Université Sorbonne Paris Nord
Institut Galilée

Initiation à la recherche

Sujet

Optimisation de la stabilité des résultats de K-means en utilisant une approche de clustering hybride avec Hierarchical Clustering (HCA)

Présenté par:

SARR Serigne Abdou Lat
NGAM Amadou

Supervisé par:

Dr. Guénaël CABANES

Abstract

Le clustering non supervisé est une technique couramment utilisée pour explorer des structures dans des données sans étiquette préalable. K-means est une méthode populaire de clustering non supervisé, cependant, il est connu pour être sensible à l'initialisation des centres, ce qui peut conduire à une faible stabilité des résultats.
Pour remédier à ce problème, nous proposons dans cet article une approche hybride de clustering qui utilise le clustering hiérarchique (HCA) pour initialiser les centres de K-means. Notre objectif est d'optimiser la stabilité des résultats de K-means tout en préservant la qualité des clusters obtenus.
Nous avons utilisé les indices de mesure ARI et Silhouette scores pour évaluer la qualité de nos clusters. Nous avons également effectué plusieurs exécutions de K-means et notre approche hybride pour vérifier la stabilité de notre solution. Les résultats de nos expériences montrent que notre approche hybride améliore significativement la stabilité des clusters par rapport à K-means seul, ce qui en fait une technique prometteuse pour une variété d'applications de clustering où la stabilité des résultats est critique.

Mot clés: clustering non supervisé, K-means, stabilité, clustering hybride, Hierarchical Clustering, ARI, Silhouette scores

Introduction

Au cours des dernières décennies, les progrès dans l'informatique et la science de données ont ouvert de nouvelles perspectives pour l'analyse des données. En particulier, l'analyse de clustering est devenue une technique courante pour regrouper des données similaires en groupes homogènes. Les résultats de clustering peuvent fournir des informations précieuses sur la structure des données et aider à comprendre les relations entre les variables.En effet le clustering est largement utilisé dans de nombreux domaines, notamment la science des données, la bioinformatique, la finance etc. K-means est l'une des méthodes les plus populaires de clustering non supervisé, mais il est souvent sensible à l'initialisation des centres, ce qui peut entraîner une faible stabilité des résultats.

La stabilité des résultats est devenue une problématique majeure dans le domaine du clustering non supervisé, car des résultats instables peuvent conduire à des conclusions incohérentes et des erreurs coûteuses. Cela a conduit à une recherche intense sur l'optimisation de la stabilité des résultats du clustering non supervisé.

Notre approche hybride de clustering combine Hierarchical Clustering (HCA) et K-means pour améliorer la stabilité des résultats de K-means. HCA est utilisé pour initialiser les centres de K-means et réduire sa sensibilité à l'initialisation aléatoire, en se basant sur la structure globale des données. Cette approche maintient une bonne qualité de clustering et présente des avantages significatifs pour améliorer la fiabilité des analyses de clustering non supervisé. Elle est innovante dans sa combinaison de deux méthodes de clustering non supervisé et offre ainsi de nouvelles perspectives pour améliorer la fiabilité des analyses de clustering.

Dans cet article, nous présentons une revue de l'état de l'art des techniques de clustering non supervisé, en mettant l'accent sur la problématique de la stabilité des résultats de K-means. Nous détaillons ensuite notre approche hybride de clustering et les expériences que nous avons menées pour valider notre contribution. Enfin, nous discutons des résultats obtenus, de leurs implications et de nos perspectives de recherche future.

Etat de l'art

Trouver une méthode effice pour l'iniatialisation des centres de clusters de k-means est un sujet de recherche actif. Il existe diverses méthodes proposées par des chercheurs. L'algorithme de clustering k-means est largement utilisé en raison de sa simplicité et de son efficacité. Il est capable de traiter de grands ensembles de données et de produire des résultats rapidement. Cependant, il présente des limitations importantes. L'initialisation aléatoire des centroïdes peut conduire à une convergence inattendue et à des résultats non optimaux. De plus, k-means nécessite de définir le nombre de clusters à l'avance, ce qui peut être un défi dans certaines situations. L'incapacité de k-means à gérer divers types de données est également une limitation majeure. Le Hierarchical Clustering (HCA) Quant à lui est une méthode de clustering qui construit une structure hiérarchique de clusters. Ses avantages incluent une vision structurée et hiérarchique des données, une flexibilité dans le choix des mesures de similarité et l'absence de besoin de spécifier le nombre de clusters à l'avance. Cependant, le HCA peut être sensible aux outliers, présente une complexité computationnelle élevée et peut être difficile à interpréter en raison de la présence de clusters imbriqués.
Pour surmonter ces limitations, certaines recherches se sont penchées sur l'utilisation de l'approche hybride avec HCA. HCA est un algorithme de clustering hiérarchique qui permet de détecter des structures complexes dans les données. En utilisant HCA pour initialiser les centroïdes de k-means, il est possible d'améliorer la stabilité des résultats et de mieux gérer les problèmes d'initialisation aléatoire. Cette approche hybride a montré des résultats prometteurs dans plusieurs études.

Pour une description plus détaillée de ces méthodes, des articles de revue plus exhaustifs sont disponibles dans la littérature, tels que ceux cités dans les références [ASI20], [JMF99], [CGBN14], [AGM+13], [Jr.63], [AI16]

Chapter 1

Proposition

1.1 Fonctionnement de k-means

Tout d'abord, le nombre de clusters K est fixé. Les K points initiaux, appelés centroïdes, sont généralement sélectionnés au hasard à partir des données. Chaque centroïde représente le centre d'un cluster. Pour chaque point de données, l'algorithme calcule sa distance par rapport à chaque centroïde et l'assigne au cluster dont le centroïde est le plus proche. Cette distance peut être mesurée en utilisant des mesures de similarité, telles que la distance euclidienne.Une fois que tous les points ont été attribués à des clusters, les nouveaux centroïdes sont calculés en prenant la moyenne des points appartenant à chaque cluster. Cela déplace les centroïdes vers le centre des clusters nouvellement formés. Les étapes d'attribution des points aux clusters et de mise à jour des centroïdes sont répétées jusqu'à ce qu'un critère d'arrêt soit atteint. Ce critère peut être le nombre maximum d'itérations prédéfini ou lorsque les centroïdes ne se déplacent plus de manière significative. Une fois que l'algorithme a convergé, les points sont finalement assignés de manière définitive à leurs clusters correspondants en utilisant les derniers centroïdes obtenus.

1.2 Fonctionnement de HCA

Le HCA construit une hiérarchie de clusters en calculant la similarité entre les points de données, fusionnant les clusters similaires et permettant une exploration structurée des données. On calcule une mesure de similarité ou de distance entre chaque paire de points de données.Une matrice de similarité est créée en utilisant les mesures de similarité calculées. Elle contient les valeurs de similarité entre toutes les paires de points de données. Le dendrogramme est une représentation graphique de la hiérarchie des clusters. Au départ, chaque point de données est considéré comme un cluster individuel. Ensuite, les clusters les plus similaires sont progressivement fusionnés pour former des clusters plus grands. Différentes méthodes peuvent être utilisées pour fusionner les clusters, telles que la fusion par lien simple, la fusion par lien complet ou la fusion par groupe moyen. Chaque méthode définit la similarité entre les clusters en fonction de la similarité entre leurs points de données.Le dendrogramme peut être coupé à différents niveaux pour obtenir un nombre spécifique de clusters. Le choix du niveau de coupure dépend du problème spécifique et des critères d'évaluation utilisés.
Une fois que le nombre de clusters est déterminé, on peut les interpréter en examinant les caractéristiques des points de données qu'ils contiennent. Des techniques de visualisation peuvent être utilisées pour mieux comprendre la structure des clusters.

1.3 Combinaison de K-means et de HCA

On utilise HCA pour effectuer une première étape de clustering sur les données. Cela permet d'obtenir une structure hiérarchique de clusters.
Ensuite, à partir de cette structure hiérarchique, on sélectionne les centres de clusters pour K-means. On peut choisir les centres à différents niveaux de la hiérarchie en fonction de critères tels que la densité des clusters ou la distance entre les points de données. Une fois que les centres de clusters sont définis, on effectue plusieurs itérations de K-means en utilisant ces centres comme points d'initialisation. Cela permet d'obtenir des clusters plus stables, car les centres sont choisis de manière plus délibérée à partir de la structure hiérarchique.
Après chaque itération de K-means, on évalue la stabilité des clusters obtenus en utilisant des indices de mesure tels que l'indice de Rand ajusté (ARI) ou le score de silhouette. Ces mesures permettent de quantifier la qualité des clusters.
En combinant K-means avec HCA, on exploite les avantages des deux approches : la capacité de K-means à trouver des partitions précises et HCA à fournir une structure hiérarchique. Cela permet de renforcer la stabilité des résultats de K-means et de réduire la dépendance à l'initialisation aléatoire des centroïdes.

On notera **Khca** l'algorithme résultant de la combinaison de K-means et de HCA.

Chapter 2

Expériences et Résultats

2.1 Protocole expérimental

Base de données utilisées: Digits, Make blob, Colposcopy

Au cours de ces experiences, nous avons lancer les algorithmes de k-means et de HCA existants sur ces bases de données. Ensuite on a tester notre solution(KHCA: l'algorithme proposé) sur ces datasets. Pour chaque base de données, nous allons appliquer notre solution en faisant varier le nombre de clusters K. En parallèle, nous allons effectuer le même test avec les algorithmes de clustering existants(Kmeans et HCA). L'objectif est de faire par la suite une comparaison en fonction d'une ou de plusieurs critères d'évaluation mesuré sur chaque expérience. Nous avons utilisé deux critères d'évaluation: silhouette score, ARI score(pour les datasets labellisé). Pour chaque base de données, on calcule la moyenne du critère d'évaluation sur les différences expériences.

Pour mesurer la stablité de notre solution par rapport au k-means, nous avons lancer plusieurs fois (50 fois) kmeans et notre solution(khca) sur une base de données étiquettée (nombre de clusters = 3). Sur chaque itération, on calcule le ARI Score ensuite on determine la moyenne des ARI Scores sur ces 50 itérations. Ces resultats vont nous permettre d'affirmer sur la performance de la solution proposée.

2.2 Mesure de performances avec Silhouette et Ari score

Les graphes ci-dessous comparent les performances entre les algorithmes de k-means, Hca et de Khca (la combinaison des deux algorithmes) avec Silhouette score.
Le score ARI (Adjusted Rand Index) est une mesure de similarité entre les clusters obtenus et les classes réelles des données. Un score ARI proche de 1 indique une correspondance parfaite entre les clusters et les classes réelles, tandis qu'un score proche de 0 indique une absence de correspondance.
Le score silhouette varie de -1 à 1, où une valeur proche de 1 indique que le point est bien adapté à son propre cluster et mal adapté aux clusters voisins, tandis qu'une valeur proche de 0 indique que le point est équidistant des clusters voisins et une valeur proche de -1 indique que le point est mal adapté à son propre cluster et bien adapté aux clusters voisins.

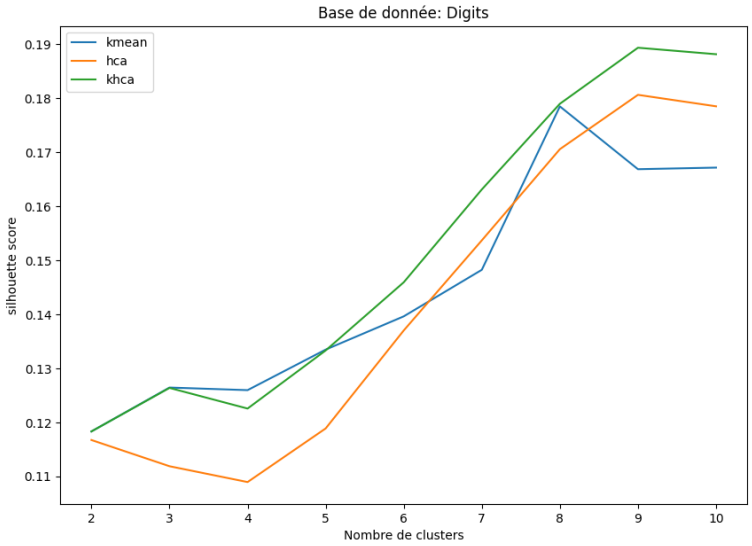


Figure 2.1: Performance Silhouette des algorithmes sur la base de donnée Digis

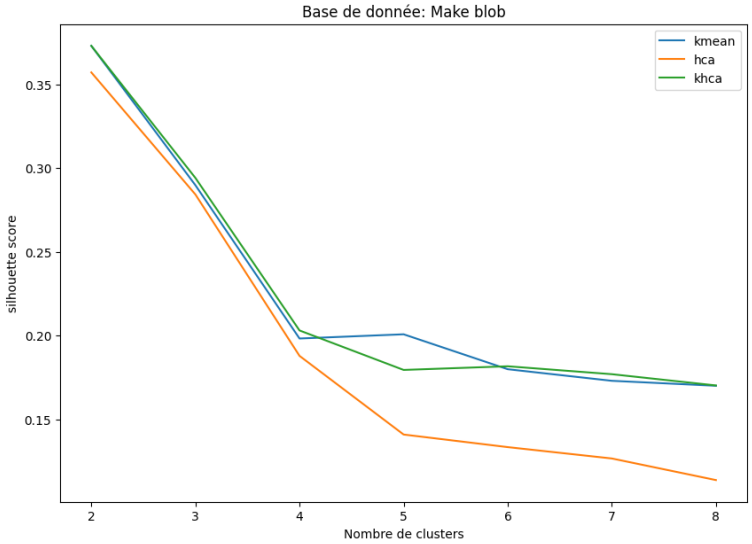


Figure 2.2: Performance Silhouette des algorithmes sur la base de donnée Make Blob

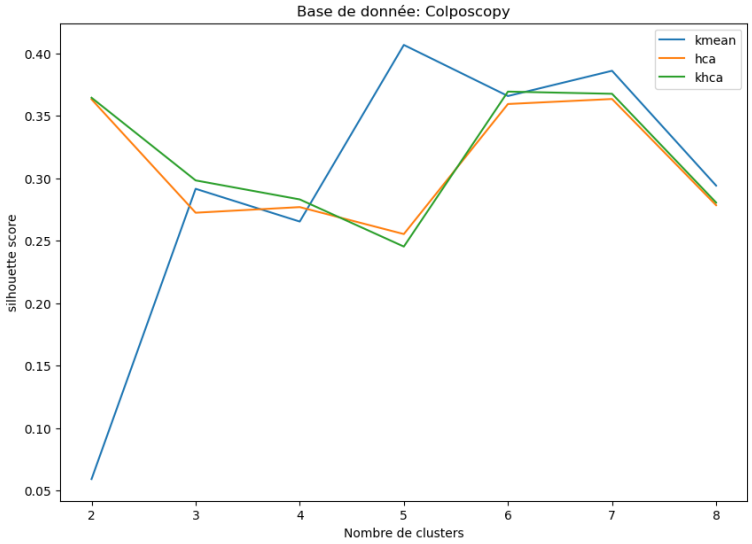


Figure 2.3: Performance Silhouette des algorithmes sur la base de donnée Colposcopy

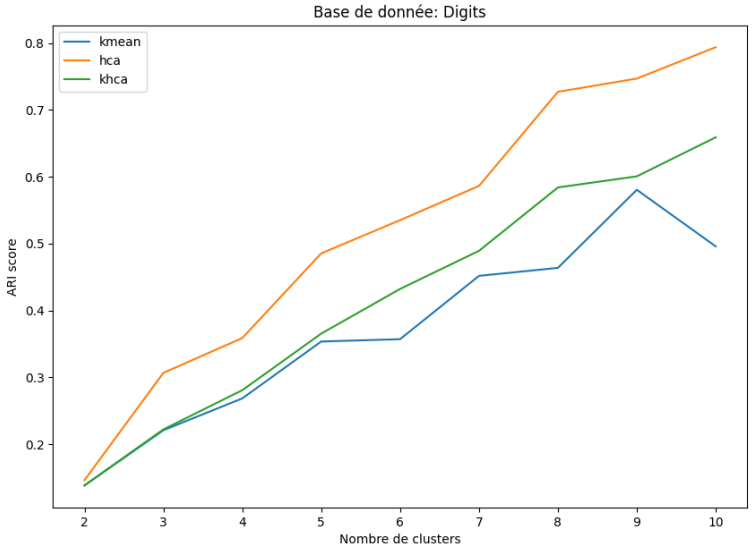


Figure 2.4: Performance Ari des algorithmes sur la base de donnée Digis

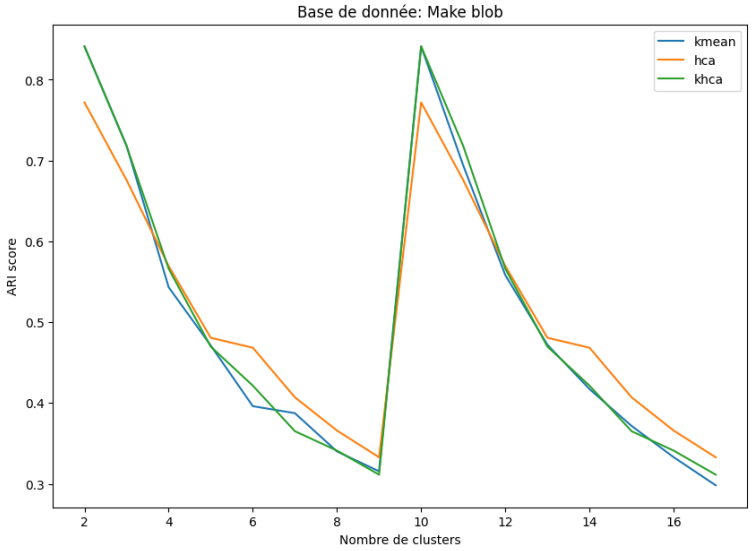


Figure 2.5: Performance Ari des algorithmes sur la base de donnée Make blob

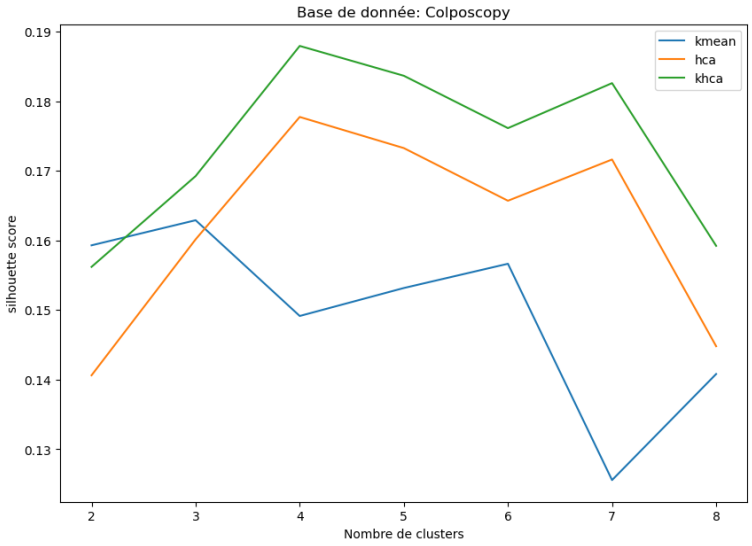


Figure 2.6: Performance Ari des algorithmes sur la base de donnée Colposcopy

2.3 Tableau récaputilatif des moyennes des indices sur chaque base de données

Les tableaux suivant représentent la moyenne des deux indices de performance sur les trois datasets (Digits, Make Blob, Colposcopy). En analysant les performances des trois algorithmes de clustering Kmeans, HCA et KHCA sur les bases de données Digits, Make blob et Colposcopy, nous pouvons observer les résultats suivants.

dataset/algo	K-MEANS	HCA	KHCA
Digits	0.136	0.141	0.151
Make blob	0.226	0.191	0.225
Colposcopy	0.149	0.161	0.173

Table 2.1: Moyennes Silhouette score des algorithmes

Globalement, nous pouvons observer que KHCA a souvent obtenu les meilleurs scores silhouette sur ces trois bases de données, suivis par HCA et Kmeans. Cependant, les différences de performance entre les trois algorithmes sont relativement faibles, ce qui suggère que le choix de l'algorithme dépendra des caractéristiques spécifiques des données et des objectifs de clustering.

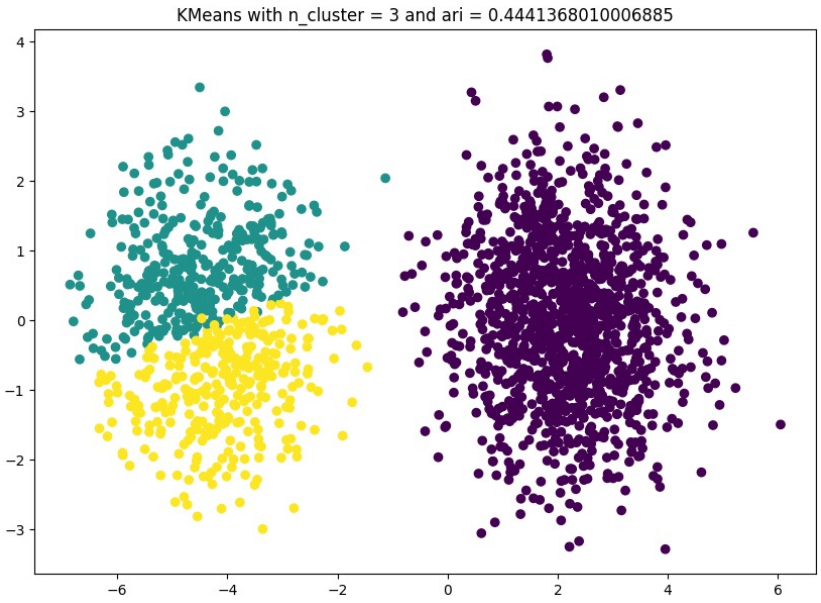
dataset/algo	K-MEANS	HCA	KHCA
Digits	0.390	0.520	0.420
Make blob	0.501	0.509	0.504
Colposcopy	0.439	0.309	0.315

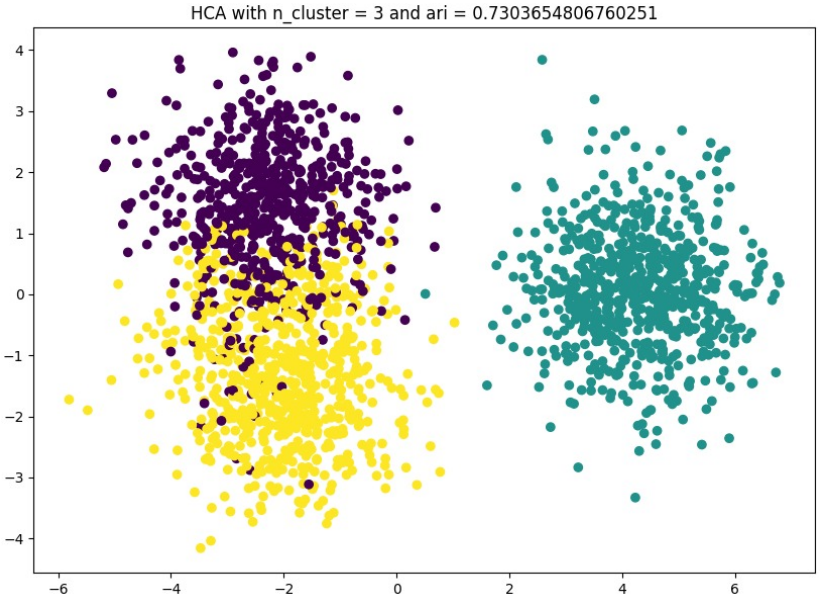
Table 2.2: Moyennes ARI score des algorithmes

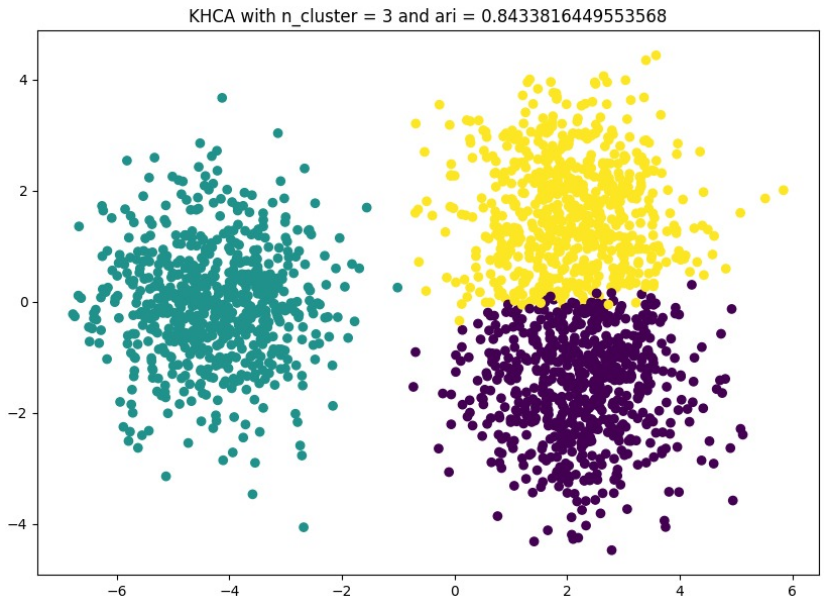
On remarque que sur les trois algorithmes,les performances sont très variables en fonction de la base de données étudiée. HCA est souvent performant sur la base de données Digits, tandis que Kmeans peut être plus performant sur la base de données Colposcopy.Ainsi, l'exécution de plusieurs fois les algorithmes de clustering, tels que Kmeans, HCA et KHCA,contribue à améliorer la qualité des résultats de clustering , en augmentant la fiabilité et la stabilité des résultats. 2.3

2.4 Score Ari aprés plusieurs exécution des différents algorithmes sur la base de donnée Make blob

Le but de compiler plusieurs fois Kmeans et HCA est de vérifier la stabilité des résultats obtenus et de réduire l'impact des conditions initiales aléatoires sur les résultats de clustering. Comme mentionné précédemment, Kmeans peut donner des résultats différents à chaque exécution en raison de l'initialisation aléatoire des centres de clusters. De même, HCA peut être sensible aux valeurs aberrantes et aux bruits dans les données, ce qui peut entraîner des variations importantes dans les résultats de clustering.


KMeans with n_cluster = 3 and ari = 0.4441368010006885


HCA with n_cluster = 3 and ari = 0.7303654806760251


KHCA with n_cluster = 3 and ari = 0.8433816449553568

En compilant plusieurs fois Kmeans et HCA, on peut observer la variabilité des résultats et vérifier si les clusters obtenus sont cohérents et stables. Cela permet également de mieux évaluer la performance des algorithmes et d'obtenir des résultats plus fiables et robustes (voir le tableau2.3 de la section 2.5).
En combinant Kmeans et HCA dans l'approche KHCA, la compilation multiple peut aider à réduire l'impact des deux méthodes sur les résultats de clustering. En appliquant Kmeans sur les clusters formés par la HCA, on peut obtenir des clusters plus précis et plus stables.

2.5 Moyenne du Score Ari aprés plusieurs exécution des différents algorithmes sur la base de donnée Make blob

La compilation multiple peut donc aider à évaluer la stabilité de cette approche hybride et à vérifier si elle améliore la performance de clustering par rapport à Kmeans et HCA utilisés seuls.

dataset/algo	K-MEANS	HCA	KHCA
Make blob	0.708	0.730	0.843

Table 2.3: Tableau des moyennes de Ari aprés 50 exécutions

Les résultats obtenus sur la base de données Make Blob montrent que l'approche hybride KHCA améliore considérablement la performance de clustering par rapport à Kmeans et HCA utilisés seuls.

Le score ARI obtenu pour Kmeans sur la base de données Make Blob est de 0.708, ce qui indique une performance de clustering modérée. Le score ARI obtenu pour HCA est de 0.730, ce qui suggère une amélioration par rapport à Kmeans. Cependant, le score ARI obtenu pour KHCA est encore plus élevé, avec un score de 0.843.
Cela indique que l'approche hybride KHCA a permis d'obtenir des clusters plus proches des classes réelles des données et donc une meilleure performance de clustering.
Ainsi, les résultats sur la base de données Make Blob montrent que KHCA peut améliorer la performance de clustering par rapport à Kmeans et HCA utilisés seuls, avec un score ARI nettement plus élevé. Cela suggère que KHCA peut être une approche prometteuse pour le clustering de données complexes.

Conclusion

En conclusion, la combinaison de Kmeans et HCA dans l'approche KHCA peut améliorer la performance de clustering par rapport à Kmeans et HCA utilisés seuls. Les résultats obtenus sur les base de données ont souvent montré que le score de KHCA était nettement plus élevé que ceux obtenus par Kmeans et HCA utilisés seuls.Cependant, il convient de noter que la combinaison de Kmeans et HCA dans l'approche KHCA peut être sensible aux paramètres tels que le nombre de clusters, le critère de distance, etc. Il est donc important de trouver les meilleurs paramètres pour cette approche pour obtenir des résultats de clustering optimaux en d'autres termes trouvé le bon cluster .
D'autres approches hybrides ont également été proposées pour combiner Kmeans et HCA, telles que la combinaison séquentielle, la combinaison pondérée, la combinaison concurrente, etc. Ces approches peuvent également améliorer la performance de clustering, mais nécessitent également une optimisation de paramètres. Enfin, d'autres méthodes de clustering peuvent également être combinées avec Kmeans ou HCA pour améliorer la performance de clustering. Par exemple, la combinaison de Kmeans avec la méthode de clustering spectral a montré des performances améliorées sur certaines bases de données. Les possibilités de combinaison de différentes méthodes de clustering sont nombreuses et peuvent dépendre des caractéristiques spécifiques des données.

Bibliography

[AGM+13] Arbelaitz, Olatz, Ibai Gurrutxaga, Javier Muguerza, Jesús M. Pérez and Iñigo Perona: *An extensive comparative study of cluster validity indices*. Pattern Recognition, 46(1):243–256, 2013.

[AI16] Alaoui Ismaili, Oumaima: *Clustering prédictif Décrire et prédire simultanément*. Theses, Université Paris Saclay (COmUE), November 2016.

[ASI20] Ahmed, Mohiuddin, Raihan Seraj and Syed Mohammed Shamsul Islam: *The k-means algorithm: A comprehensive survey and performance evaluation*. Electronics, 9(8):1295, 2020.

[CGBN14] Charrad, Malika, Nadia Ghazzali, Véronique Boiteau and Azam Niknafs: *NbClust: An R Package for Determining the Relevant Number of Clusters in a Data Set*. Journal of Statistical Software, 61(6):1–36, 2014.

[JMF99] Jain, A. K., M. N. Murty and P. J. Flynn: *Data Clustering: A Review*. ACM Comput. Surv., 31(3):264–323, sep 1999.

[Jr.63] Jr., Joe H. Ward: *Hierarchical Grouping to Optimize an Objective Function*. Journal of the American Statistical Association, 58(301):236–244, 1963.

Contributions des membres

Tests de kmeans et de HCA sur les trois base de données avec ARI et :**NGAM Amadou**
Combinaison de kmeans et de hca : **NGAM Amadou et SARR Serigne Abdou Lat (réunion Teams)**
Tests de kmeans et de HCA sur les trois base de données avec Silhouette :**SARR Serigne Abdou Lat**
Tests de KHCA sur make blob, Digits et Coploscopy avec ARI :**NGAM AMADOU**
Exécution multiple des algorithmes sur les bases de donneés avec ARI :**NGAM AMADOU**
Tests de KHCA sur make blob, Digits et Coploscopy avec Silhouette :**SARR Serigne Abdou Lat**
Slides de présentation et rapport :**NGAM Amadou et SARR Serigne Abdou Lat (réunion Teams)**