes.

Sans ambition d'exhaustivité, ce papier a tenté de mettre en évidence et d'organiser les principaux problèmes liés au développement de méthodes collaboratives de clustering. C'est un domaine passionnant qui nous semble prometteur pour la résolution des problèmes de découverte de connaissances dans les masses de données. Nous croyons que le moment est venu de promouvoir largement l'utilisation de ces méthodes. De plus, nous sommes persuadés que le clustering collaboratif est aussi un domaine de recherche riche de questions qui vont au-delà de la théorie actuelle de l'apprentissage statistique et que cela devrait stimuler de nouvelles pistes ou domaines de recherche très intéressants.

RÉFÉRENCES

[1] Margareta Ackerman, Shai Ben-David, and David Loker. 2010. Towards property-based classification of clustering paradigms. In *Advances in Neural Information Processing Systems*. 10–18.

[2] Ethem Alpaydin and Cenk Kaynak. 1998. Cascading Classifiers. *Kybernetika* 34, 4 (1998), 369–374.

[3] Hanan Ayad and Mohamed S. Kamel. 2008. Cumulative Voting Consensus Method for Partitions with Variable Number of Clusters. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 30, 1 (2008), 160–173.

[4] Avron Barr, Edward Feigenbaum, and C Roads. 1982. The Handbook of Artificial Intelligence, Volume 1. (1982).

[5] Avrim Blum and Tom Mitchell. 1998. Combining labeled and unlabeled data with co-training. In *Proceedings of the eleventh annual conference on Computational learning theory*. ACM, 92–100.

[6] Guillaume Cleuziou, Matthieu Exbrayat, Lionel Martin, and Jacques-Henri Sublemontier. 2009. CoFKM : A centralized method for multiple-view clustering. In *Data Mining, 2009. ICDM'09. Ninth IEEE International Conference on*. IEEE, 752–757.

[7] Antoine Cornuéjols and Christine Martin. 2015. Unsupervised one class identification by selecting and combining ranking functions. In *Proceedings of the 2015 Conférence sur l'Apprentissage Automatique*.

[8] Antoine Cornuejols, Cédric Wemmert, Pierre Gançarski, and Younès Bennani. 2017. Collaborative Clustering : Why, When, What and How. *Information Fusion* 39 (Apr 2017), 81–95. https://doi.org/10.1016/j.inffus.2017.04.008

[9] Huimin Cui, Gong Ruan, Jingling Xue, Rui Xie, Lei Wang, and Xiaobing Feng. 2014. A Collaborative Divide-and-conquer K-means Clustering Algorithm for Processing Large Data. In *Proceedings of the 11th ACM Conference on Computing Frontiers*. 20 :1–20 :10.

[10] Benoît Depaire, Rafael Falcón, Koen Vanhoof, and Geert Wets. 2011. PSO driven collaborative clustering : A clustering algorithm for ubiquitous environments. *Intelligent Data Analysis* 15, 1 (2011), 49–68.

[11] Thomas G. Dietterich. 2000. Ensemble Methods in Machine Learning. In *Multiple Classifier Systems, LBCS-1857*. Springer, 1–15.

[12] Carlotta Domeniconi and Muna Al-Razgan. 2009. Weighted cluster ensembles : Methods and analysis. *ACM Transactions on Knowledge Discovery from Data (TKDD)* 2, 4 (2009), 17.

[13] Ehsan Elhamifar and Rene Vidal. 2013. Sparse subspace clustering : Algorithm, theory, and applications. *Pattern Analysis and Machine Intelligence, IEEE Transactions on* 35, 11 (2013), 2765–2781.

[14] Fabio Fagnani, Sophie M Fosson, and Chiara Ravazzi. 2014. Consensus-like algorithms for estimation of Gaussian mixtures over large scale networks. *Mathematical Models and Methods in Applied Sciences* 24, 02 (2014), 381–404.

[15] Rafael Falcón, Benoît Depaire, Koen Vanhoof, and Ajith Abraham. 2008. Towards a suitable reconciliation of the findings in collaborative fuzzy clustering. In *Intelligent Systems Design and Applications, 2008. ISDA'08. Eighth International Conference on*, Vol. 3. IEEE, 652–657.

[16] Germain Forestier, Pierre Gançarski, and Cédric Wemmert. 2010. Collaborative clustering with background knowledge. *Data & Knowledge Engineering* 69, 2 (2010), 211–228.

[17] Germain Forestier, Cédric Wemmert, and Pierre Gançarski. 2008. Collaborative Multi-Strategical Clustering for Object-Oriented Image Analysis. In *Studies in Computational Intelligence*, Vol. 126. Springer, 71–88.

[18] Ana L. Fred and Anil K. Jain. 2005. Combining multiple clusterings using evidence accumulation. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 27, 6 (2005), 835–850.

[19] Pierre Gançarski and Cédric Wemmert. 2007. Collaborative Multi-step Mono-level Multi-strategy Classification. *Multimedia Tools and Applications* 35 (Oct 2007), 1–27. https://doi.org/10.1007/s11042-007-0115-x

[20] Mohamad Ghassany, Nistor Grozavu, and Younès Bennani. 2012. Collaborative generative topographic mapping. In *Neural Information Processing*. Springer, 591–598.

[21] Kevin Gimpel and Noah A Smith. 2012. Concavity and initialization for unsupervised dependency parsing. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies*. Association for Computational Linguistics, 577–581.

[22] Aristides Gionis, Heikki Mannila, and Panayiotis Tsaparas. 2007. Clustering aggregation. *ACM Transactions on Knowledge Discovery from Data (TKDD)* 1, 1 (2007), 4.

[23] Nistor Grozavu and Younes Bennani. 2010. Topological collaborative clustering. *Australian Journal of Intelligent Information Processing Systems* 12, 2 (2010).

[24] John A Hartigan. 1972. Direct clustering of a data matrix. *Journal of the american statistical association* 67, 337 (1972), 123–129.

[25] Jeffrey Horn, Nicholas Nafpliotis, and David E Goldberg. 1994. A niched Pareto genetic algorithm for multiobjective optimization. In *Evolutionary Computation, 1994. IEEE World Congress on Computational Intelligence., Proceedings of the First IEEE Conference on*. Ieee, 82–87.

[26] Anil K. Jain. 2010. Data Clustering : 50 Years Beyond K-means. *Pattern Recogn. Lett.* 31, 8 (2010), 651–666.

[27] Yizhang Jiang, Fu-Lai Chung, Shitong Wang, Zhaohong Deng, Jun Wang, and Pengjiang Qian. 2015. Collaborative fuzzy clustering from multiple weighted views. *IEEE transactions on cybernetics* 45, 4 (2015), 688–701.

[28] Erik L. Johnson and Hillol Kargupta. 2000. Collective, hierarchical clustering from distributed, heterogeneous data. In *Large-Scale Parallel Data Mining*. Springer, 221–244.

[29] Rasha Kashef and Mohamed S Kamel. 2010. Cooperative clustering. *Pattern Recognition* 43, 6 (2010), 2315–2329.

[30] Josef Kittler, Mohamad Hatef, Robert P. W. Duin, and Jiri Matas. 1998. On Combining Classifiers. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 20, 3 (1998), 226–239.

[31] Teuvo Kohonen. 1990. The self-organizing map. *Proc. IEEE* 78, 9 (1990), 1464–1480.

[32] Ludmila I. Kuncheva. 2004. *Combining Pattern Classifiers : Methods and Algorithms*. Wiley-Interscience.

[33] Ludmila I. Kuncheva, Stephan T. Hadjitodorov, and Ludmila P. Todorova. 2006. Experimental Comparison of Cluster Ensemble Methods. In *Information Fusion, 2006 9th International Conference on*. 1–7.

[34] Andrea Lancichinetti and Santo Fortunato. 2012. Consensus clustering in complex networks. *Scientific reports* 2 (2012).

[35] Zhuwen Li, Jiaming Guo, Loong-Fah Cheong, and Steven Zhou. 2013. Perspective motion segmentation via collaborative clustering. In *Proceedings of the IEEE International Conference on Computer Vision*. 1369–1376.

[36] Vincenzo Loia, Witold Pedrycz, and Sabrina Senatore. 2007. Semantic Web Content Analysis : A Study in Proximity-Based Collaborative Clustering. *Fuzzy Systems, IEEE Transactions on* 15, 6 (2007), 1294–1312.

[37] David J.C. MacKay. 2003. *Information theory, inference and learning algorithms*. Cambridge University Press.

[38] Sushmita Mitra, Haider Banka, and Witold Pedrycz. 2006. Rough-fuzzy collaborative clustering. *IEEE Transactions on Systems, Man, and Cybernetics* 36 (2006), 795–805.

[39] Stefano Monti, Pablo Tamayo, Jill Mesirov, and Todd Golub. 2003. Consensus clustering : a resampling-based method for class discovery and visualization of gene expression microarray data. *Machine learning* 52, 1-2 (2003), 91–118.

[40] Radford M Neal and Geoffrey E Hinton. 1998. A view of the EM algorithm that justifies incremental, sparse, and other variants. In *Learning in graphical models*. Springer, 355–368.

[41] Nam Nguyen and Rich Caruana. 2007. Consensus Clusterings.. In *International Conference on Data Mining*. IEEE Computer Society, 607–612.

[42] Nils J. Nilsson. 2010. *The quest for artificial intelligence*. Cambridge University Press.

[43] Richard Nock and Frank Nielsen. 2004. An abstract weighting framework for clustering algorithms. In *Proceedings of the 2004 SIAM International Conference on Data Mining*. SIAM, 200–209.

[44] Witold Pedrycz. 2002. Collaborative fuzzy clustering. *Pattern Recognition Letters* 23, 14 (2002), 1675–1686.

[45] Wiltord Pedrycz. 2007. *Knowledge-Based Clustering : From Data to Information Granules* . Wiley-Interscience.

[46] Witold Pedrycz and P. Rai. 2008. A Multifaceted Perspective at Data Analysis : A Study in Collaborative Intelligent Agents. *Systems, Man, and Cybernetics, Part B : Cybernetics, IEEE Transactions on* 38, 4 (2008), 1062–1072.

[47] Robert E. Schapire and Yoav Freund. 2012. *Boosting : Foundations and Algorithms*. MIT Press.

[48] Oliver G. Selfridge. 1958. *Pandemonium : a paradigm for learning in mechanisation of thought processes*. HMSO.

[49] Shai Shalev-Shwartz and Shai Ben-David. 2014. *Understanding Machine Learning : From Theory to Algorithms*. Cambridge University Press.

[50] Valentin I Spitkovsky, Hiyan Alshawi, and Daniel Jurafsky. 2011. Lateen EM : Unsupervised training with multiple objectives, applied to dependency grammar induction. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 1269–1280.

[51] Alexander Strehl and Joydeep Ghosh. 2002. Cluster Ensembles – A Knowledge Reuse Framework for Combining Multiple Partitions. *Journal on Machine Learning Research* 3 (2002), 583–617.

[52] Jacques-Henri Sublemontier. 2013. Unsupervised collaborative boosting of clustering : a unifying framework for multi-view clustering, multiple consensus clusterings and alternative clustering. In *International Joint Conference on Neural Networks (IJCNN 2013)*. Dallas, United States.

[53] Jérémie Sublime, Nistor Grozavu, Younes Bennani, and Antoine Cornuéjols. 2015. Collaborative clustering with heterogeneous algorithms. In *Neural Networks (IJCNN), 2015 International Joint Conference on*. IEEE, 1–8.

[54] Alexander Topchy, Anil K Jain, and William Punch. 2003. Combining multiple weak clusterings. In *Data Mining, 2003. ICDM 2003. Third IEEE International Conference on*. IEEE, 331–338.

[55] Alexander P. Topchy, Anil K. Jain, and William F. Punch. 2003. Combining Multiple Weak Clusterings.. In *International Conference on Data Mining*. IEEE Computer Society, 331–338.

[56] Grigorios Tsoumakas, Lefteris Angelis, and Ioannis Vlahavas. 2004. Clustering classifiers for knowledge discovery from physically distributed databases. *Data & Knowledge Engineering* 49, 3 (2004), 223–242.

[57] Sandro Vega-Pons and José Ruiz-Shulcloper. 2011. A survey of clustering ensemble algorithms. *International Journal of Pattern Recognition and Artificial Intelligence* 25, 03 (2011), 337–372.

[58] Cédric Wemmert and Pierre Gançarski. 2002. A Multi-View Voting Method to Combine Unsupervised Classifications. In *Artificial Intelligence and Applications*. Malaga, Spain, 447–452.

[59] Yang Yan, Lihui Chen, and William-Chandra Tjhi. 2013. Fuzzy semi-supervised co-clustering for text documents. *Fuzzy Sets and Systems* 215 (2013), 74–89.

[60] Zhou Zhi-Hua. 2012. *Ensemble Methods : Foundations and Algorithms*. CRC Press.

[61] Zhi-Hua Zhou and Wei Tang. 2006. Clusterer ensemble. *Knowledge-Based Systems* 19, 1 (2006), 77–83.