Des proportions logiques à un nouvel algorithme de clustering

Myriam Bounhas[1], Henri Prade[2]

[1] LARODEC, ISG, Université de Tunis, Tunisie
[2] IRIT, Univ. Toulouse, Toulouse INP, CNRS, Toulouse, France

myriam_bounhas@yahoo.fr, henri.prade@irit.fr

Résumé

Cet article propose une nouvelle méthode de regroupement (clustering) capable de rivaliser avec des algorithmes comme HDBSCAN ou k-means. L'algorithme regroupe des paires d'éléments plutôt que des éléments isolés et s'appuie sur deux proportions logiques : la paralogie et la paralogie inverse. Ce sont des connecteurs logiques booléens, quaternaires, qui peuvent être étendus à des valeurs numériques. La paralogie évalue dans quelle mesure deux paires d'éléments partagent des caractéristiques similaires, suggérant qu'elles peuvent appartenir au même cluster. En revanche, la paralogie inverse évalue si ces paires sont orthogonales et devraient appartenir à des clusters différents. L'algorithme commence par créer de petits clusters à l'aide d'un indice de paralogie inverse, puis les fusionne progressivement grâce à un indice de paralogie.

Mots-clés

Regroupement, Paralogie, Inverse Paralogie.

Abstract

This article proposes a new clustering method capable of competing with algorithms such as HDBSCAN or k-means. The algorithm clusters pairs of elements rather than isolated elements and is based on two logical proportions : paralogy and inverse paralogy. These are Boolean, quaternary logical connectors that can be extended to numerical values. Paralogy assesses the extent to which two pairs of elements share similar characteristics, suggesting that they may belong to the same cluster. In contrast, inverse paralogy assesses whether these pairs are orthogonal and should belong to different clusters. The algorithm begins by creating small clusters using a reverse paralogy index, then gradually merges them using a paralogy index.

Keywords

Clustering, Paralogy, Inverse Paralogy.

1 Introduction

La formalisation logique du concept de proportions analogiques, qui sont des énoncés de la forme « a est à b comme c est à d », a conduit à la définition des proportions logiques [8]. Les proportions logiques comparent la similitude / dissemblance des éléments d'une paire avec la similitude / dissemblance des éléments d'une autre paire. Parmi les exemples de telles proportions, on peut citer les proportions analogiques qui expriment que « a diffère de b comme c diffère de d et b diffère de a comme d diffère de c », ainsi que d'autres proportions remarquables.

Parmi celles-ci, celle appelée « paralogie » présente un intérêt particulier : elle exprime que « a et b sont similaires comme c et d sont similaires », où la similitude peut faire référence à des caractéristiques présentes ou absentes, établissant ainsi un parallèle entre les paires (a, b) et (c, d). Une autre, appelée « paralogie inverse » [9], exprime plutôt une forme d'orthogonalité entre les paires : « ce que a et b ont en commun, c et d ne l'ont pas, et vice versa ».

Ces deux proportions logiques nous rappellent ce qui caractérise l'idée de cluster. En effet, en apprentissage non supervisé, les techniques de regroupement visent à structurer un ensemble de données en groupes de telle sorte que les éléments d'un groupe aient plus en commun entre eux qu'avec les autres éléments, c'est-à-dire qu'ils soient plus similaires entre eux qu'avec ceux d'autres groupes. Cela suggère la possibilité de construire un algorithme de regroupement qui repose sur l'utilisation de ces proportions. C'est ce que nous explorons dans cet article.

Le document est organisé comme suit. Après une brève présentation des proportions logiques en Section 2, la notion plus ciblée de comparateur de paires est introduite en Section 3 et appliquée à la paralogie et à la paralogie inverse, d'abord dans un cadre booléen, puis étendue en logique multivalente. La Section 4 présente un nouvel algorithme de regroupement basé sur la paralogie[1], et la Section 5 rend compte d'expériences ayant donné des résultats compétitifs. La Section 6 compare les résultats obtenus par l'approche proposée à ceux obtenus avec HDBSCAN et la méthode $k\text{-}means$.

2 Proportions logiques : une brève présentation

Les proportions logiques booléennes [9, 11] sont des connecteurs T reliant quatre variables booléennes, disons a, b, c, d, qui lient par une équivalence deux indicateurs de comparaison entre a et b à deux indicateurs de comparaison entre c et d. Les indicateurs de comparaison sont notés

1. Un premier algorithme exclusivement basé sur la paralogie inverse a été présenté dans [2] et ses performances y sont comparées à l'approche $k\text{-}means$ uniquement.

ci dans l'expression (*) ci-après. Plus précisément, une proportion logique T est de la forme

$$T(a,b,c,d) = [ci_1(a,b) \equiv ci_2(c,d)] \wedge [ci_3(a,b) \equiv ci_4(c,d)] \quad (*)$$

où chacun des ci_i est l'un des quatre indicateurs de comparaison booléenne possibles, à savoir $ci_i(a,b)$ est l'un quelconque parmi $a \wedge b$, $\neg a \wedge \neg b$, $\neg a \wedge b$ ou $a \wedge \neg b$ (avec $ci_1 \neq ci_3$ ou $ci_2 \neq ci_4$ pour s'assurer que la conjonction n'est pas triviale). Notez que les deux premiers indicateurs, $a \wedge b$ et $\neg a \wedge \neg b$, évaluent respectivement la similarité positive et négative, tandis que les deux derniers $\neg a \wedge b$ et $a \wedge \neg b$ sont des indicateurs de dissemblance. Prenons les exemples de la paralogie et de la paralogie inverse :

$$P(a,b,c,d) = [(a \wedge b) \equiv (c \wedge d)] \wedge [(\neg a \wedge \neg b) \equiv (\neg c \wedge \neg d)]$$

$$I(a,b,c,d) = [(a \wedge b) \equiv (\neg c \wedge \neg d)] \wedge [(\neg a \wedge \neg b) \equiv (c \wedge d)]$$

La paralogie P stipule clairement que « ce que a et b ont en commun, c et d l'ont également, positivement ou négativement ». La paralogie inverse I affirme le contraire : « lorsque a et b sont vrais, c et d sont faux, et vice-versa ». Comme toute proportion logique, P et I ne sont vraies que pour 6 valuations (parmi les 2^4 possibles) [10], données dans la Table 1.

TABLE 1 –Tables de vérité : Paralogie (**P**), Paralogie Inverse (**I**)

P				**I**			
a	b	c	d	a	b	c	d
0	0	0	0	0	0	1	1
1	1	1	1	1	1	0	0
0	1	1	0	0	1	1	0
1	0	0	1	1	0	0	1
0	1	0	1	0	1	0	1
1	0	1	0	1	0	1	0

Il convient de noter que la paralogie est étroitement liée à la proportion analogique A par une simple permutation, puisque nous avons $A(a,b,c,d) \equiv P(a,d,c,b)$ [11].

3 Comparateurs de paires

Les paires peuvent être comparées non seulement à l'aide des équivalences $\equiv$ comme dans (*), mais aussi à l'aide du connecteur opposé, l'opérateur du « ou exclusif » $\oplus$. Ainsi, si nous substituons $\oplus$ à $\equiv$ dans (*), T devient

$$\mathcal{S}^{\oplus}(T)(a,b,c,d) = [ci_1(a,b) \oplus ci_2(c,d)] \wedge [ci_3(a,b) \oplus ci_4(c,d)] \quad (**)$$

Pour une étude systématique des comparateurs de paires, obtenus par cette substitution à partir des proportions logiques, le lecteur pourra consulter [12].
Si nous appliquons la substitution $\mathcal{S}^{\oplus}$ à P et I, nous obtenons :
$\underline{\underline{P}}(a,b,c,d) = [(a \wedge b) \oplus (\neg c \wedge \neg d)] \wedge [(\neg a \wedge \neg b) \oplus (c \wedge d)]$
$\underline{\underline{I}}(a,b,c,d) = [(a \wedge b) \oplus (c \wedge d)] \wedge [(\neg a \wedge \neg b) \oplus (\neg c \wedge \neg d)] = [(a \wedge b) \oplus (c \wedge d)] \wedge [(a \vee b) \oplus (c \vee d)]$
où

$\underline{\underline{P}}(a,b,c,d) = \mathcal{S}^{\oplus}(I)(a,b,c,d)$,
$\underline{\underline{I}}(a,b,c,d) = \mathcal{S}^{\oplus}(P)(a,b,c,d)$.

Les tables de vérité de $\underline{\underline{P}}$ et $\underline{\underline{I}}$ sont données dans la Table 2, où seuls les deux valuations qui les rendent vrais apparaissent, tandis que les 14 autres valuations possibles rendent ces connecteurs faux.

TABLE 2 – Tables de vérité de $\underline{\underline{P}}$ et $\underline{\underline{I}}$

$\underline{\underline{P}}$				$\underline{\underline{I}}$			
a	b	c	d	a	b	c	d
0	0	0	0	0	0	1	1
1	1	1	1	1	1	0	0

Comme on peut le voir, $\underline{\underline{P}}$ et $\underline{\underline{I}}$ sont beaucoup plus ciblés que P et I, puisqu'ils ne conservent que les deux premières lignes de la Table 1, omettant les évaluations communes à P et I et ne conservant que les valuations qui expriment leur véritable identité.
Les expressions booléennes de P et I peuvent être étendues à des connecteurs à valeurs multiples avec des valeurs dans l'intervalle $[0,1]$. Nous utilisons les connecteurs de Łukasiewicz [5], à savoir

- $\neg x = 1 - x$;
- $x \wedge y = \min(x,y)$; $x \vee y = \max(x,y)$;
- $x \to y = \min(1, 1 - x + y)$;
- $x \equiv y = (x \to y) \wedge (y \to x) = 1 - |x - y|$;
- $x \oplus y = \neg(x \equiv y) = |x - y|$.

Cela conduit aux extensions suivantes pour $\underline{\underline{P}}$ et $\underline{\underline{I}}$, en conservant *la même structure de comparaison par paires* : Pour $(x,y,z,t) \in [0,1]^4$,

$$\underline{\underline{P}}(x,y,z,t) = \min(|\min(x,y) - \min(1-z, 1-t)|, |\min(1-x, 1-y) - \min(z,t)|) \quad (1)$$

De façon similaire, on obtient pour $\underline{\underline{I}}$

$$\underline{\underline{I}}(x,y,z,t) = \min(|\min(x,y) - \min(z,t)|, |\max(x,y) - \max(z,t)|) \quad (2)$$

Il existe une autre façon d'écrire ces deux connecteurs qui ne sont vrais que pour les deux valuations de la Table 2 :

$$\underline{\underline{P}}(a,b,c,d) = (a \equiv b) \wedge (c \equiv d) \wedge (a \equiv c) \wedge (b \equiv d)$$

$$\underline{\underline{I}}(a,b,c,d) = (a \equiv b) \wedge (c \equiv d) \wedge (a \equiv \neg c) \wedge (b \equiv \neg d)$$

La structure des formules ci-dessus diffère clairement de celle obtenue à partir de (**), bien qu'elles aient les mêmes tables de vérité : les expressions ne se concentrent pas uniquement sur les paires (a,b) et (c,d), mais impliquent également les paires (a,c) et (b,d). L'expression pour $\underline{\underline{P}}$ s'inscrit dans l'esprit de la vision en termes de parallélogramme des proportions analogiques [13] (considérant les paires comme des côtés). La condition $b \equiv d$ est superflue ($\equiv$

est transitif), mais a été ajoutée pour un traitement égal des littéraux. Cela conduit aux extensions graduelles suivantes :

$$\underline{\underline{P}}(x,y,z,t) = 1-\max(|x-y|,|z-t|,|x-z|,|y-t|) \quad (3)$$

$$\underline{\underline{I}}(x,y,z,t) = 1-\max(|x-y|,|z-t|,|x-1+z|,|y-1+t|) \quad (4)$$

Notons que (2) et (3) n'impliquent pas de négation interne, contrairement à (1) et (4) : nous privilégions donc leur utilisation dans la pratique. $\underline{\underline{P}}$ diminue dès qu'une valeur s'écarte des autres.

La paralogie et l'orthogonalité s'étendent aux *paires de vecteurs*, avec n composantes valuées dans $[0,1]$:

$$\mathcal{P}((\vec{x},\vec{y}),(\vec{z},\vec{t})) = \sum_{i=1}^{n} \underline{\underline{P}}(x_i,y_i,z_i,t_i)/n \quad (5)$$

et

$$\mathcal{O}((\vec{x},\vec{y}),(\vec{z},\vec{t})) = \sum_{i=1}^{n} \underline{\underline{I}}(x_i,y_i,z_i,t_i)/n. \quad (6)$$

4 Algorithme de regroupement basé sur la paralogie

Nous présentons une nouvelle procédure de regroupement qui opère sur des paires de points plutôt que sur des points individuels. Elle utilise la paralogie $\mathcal{P}$ entre les paires (formule (5)) et l'orthogonalité $\mathcal{O}$ entre les paires (formule (6)) , définies à la fin de la Section 3. $\mathcal{P}$ estime la cohérence structurelle entre deux paires, et $\mathcal{O}$ leur séparation.

Notations et préliminaires. Soit $\mathcal{X} = \{x_1,\ldots,x_N\} \subset \mathbb{R}^d$ l'ensemble de données. Soit $\mathbb{P} = \{(a^j,b^j)\}_{j=1}^m$ un ensemble de paires de points (représentant des données), généré dans une étape préliminaire. Chaque paire $p = (a^j,b^j) \in \mathbb{P}$ est définie de telle sorte que b^j soit le plus proche voisin de a^j par rapport à la distance euclidienne, et que les doublons ou chevauchements de paires soient évités. À une itération donnée, le regroupement est représenté par une partition $\mathcal{C} = \{C_1,\ldots,C_k\}$.

4.1 Procédure de base

L'algorithme fonctionne en créant, ajustant et consolidant, de manière répétée, des (petits) clusters intermédiaires. Ce processus implique : (i) la détection de paires de points qui présentent une forte orthogonalité, (ii) placer ces paires dans des mini-clusters initiaux et provisoires, et (iii) fusionner progressivement les mini-clusters dont les contenus sont mutuellement cohérents en termes de paralogie, ce qui garantit "la cohérence relationnelle" en termes de "similarité".

La consolidation des clusters obéit à une règle de préservation de la cohérence régie par un paramètre de tolérance α spécifié par l'utilisateur, garantissant que seuls les mini-clusters mutuellement compatibles sont fusionnés. Nous fournissons maintenant une description détaillée de chaque étape.

Prétraitement des données. Soit $\mathcal{X}$ un ensemble de données de points dans un espace à N dimensions. Dans un premier temps, toutes les caractéristiques sont mises à l'échelle afin que leurs valeurs se situent dans des plages comparables. Cette normalisation est essentielle pour garantir une utilisation homogène des indices de paralogie et d'orthogonalité entre les dimensions. Ensuite, l'ensemble $\mathbb{P}$ des paires de points les plus proches est généré.

Calcul des matrices d'orthogonalité et de paralogie. Au cours d'une phase préliminaire hors ligne, l'indice d'orthogonalité est évalué pour chaque paire ordonnée de paires $(p^j,q^j) \in \mathbb{P}\times\mathbb{P}$ à l'aide de la définition donnée à la fin de la section 3. Les valeurs obtenues sont stockées dans la matrice d'orthogonalité $\boldsymbol{\mathcal{O}}$. Une procédure analogue est effectuée pour calculer et stocker les indices de paralogie correspondants dans la matrice de paralogie $\boldsymbol{\mathcal{P}}$.

Initialisation des mini-clusters. Cette étape commence par la sélection des deux paires qui présentent le plus haut degré d'orthogonalité selon la matrice $\boldsymbol{\mathcal{O}}$. Ces paires définissent les deux premiers mini-clusters et leurs centroïdes initiaux, représentant la configuration la plus dissemblable sur le plan structurel. D'autres mini-clusters sont introduits progressivement en identifiant les paires qui sont suffisamment orthogonales par rapport à tous les centroïdes existants. Une paire $p \in \mathbb{P}$ est choisie comme nouveau centroïde lorsque son orthogonalité par rapport à chaque centroïde actuel dépasse un seuil donné β (β est maintenu faible, par exemple 0,1).

Affectation des paires aux clusters. Chaque paire restante non affectée $p \in \mathbb{P}$ est successivement attribuée au mini-cluster dont le centroïde donne la plus petite valeur d'orthogonalité avec p. Le recours à des comparaisons basées sur les centroïdes réduit considérablement la charge de calcul en évitant les comparaisons exhaustives par paires au sein des clusters.

Étape de raffinement. Régulièrement, les centroïdes des mini-clusters sont recalculés en faisant la moyenne des coordonnées des points qui constituent les paires attribuées à chaque cluster. Cette mise à jour est effectuée après qu'une fraction spécifiée ($level\ l$) du total des paires a été traitée (par exemple, $l = 50\%$ ou 75%).

Fusion des mini-clusters. Une fois que toutes les paires ont été attribuées à des mini-clusters, une phase de fusion est appliquée afin de réduire la sur-segmentation et de produire des clusters plus stables. Cette phase suit une stratégie agglomérative contrainte par un critère de préservation de la cohérence. Soit $\mathcal{C} = \{C_1,\ldots,C_k\}$ l'ensemble actuel de mini-clusters et soit $\mathbb{P}_i \subset \mathbb{P}$ le sous-ensemble de paires dont les deux points appartiennent au cluster C_i.

Cohérence intra-cluster. Pour chaque mini-cluster C_i, un score de cohérence intra-cluster est défini comme la paralogie moyenne entre toutes les paires distinctes de l'ensenble

$\mathbb{P}_i$ appartenant au cluster C_i :

$$\mathrm{coh}_{\mathrm{intra}}(C_i) = \frac{1}{|\mathbb{P}_i|(|\mathbb{P}_i|-1)} \sum_{\substack{p,q \in \mathbb{P}_i \\ p \neq q}} \mathcal{P}(p,q), \qquad (7)$$

où $\mathcal{P}$ est la mesure de paralogie entre deux paires. Ce score quantifie l'homogénéité structurelle interne de C_i.

Étant donnés deux mini-clusters distincts C_i et C_j, leur similitude est approximée par la paralogie entre leurs centroïdes respectifs :

$$\mathrm{coh}_{\mathrm{inter}}(C_i, C_j) \approx \mathcal{P}(\mu_i, \mu_j), \qquad (8)$$

où μ_i et μ_j désignent les centroïdes (paires de points) associés aux clusters C_i et C_j. Cette approximation basée sur les centroïdes évite le coût computationnel lié à l'évaluation de la paralogie entre toutes les paires de tous les clusters.

Critère de fusion. Une fusion entre les clusters C_i et C_j est considérée comme admissible si elle ne détériore pas la cohérence interne de chaque cluster individuel, c'est-à-dire si la similarité inter-clusters dépasse la cohérence interne maximale entre les deux clusters individuels (avec un certain degré de tolérance α) :

$$\mathrm{coh}_{\mathrm{inter}}(C_i, C_j) > \alpha * \max\left(\mathrm{coh}_{\mathrm{intra}}(C_i), \mathrm{coh}_{\mathrm{intra}}(C_j)\right) \qquad (9)$$

où $\alpha > 0$ est un paramètre de tolérance défini par l'utilisateur qui contrôle la rigueur de la préservation de la cohérence.

Parmi toutes les paires de clusters admissibles, on choisit la fusion qui maximise $\mathrm{coh}_{\mathrm{inter}}$. Après la fusion, les étiquettes des clusters sont mises à jour, le centroïde du cluster fusionné est recalculé, ainsi que sa cohérence intra-cluster. La procédure est répétée de manière itérative jusqu'à ce qu'il n'y ait plus de fusion admissible.

Arrêt. L'algorithme se termine lorsqu'aucune paire de mini-clusters ne satisfait au critère de fusion préservant la cohérence. Le résultat final comprend le nombre final de clusters $k^\star$, les centroïdes finaux des clusters $\mu^\star$ et les affectations finales des clusters $\mathcal{C}^\star$ pour tous les points .

4.2 Algorithme

L'algorithme 1 décrit la procédure principale de l'approche proposée. Cet algorithme appelle la fonction $init\_centroids$ pour initialiser les mini-clusters et la fonction $merge\_clusters$ pour les fusionner lorsque cela est nécessaire, comme expliqué ci-dessus.

Algorithm 1 Paralogy, un Algorithme de Clustering

Input : $\mathcal{X}$: Données, β : seuil init. , α : seuil fusion.
Output : $k^\star, \mathcal{C}^\star, \mu^\star$: nb final de clusters, clusters, et centroïdes.

$\mathcal{X} \leftarrow$ Normalize($\mathcal{X}$) {Prétraitement}
$\mathbb{P} \leftarrow \{(a,b) \mid a,b \in \mathcal{X} \times \mathcal{X}, b =$ Nearest-Neighbor$(a)\}$
$\boldsymbol{\mathcal{O}} \leftarrow \{\mathcal{O}_{i,j} \mid p_i, p_j \in \mathbb{P}^2, \mathcal{O}_{i,j} = \mathcal{O}((a,b),(c,d))\}$
$\boldsymbol{\mathcal{P}} \leftarrow \{\mathcal{P}_{i,j} \mid p_i, p_j \in \mathbb{P}^2, \mathcal{P}_{i,j} = \mathcal{P}((a,b),(c,d))\}$ {Matrices}
$k, \mathcal{C}, \mu \leftarrow$ init_centroids($\boldsymbol{\mathcal{O}}$) {Initialisation}
for each unclustered $p = (a,b) \in \mathbb{P}$ **do** {Assignation}
  $C \leftarrow \arg\min_j(\mathcal{O}((a,b),(c,d))|\mu_j = (c,d),\ j = 1,\ldots,k)$; C.append(p)
end for
if $|\text{Clustered}(\mathbb{P})|/|\mathbb{P}| > l$ **then** UpdateCentroids(μ) {Refine}
$k^\star, \mathcal{C}^\star, \mu^\star \leftarrow$ merge_clusters($k, \mathcal{C}, \mu, \alpha$) {Fusion}
return $k^\star$, $\mathcal{C}^\star$, $\mu^\star$

5 Résultats expérimentaux

Cette Section présente une analyse des résultats empiriques obtenus à l'aide de la méthode proposée de *regroupement basé sur la paralogie* sur des ensembles de données synthétiques ou réels. Nous comparons également l'efficacité de cette méthode avec l'algorithme HDBSCAN [4, 3].

Nous expérimentons une variété d'ensembles de données présentant des caractéristiques différentes : un ensemble de données synthétiques (Blobs) et trois ensembles de données réelles (Iris, Zoo et Breast Cancer).

Pour évaluer les deux méthodes de regroupement comparées, nous utilisons trois mesures d'évaluation : le *score Silhouette (Sil)*, l'*indice de Calinski-Harabasz (CH)* et le *score de Davies-Bouldin (DB)* pour les ensembles de données d'apprentissage non supervisé (Blobs), et nous appliquons l'*Adjusted Rand Index (ARI)* plus adapté aux trois ensembles de données d'apprentissage supervisé. Des valeurs plus élevées (proches de 1 pour Sil et ARI) pour Sil, CHI et ARI sont préférables, tandis que des valeurs plus faibles (proches de 0) pour DB sont préférables. Les résultats sont présentés respectivement dans en Figure 1 pour Blobs, en Figure 2 pour Iris, en Figure 3 pour Zoo et en Figure 4 pour les ensembles de données sur le cancer du sein. Ces résultats correspondent aux meilleurs paramètres obtenus pour chaque algorithme. Pour le regroupement basé sur la paralogie, nous affichons les centroïdes des mini-clusters (sous forme d'étoiles noires) afin de suivre le processus de fusion. Nous indiquons d'abord les performances de l'algorithme proposé pour chaque ensemble de données.

5.1 Blobs - 5 amas gaussiens isotropes

Dans les ensembles de données Blobs, tout point d'un cluster est plus proche de son centroïde que du centroïde de tout autre cluster. L'algorithme Paralogy obtient des mesures de validation meilleures (Sil = 0,80, CHI = 10433,6,

DB = 0,27) que HDBSCAN (Sil= 0,76, CHI = 8404,6, DB = 0,98), ce qui indique un meilleur regroupement. Alors que HDBSCAN reste sensible aux variations de densité locales et introduit un certain bruit (2 points noirs), l'algorithme Paralogy préserve l'intégrité des clusters en s'appuyant sur la cohérence relationnelle plutôt que sur les écarts de densité. Les centroïdes s'alignent correctement avec la structure réelle des blobs, démontrant ainsi sa robustesse pour cet ensemble de données.

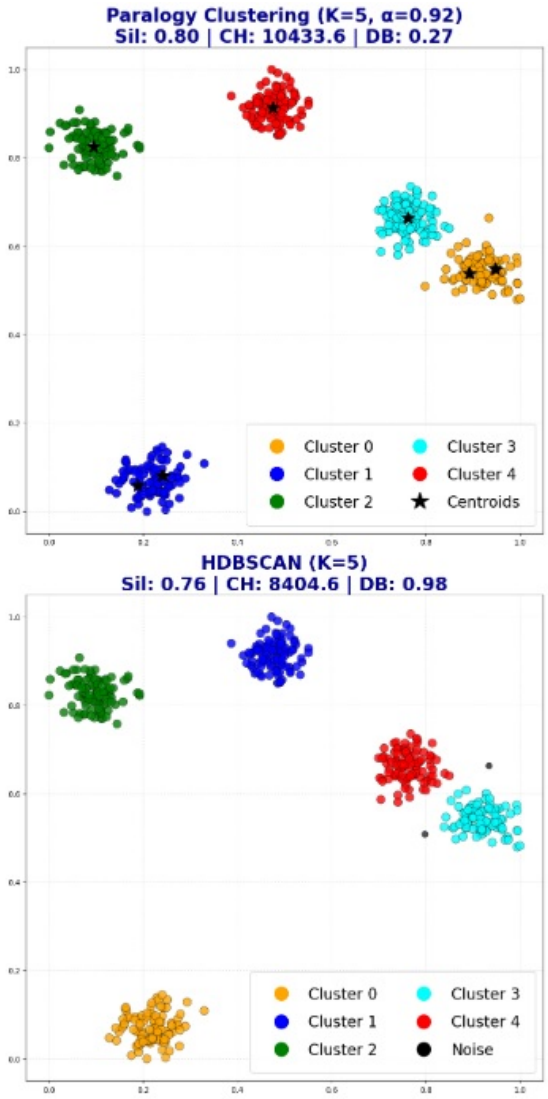

FIGURE 1 – Résultats pour l'ensemble de données Blobs

5.2 Iris - Relations d'échelle allométriques

L'ensemble de données Iris contient 3 classes, dont l'une est linéairement séparable des deux autres ; ces dernières ne sont pas linéairement séparables l'une de l'autre. L'algorithme Paralogy obtient un score ARI beaucoup plus élevé (0,886) que HDBSCAN (0,568). Bien que les espèces Versicolor et Virginica se chevauchent fortement dans l'espace, chaque espèce suit une relation de croissance allométrique distincte entre la longueur et la largeur des pétales. Même lorsque les échantillons sont proches en distance absolue, leurs taux de croissance locaux diffèrent. L'algorithme Paralogy capture ces caractéristiques invariables et réussit à séparer les espèces, tandis que HDBSCAN traite le chevauchement comme une seule région à haute densité, empêchant ainsi la séparation. pour cet ensemble de données.

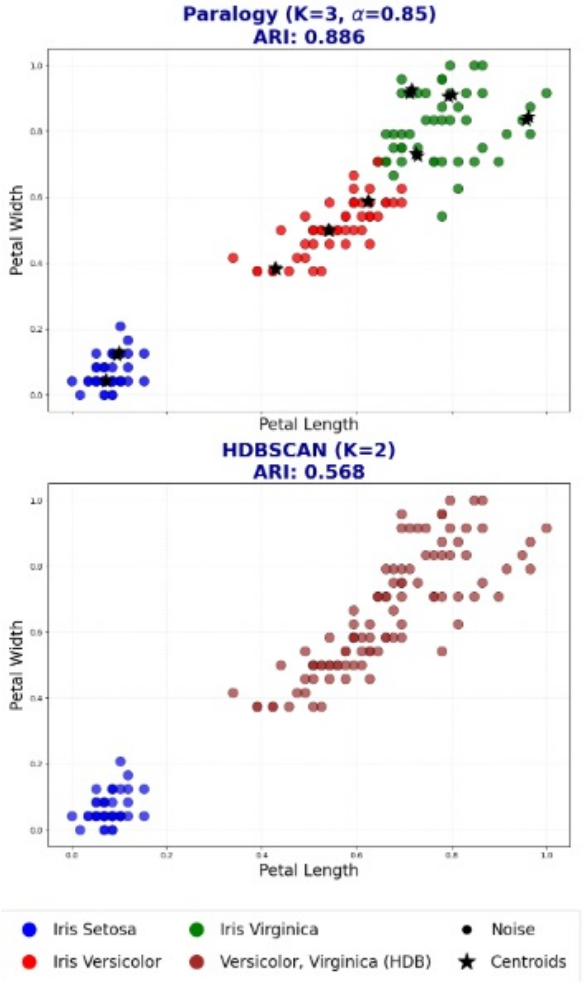

FIGURE 2 – Résultats pour l'ensemble de données Iris

5.3 Zoo - Espace discret avec descriptions booléennes

Dans cet ensemble de données, 101 animaux appartenant à 7 classes sont décrits à l'aide de 16 attributs booléens. L'algorithme Paralogy surpasse HDBSCAN avec un ARI de 0,868 contre 0,475 (les écarts de densité sont mal définis). Paralogy maintient la cohérence relationnelle au sein des classes biologiques. Cette approche préserve correctement le groupe des mammifères, même pour des anomalies comme l'ornithorynque, qui est regroupé correctement grâce à des traits décisifs (poils, lait) malgré ses caractéristiques reptiliennes (voir le point brun au centre). De plus, Paralogy réussit à démêler les groupes qui se chevauchent comme les amphibiens et les reptiles, en identifiant huit clusters cohérents. En revanche, HDBSCAN a du mal dans cet espace discret ; il fragmente la classe des mammifères, crée des clusters supplémentaires arbitraires, fusionne inutilement les amphibiens et les reptiles et étiquette à tort comme du bruit des échantillons isolés mais logiquement cohérents.

5.4 Breast Cancer Wisconsin

Dans cet ensemble de données à deux classes (malignes vs. bénignes) avec de multiples dimensions (30 caractéristiques) présentant un chevauchement clair des clusters, Paralogy (ARI = 0,501) surpasse largement HDBSCAN (ARI = 0,168). Au lieu d'effectuer un regroupement basé sur la densité, Paralogy capture les lois de corrélation stables entre les caractéristiques des noyaux cellulaires (par exemple, le rayon et la surface), ce qui nous permet de séparer les échantillons qui se chevauchent spatialement mais qui diffèrent relationnellement. Il maintient la continuité des classes malgré la présence de bruit et identifie un petit sous-groupe cohérent supplémentaire (Extra Cluster 1) qui peut représenter un sous-groupe spécial cohérent. En revanche, HDBSCAN fragmente la structure et étiquette de nombreux points ambigus comme du bruit.

6 Discussion / Comparaison avec HDBSCAN et k-means

Contrairement à HDBSCAN, qui dépend de la densité des points et peut échouer lorsque les écarts de densité entre les groupes sont faibles ou inexistants (comme dans le cas d'Iris ou de Breast Cancer), Paralogy privilégie la cohérence relationnelle plutôt que la proximité spatiale. En passant d'une approche basée sur la proximité à un paradigme structurel relationnel, il construit des clusters en fonction de caractéristiques communes. Cela lui permet de préserver l'intégrité des groupes indépendamment de la densité locale, en intégrant les points isolés ou en séparant les groupes qui se chevauchent, à condition qu'ils suivent une logique interne cohérente. En outre, en cas d'ambiguïté, HDBSCAN peut laisser de nombreux points en dehors des clusters en tant que bruit, ce qui n'est pas le cas de la méthode basée sur la paralogie qui affecte tous les points à des clusters.

Contrairement à k-$means$ [7, 6], qui nécessite de spécifier à l'avance le nombre de clusters k (par exemple, via la méthode "Elbow"), Paralogy estime le nombre de clusters directement pendant le processus de clustering, sans aucune procédure supplémentaire. Les résultats préliminaires indiquent que Paralogy surpasse nettement k-$means$ sur les ensembles de données Blobs et Iris (pour Blobs, k-$means$ ne montre que trois clusters au lieu de cinq avec des scores Sil = 0,74, CH = 1925,4, DB = 0,43 et pour Iris seulement 2 clusters avec ARI= 0,568) et reste compétitif sur le jeu de données Moons, où les méthodes basées sur la densité capturent mieux les deux demi-lunes.

Pour traiter les structures non convexes (par exemple Moons), nous avons développé une variante de l'algorithme Paralogy. Au lieu d'estimer la cohérence inter-clusters uniquement à partir des scores de paralogie des centroïdes, nous (i) définissons 2 ou 3 paires de limites au sein de chaque mini-cluster et (ii) calculons la cohérence inter-clusters comme le score de paralogie maximal sur les paires de limites. Le nouveau critère de fusion garantit que seuls les mini-clusters présentant une forte cohérence des limites sont fusionnés. Les résultats montrent que cette stra-

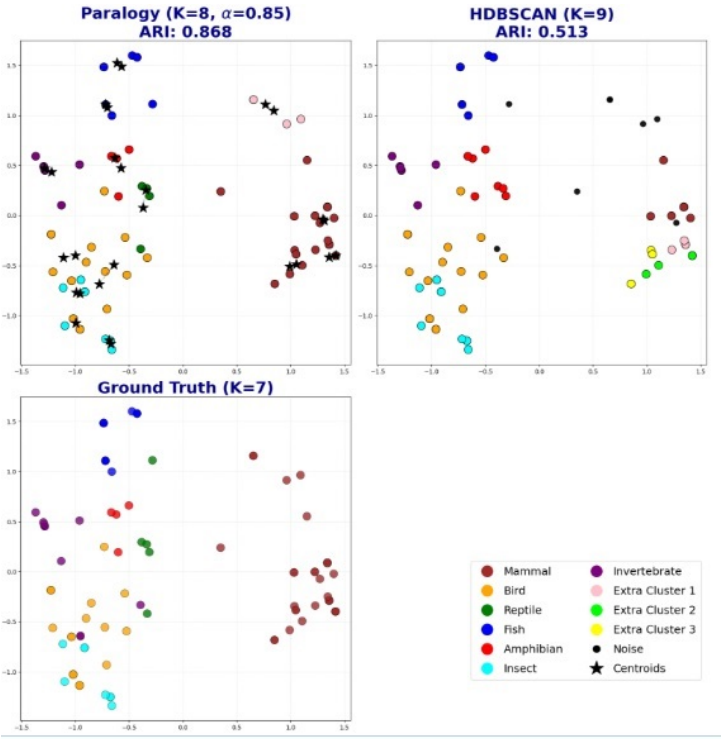

FIGURE 3 – Résultats pour l'ensemble de données Zoo

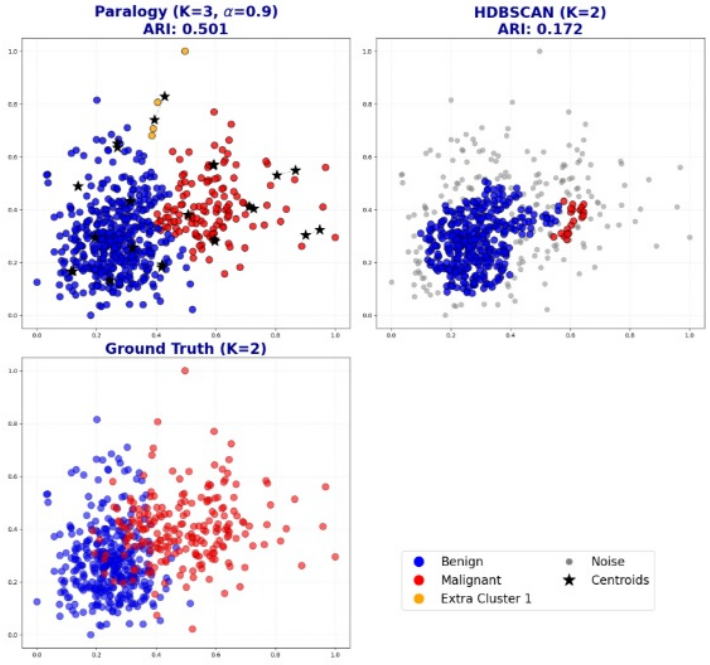

FIGURE 4 – Résultats pour l'ensemble de données Breast Cancer

tégie identifie les deux structures en demi-lune dans l'ensemble de données Moons. L'évaluation de son efficacité sur d'autres ensembles de données fera l'objet de travaux futurs.
Les expérimentations à l'appui de la discussion ci-dessus sont données en Section 8.

7 Conclusion

Une analyse logique de l'idée de cluster a conduit à une nouvelle méthode de regroupement. Le regroupement basé sur la paralogie introduit une nouvelle perspective en apprentissage non supervisé en identifiant les clusters grâce à la cohérence relationnelle plutôt qu'en se basant uniquement sur la densité ou la distance.
Dans l'ensemble, les résultats expérimentaux suggèrent que cette approche est particulièrement intéressante pour les données dont l'organisation est régie par des relations sémantiques, permettant l'émergence de clusters significatifs même en l'absence de séparations claires en termes de densité. Les méthodes basées sur la densité, telles que HDBSCAN, restent compétitives pour les structures essentiellement définies par la densité. Plutôt que de se faire concurrence directement, les deux approches doivent être considérées comme complémentaires : l'une découvre les clusters comme des régions denses, l'autre comme des groupes régis par des relations sémantiques locales cohérentes.
Il serait aussi intéressant d'examiner si les idées de « clusters flous », à la base de l'algorithme « fuzzy c-means » [1], peuvent être introduites avec profit dans l'approche présentée ici.

8 Annexe : Résultats supplémentaires concernant la discussion de la Section 6

Dans cet annexe, nous fournissons les figures supplémentaires qui appuient les affirmations faites dans la Section discussion : i) concernant k-means pour les jeux de données Blobs et Iris et ii) concernant la variante de l'algorithme Paralogy exploitant des paires de frontière sur Moons.

8.1 Comparaison avec k-means

Les résultats de l'étude comparative avec l'algorithme k-means sont présentés dans la Figure 5 pour Blobs et la Figure 6 pour Iris. Nous rappelons d'abord que k-means exploite la méthode du coude (Elbow) pour estimer le nombre de clusters. Comme indiqué précédemment dans la Section 5, la Paralogy montre une forte capacité à détecter le nombre correct de clusters pour le jeu de données Blobs. En particulier, l'approche proposée surpasse k-means pour toutes les métriques utilisées. En revanche, la méthode du Elbow sous-estime le nombre de clusters, ce qui force par conséquent k-means à fusionner « inutilement » certains clusters distincts.
De plus, comme illustré dans la Figure 6, la Paralogy présente une capacité plus élevée à retrouver la structure intrinsèque du jeu de données Iris ($k = 3$, ARI = 0.886) comparé à k-means lorsque ce dernier est contraint par le critère de sélection du Elbow. Bien que l'heuristique du Elbow identifie $k = 2$ comme optimal, ce choix conduit à une sous-estimation claire du nombre réel de clusters, forçant à regrouper les espèces Versicolor et Virginica dans une seule partition (ARI = 0.568).

8.2 Paralogy : Stratégies Centroïde vs Frontière

Pour traiter des structures non convexes, dans l'expérience Moons (Figure 7), nous avons développé une variante de l'algorithme de paralogie exploitant des paires de frontière. Au lieu d'utiliser uniquement les scores de paralogie des centroïdes, nous mesurons la cohérence inter-clusters par le score maximal entre les paires de frontière de chaque mini-cluster. Dans la Figure 7, les points jaunes indiquent les paires de frontière utilisées lors de la dernière fusion.
Les deux stratégies — basée sur le centroïde et sur les paires de frontière — cherchent à capturer la séparation entre les deux arcs non convexes, mais présentent des sensibilités différentes à la structure des données. La stratégie par centroïde fournit une représentation globale stable, tandis que la stratégie de frontière reflète mieux la configuration structurelle locale. Empiriquement, cette approche montre des résultats prometteurs pour identifier les deux arcs, suggérant que l'intégration d'informations de frontière peut améliorer le clustering de structures non convexes.

Remerciements

Cette recherche a été financée par le projet de l'ANR "Analogies: from theory to tools and applications" (AT2TA), ANR-22-CE23-0023.

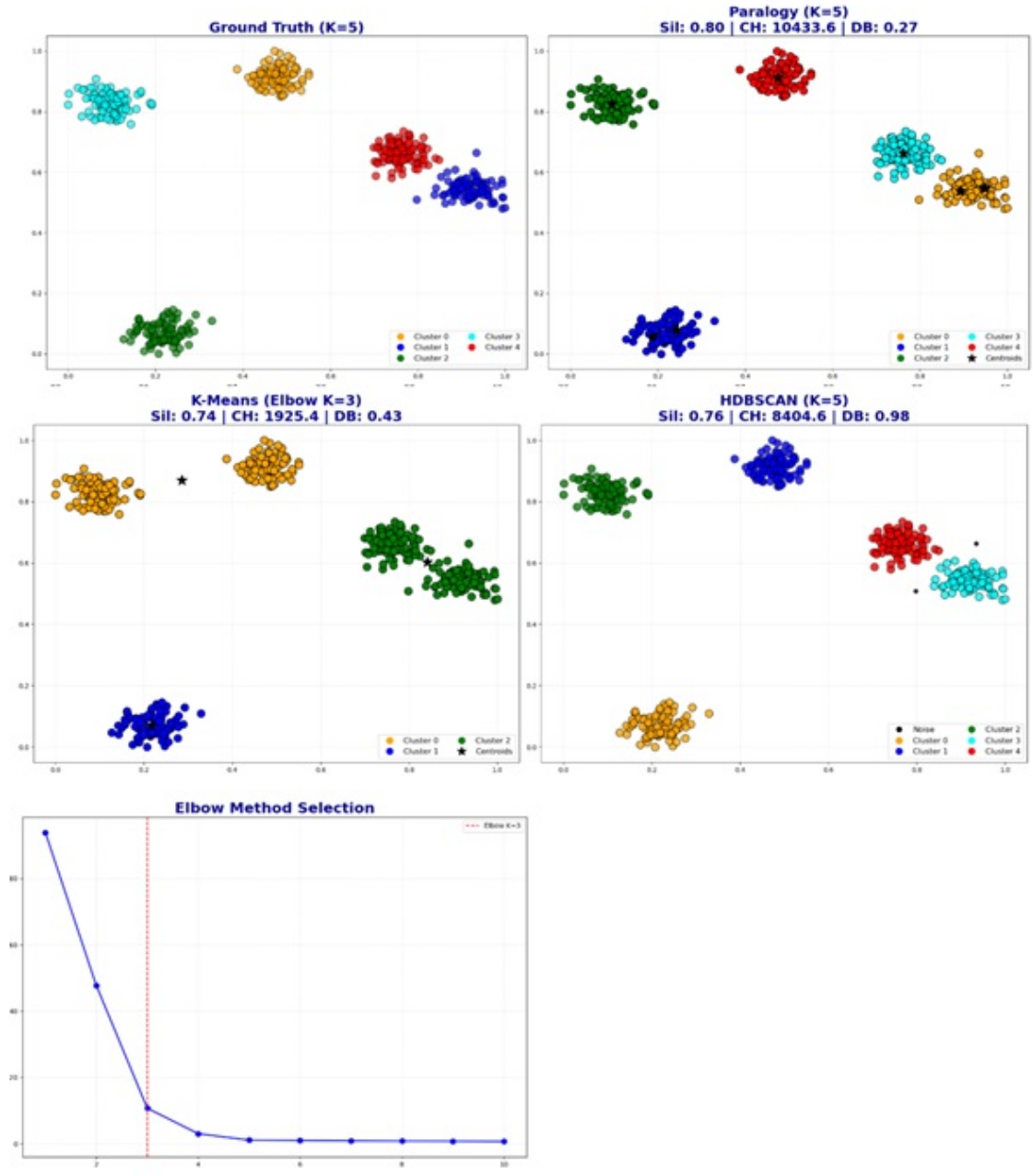

FIGURE 5 – Paralogy/HDBSCAN/k-means : Résultats pour le jeu de données Blobs

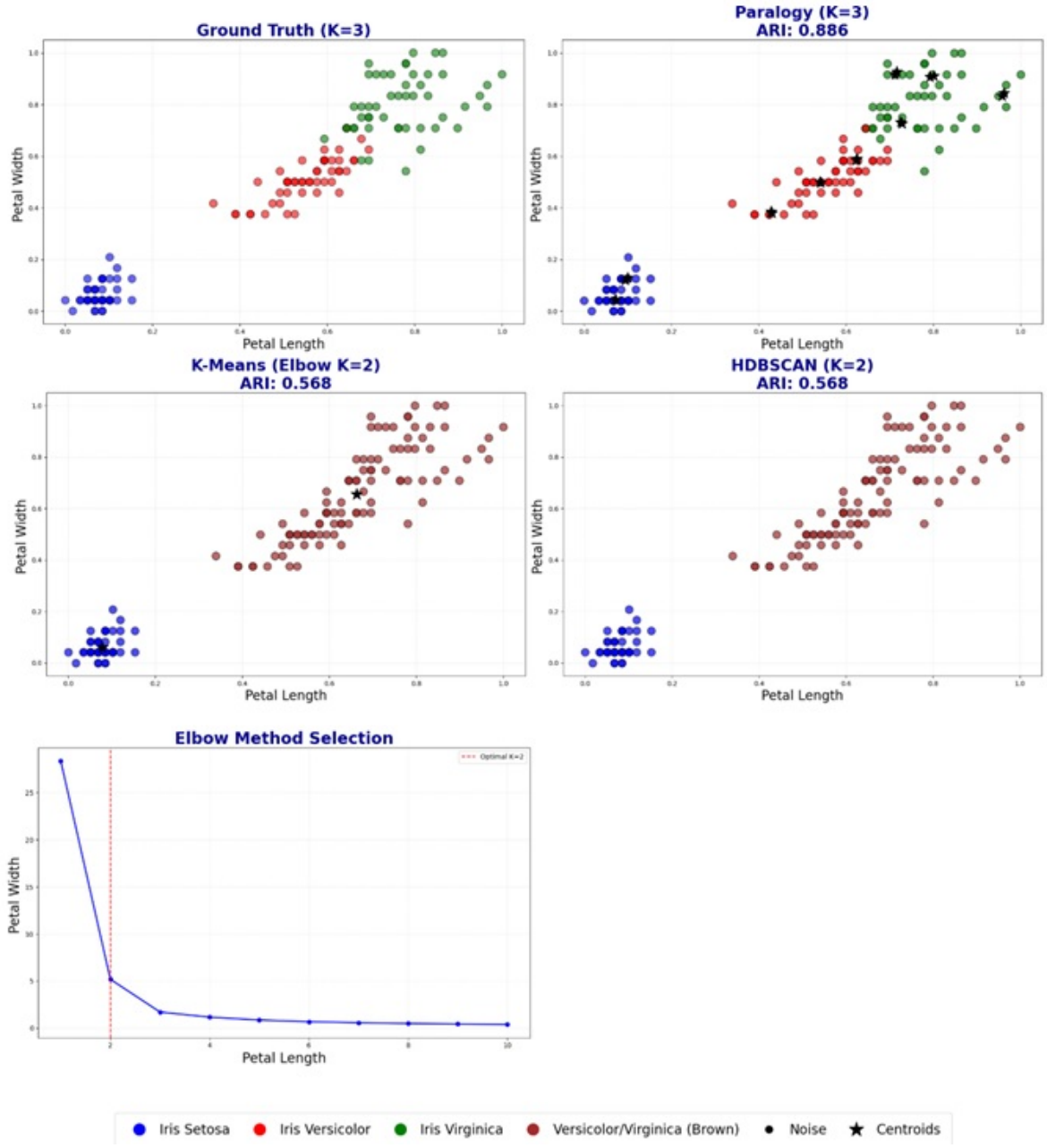

FIGURE 6 – Paralogy/HDBSCAN/k-means : Résultats pour le jeu de données Iris

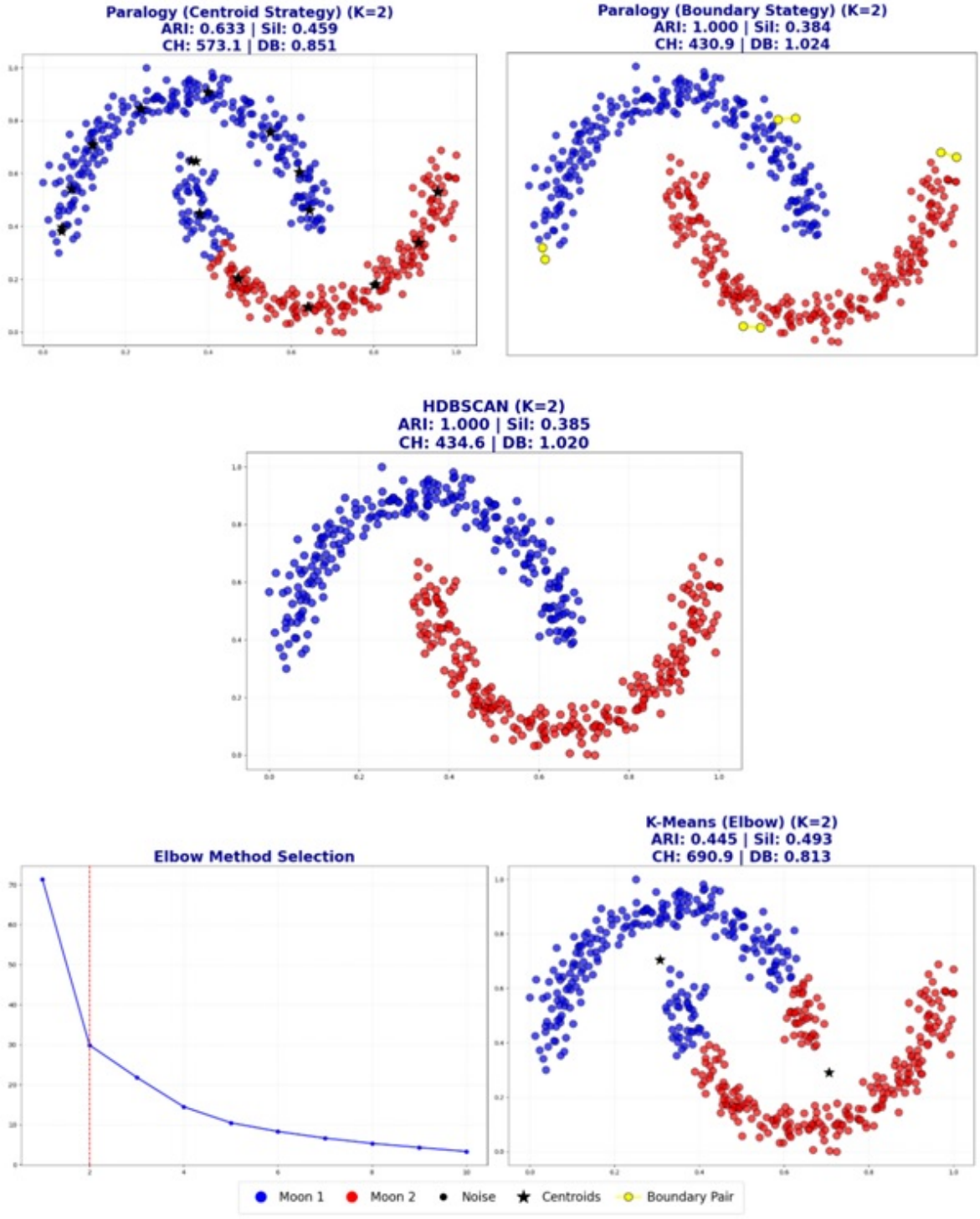

FIGURE 7 – Paralogy (stratégie centroïde)/Paralogy (stratégie frontière) /HDBSCAN/k-means : Résultats pour le jeu de données Moons

Références

[1] James C. Bezdek, Robert Ehrlich, and William Full. FCM : The fuzzy c-means clustering algorithm. *Computers & Geosciences*, 10(2-3) :191–203, 1984.

[2] Myriam Bouhnas and Henri Prade. A new approach to clustering based on an analogy-related logical proportion. In *4th Workshop Interactions between Analogical Reasoning and Machine Learning, IJCAI, 17 Aug., Montréal*. AAAI Press, 2025. Version française abrégée : Une méthode de clustering basée sur la paralogie inverse graduelle, Proc. Rencontres Francophones Logique Floue et Applications 2025, Clermont-Ferrand, Nov. 6-7, Cépaduès, 135-142.

[3] Ricardo J. G. B. Campello, Davoud Moulavi, and Joerg Sander. Density-based clustering based on hierarchical density estimates. In Jian Pei, Vincent S. Tseng, Longbing Cao, Hiroshi Motoda, and Guandong Xu, editors, *Advances in Knowledge Discovery and Data Mining : Proc. 17th Pacific-Asia Conf.(PAKDD'2013), Part II*, volume 7819 of *LNCS*, pages 160–172. Springer, 2013.

[4] Martin Ester, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proc. 2nd Int. Conf. on Knowledge Discovery and Data Mining (KDD'96), Portland, Aug. 2-4*, page 226–231, 1996.

[5] Peter Hájek. Metamathematics of fuzzy logic. Kluwer, 1998.

[6] Anil K Jain. Data clustering : 50 years beyond k-means. *Pattern Recognition Letters*, 31(8) :651–666, 2010.

[7] Stuart P. Lloyd. Least squares quantization in PCM. *IEEE Trans. Inf. Theory*, 28(2) :129–136, 1982.

[8] Henri Prade and Gilles Richard. Reasoning with logical proportions. In Fangzhen Lin, Ulrike Sattler, and Miroslaw Truszczynski, editors, *Proc. 12th Int. Conf. on Principles of Knowledge Representation and Reasoning (KR 2010), Toronto, May 9-13*. AAAI Press, 2010.

[9] Henri Prade and Gilles Richard. Homogeneous logical proportions : Their uniqueness and their role in similarity-based prediction. In Gerhard Brewka, Thomas Eiter, and Sheila A. McIlraith, editors, *Proc. 13th Int. Conf. on Principles of Knowledge Representation and Reasoning (KR 2012), Rome, June 10-14*. AAAI Press, 2012.

[10] Henri Prade and Gilles Richard. From analogical proportion to logical proportions. *Logica Universalis*, 7(4) :441–505, 2013.

[11] Henri Prade and Gilles Richard. Homogenous and heterogeneous logical proportions. *IfCoLog J. of Logics and their Applications*, 1(1) :1–52, 2014.

[12] Henri Prade and Gilles Richard. Pair comparators : a family of connectives related to logical proportions. In B. Vantaggi *et al.*, editor, *Proc. 21st Int. Conf. on Information Processing and Management of Uncertainty in Knowledge-Based Systems (IPMU'26), Rome, June 15-19*, Communications in Computer and Information Science. Springer, 2026.

[13] David E. Rumelhart and Adele A. Abrahamson. A model for analogical reasoning. *Cogn. Psychol.*, 5 :1–28, 1973.