ENSA
ÉCOLE NATIONALE DES SCIENCES APPLIQUÉES
KHOURIBGA

Université Sultan Moulay Slimane

Master : Big Data & Aide à la Décision
Rapport Data Maining

Clustring
Classification hiérarchique ascendante (CHA)

Effectuée par :

Aouatif Ait Boulahcen
Fouad Khattari
Zizouan Azizan

Année Universitaire : 2023-2024

TABLE DES MATIÈRES

1 Résumé **3**

2 Clustring **4**
2.1 Introduction 4
2.2 Objectif 4
2.3 Démarche classique 4
2.4 Domaines d'applications 5
2.5 Clustering : par partitionnement 5
2.6 Clustering : hiérarchique 6

3 Classification hiérarchique ascendante **7**
3.1 Introduction 7
3.2 Algorithme 7
3.3 Propriétés des dendrogrammes 7
3.4 Choix du nombre de clusters : 8

4 Métrique **9**
4.1 Probléme 9
4.2 Les avantages : 9
4.3 Les inconvénients : 10
4.4 Exemple 10

5 Conclusion **15**

CHAPITRE 1

RÉSUMÉ

La classification hiérarchique ascendante (CHA) est une méthode d'analyse de données utilisée pour regrouper des individus ou des objets similaires en groupes. Cette méthode commence par considérer chaque individu ou objet comme un cluster distinct, puis fusionne progressivement les clusters les plus similaires jusqu'à ce qu'un seul cluster contenant tous les individus soit formé. La CHA peut être utilisée pour explorer des données et découvrir des structures cachées, ainsi que pour la segmentation de marché et l'analyse de la biologie. Cette méthode peut également être utilisêe comme une étape de prêtraitement pour des algorithmes de prédiction, tels que la classification ou la régression.La CHA peut être mise en oeuvre à l'aide de différentes mesures de similarité, telles que la distance euclidienne ou la corrélation, ainsi que de différentes méthodes de fusion de clusters, telles que la méthode de Ward ou la méthode de liaison complète. La CHA peut également être visualisée sous forme de dendrogramme pour aider à interpréter les résultats. En somme, la CHA est une méthode d'analyse de données utile pour la classification et la segmentation de groupes similaires d'individus ou d'objets, ainsi que pour la découverte de structures cachèes dans les données.

CHAPITRE 2

CLUSTRING

2.1 Introduction

Le clustering, ou regroupement, est une méthode d'analyse de données qui vise à regrouper des objets similaires en sous-groupes. Cette technique d'apprentissage non supervisé permet d'explorer et de découvrir des structures cachées dans les données en se basant sur des critères de similarité. Les méthodes de clustering incluent le clustering hiérarchique, le clustering partitionné, le clustering basé sur la densité et le clustering de modèle mixte. Le clustering est utilisé dans divers domaines pour réduire la dimensionnalité des données et peut aider à la segmentation de marché, l'analyse biologique, l'exploration de données, la prédiction, etc.

2.2 Objectif

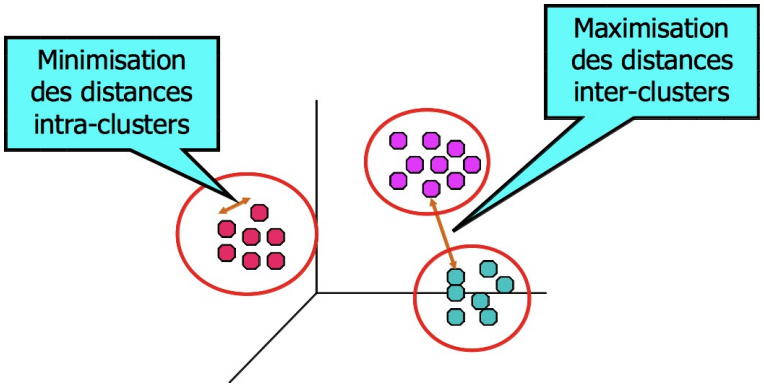
Le clustering est une méthode d'analyse de données polyvalente qui peut être utilisée pour diverses applications. Il peut aider à découvrir des structures cachées ou des motifs dans les données, à identifier des groupes de données similaires ou à détecter des anomalies. De plus, le clustering peut réduire la dimensionnalité des données, ce qui peut être utile pour la compression de données et le stockage de données volumineuses. Le clustering peut également être utilisé comme une étape de prétraitement pour des algorithmes de prédiction, comme la classification ou la régression, pour augmenter la précision de la prédiction. Enfin, le clustering peut aider les entreprises à personnaliser leur marketing en identifiant des groupes de clients ayant des comportements d'achat similaires, ainsi qu'àidentifier des groupes de gènes ou de patients ayant des caractéristiques similaires dans les études cliniques pour la recherche biologique et le développement de nouveaux traitements médicaux.

2.3 Démarche classique

Former des groupes homogénes à l'intérieur d'une population

- Etant donné un ensemble de points, chacun ayant un ensemble d'attributs, et une mesure de similarité définie sur eux, trouver des groupes (classes, segments, clusters) tels que :
  - — Les points à l'intérieur d'un méme groupe sont trés similaires entre eux.
  - — Les points appartenant à des groupes différents sont trés dissimilaires.

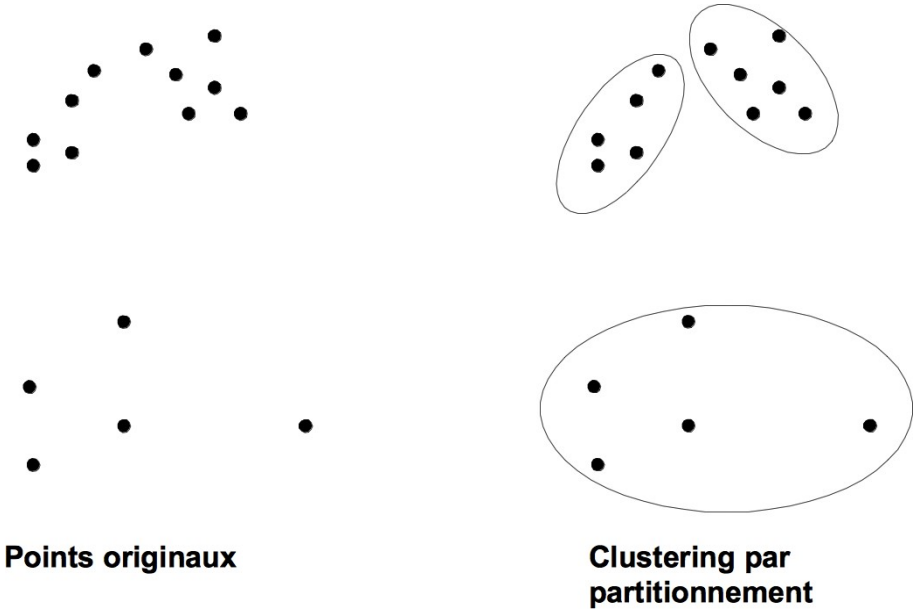
- Le choix de la mesure de similarité est important.



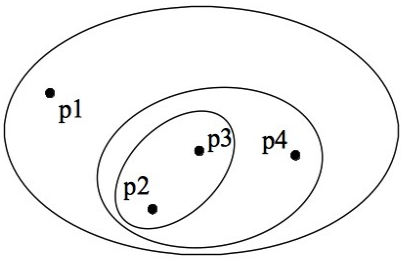
2.4 Domaines d'applications

- Text mining : textes proches, dossiers automatiques
- Web mining : pages web proches
- Marketing : segmentation de la clientéle
- Web lot analysis : profils utilisateurs
- BioInformatique : génes ressemblants

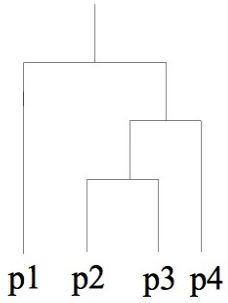
2.5 Clustering : par partitionnement



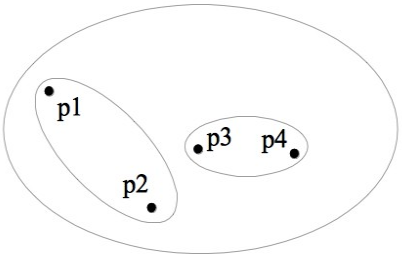
2.6 Clustering : hiérarchique



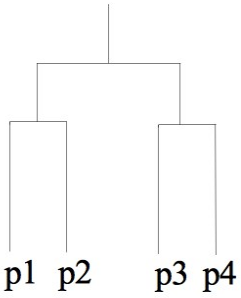
Clustering hiérarchique classique



Dendrogram classique



Clustering hiérarchique non classique



Dendrogram non classique

CHAPITRE 3

CLASSIFICATION HIÉRARCHIQUE ASCENDANTE

3.1 Introduction

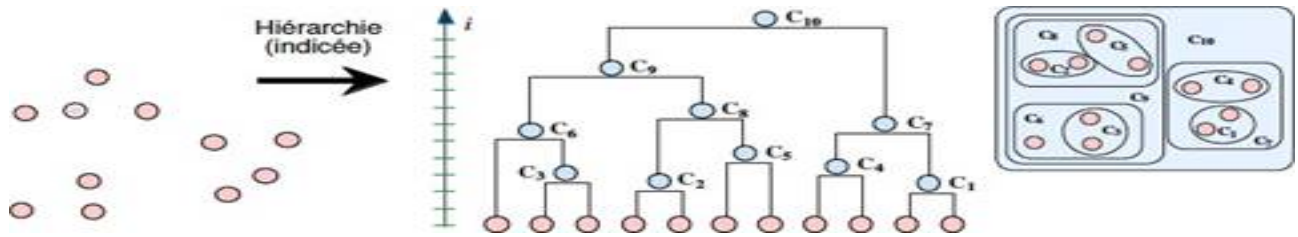
La Classification Ascendante Hiérarchique (CAH) est une méthode d'apprentissage non supervisée dont le but est la partition automatique d'objets (ou individus) en sousgroupes d'individus similaires pour une certaine mesure de ressemblance. Elle fait partie des méthodes dites de regroupement hiérarchique. Dans la suite, on notera Ω un ensemble de n individus, x1, . . . , xn, à partitionner.

3.2 Algorithme

- Initialisation
  - Chaque individu est placé dans son propre cluster
  - Calcul de la matrice de ressemblance M entre chaque couple de clusters
- Répétition
  - Sélection dans M des deux clusters les plus proches Ci et Cj .
  - Fusion de Ci et Cj pour former un cluster Cg.
  - Mise à jour de M en calculant la ressemblance entre Cg et les clusters existants.
- Jusqu'a fusion des 2 derniers clusters.

3.3 Propriétés des dendrogrammes

Dans cette partie nous introduisons la notion de dendrogramme qui est un outil classique de représentation graphique des résultats de classification ascendante hiérarchique. Un dendrogramme est une représentation sous forme d'arbre de la hiérarchie induite par la CAH. Cet arbre peut-être pondéré ou non, ce qui induit deux points de vue, topologique ou pondéré.



— Dendogramme = représentation des fusions successives.

— Hauteur d'un cluster dans le dendogramme = similarité entre les deux clusters avant la fusion (sauf pour certaines mesures de similarite).

3.4 Choix du nombre de clusters :

La détermination du nombre de clusters à partir d'un dendrogramme de Classification Ascendante Hiérarchique (CAH) peut être réalisée en traçant une ligne horizontale à travers le dendrogramme et en coupant les liens entre les clusters à une hauteur spécifique. Le choix de la hauteur de coupure dépendra de la méthode utilisée pour la CAH et des objectifs de l'analyse.

Pour la méthode "Ward"*, la hauteur de coupure peut être choisie de sorte que le nombre de clusters formés soit le plus petit possible tout en maintenant la cohérence interne de chaque cluster élevée. D'autres mesures de similarité ou de dissimilarité peuvent également être utilisées pour choisir la hauteur de coupure, telles que la distance euclidienne ou la distance de Manhattan.

Il n'existe pas de méthode unique et universelle pour choisir le nombre de clusters à partir d'un dendrogramme de CAH, et cela dépendra de la nature des données, des objectifs de l'analyse et de l'expertise du chercheur.

* : La méthode de Ward est une méthode de la classification ascendante hiérarchique (CAH) qui vise à minimiser la somme des carrés des différences entre chaque observation et la moyenne de son groupe, dans le but de maximiser la cohérence interne de chaque groupe.

Cette méthode est considérée comme une méthode agglomérative, car elle commence par considérer chaque observation comme son propre groupe et fusionne ensuite les groupes en utilisant un critère de similarité jusqu'à ce qu'il ne reste qu'un seul groupe.

La méthode de Ward est souvent préférée car elle permet de minimiser les effets des variations aléatoires sur les données et elle donne des groupes de tailles similaires. Elle est largement utilisée dans les domaines de la biologie, de la psychologie et des sciences sociales pour regrouper des individus ou des objets en fonction de leurs caractéristiques communes.

CHAPITRE 4

MÉTRIQUE

4.1 Probléme

Trouver la métrique entre les clusters la plus proche de la métrique utilisée entre les individus : min, max, moyenne, ...

— **Saut minimal (Single linkage) :** Se base sur dmin(C1, C2), distance entre les deux points les plus proches de chaque cluster
  — Tendance à produire des classes générales

— **Saut maximal (Complete linkage) :** Se base sur la distance dmax(C1, C2), distance entre les deux points les plus éloignés des deux clusters.
  — Tendance à produire des classes spécifiques (on ne regroupe que des classes trés proches).

4.2 Les avantages :

La classification ascendante hiérarchique (CAH) est une méthode de classification qui présente les avantages suivants :

- **Aide au choix du nombre de groupes :** La plupart des méthodes de clustering demandent à l'utilisateur de choisir le nombre de groupes qu'il souhaite créer. Ce n'est pas le cas de la CAH qui va calculer toutes les combinaisons possibles. Elle les représente ensuite via un dendrogramme qui permettra au Data Scientist de choisir le nombre de clusters le plus adapté à ses données et à son objectif.
- **Hiérarchique :** La construction des groupes se fait étape par étape. Ce qui veut dire que vous pourrez d'abord faire le choix de segmenter vos données en 3 groupes par exemple. Si vous analysez le contenu de ces 3 groupes et qu'un groupe ne vous parait pas homogène, vous pourrez décider de passer à un découpage à 4 groupes sans avoir à tout recommencer.
- **Facile à utiliser :** La CAH construit systématiquement un dendrogramme (une sorte d'arbre) qui va résumer tous les regroupements qui ont été faits. Ce dendrogramme est vraiment très utile et facilite l'utilisation de la CAH.
- **Choix de la distance :** C'est un dernier point qui peut avoir son importance. Pour décider si

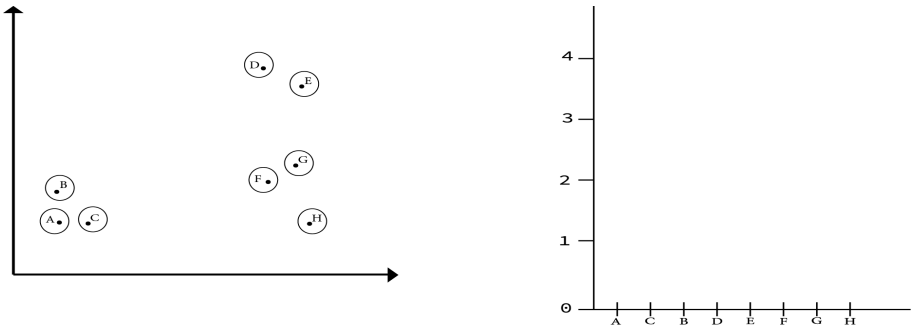
des individus ou des groupes sont proches ou éloignés, vous pourrez choisir votre distance. En particulier j'utilise souvent la distance de Ward pour les segmentations client. Elle permet d'éviter que des individus atypiques se retrouvent seuls isolés dans un groupe.

4.3 Les inconvénients :

- **Pas adapté aux grands volumes de données :** Malheureusement la CAH ne peut pas être utilisée sur des volumes de données importants. Les temps de calcul explosent et cela s'explique complètement par les méthodes de calcul de l'algorithme. Il existe quand même des solutions dans ce cas, on peut coupler la CAH à un k-means.
- **Sensibilité aux valeurs aberrantes :**la CAH est sensible aux valeurs aberrantes, qui peuvent perturber la structure des groupes. Il est donc important de prétraiter les données pour supprimer les valeurs aberrantes ou d'utiliser des méthodes robustes qui peuvent traiter ces valeurs.
- **Problèmes de stabilité :** la CAH peut être instable et produire des résultats différents pour des échantillons de données différents. Il est donc important de valider les résultats de la CAH à l'aide de méthodes de validation croisée .

4.4 Exemple

■ **Etape 0 :**
On considère que chaque point est seul et isolé dans son groupe.

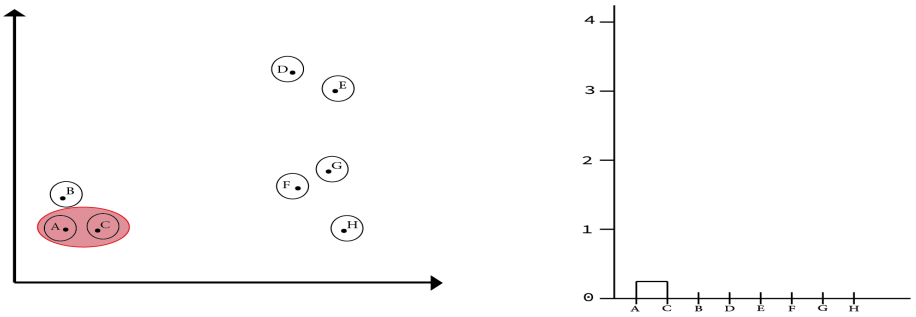


	A	B	C	D	E	F	G	H
A	0							
B	0.5	0						
C	0.25	0.56	0					
D	5	4.72	4.80	0				
E	4.78	5.55	5.57	1	0			
F	4.32	4.23	4.07	2.01	2.06	0		
G	4.92	4.84	4.68	2.06	1.81	0.61	0	
H	5	5.02	4.75	3.16	2.90	1.28	1.12	0

FIGURE 4.1 – Etape 0 : Initialisation

■ **Etape itérative :**

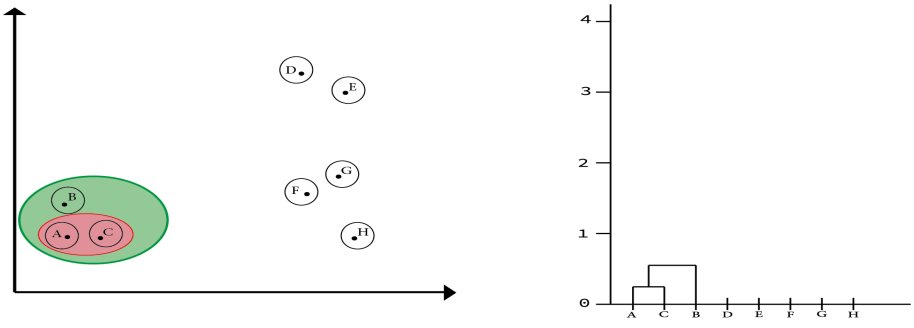
Dans cette étape de la CAH, nous calculons d'abord la distance entre chaque groupe. Il est possible d'utiliser différentes distances pour évaluer la proximité entre les groupes. Cependant, pour la CAH, nous préférons utiliser la distance euclidienne qui permet de prendre en compte la taille des groupes lors du calcul de la distance. Cela est important car cela permet d'éviter d'isoler les valeurs extrêmes dans un groupe, ce qui aurait peu de sens métier en général.



	AC	B	D	E	F	G	H
AC	0						
B	0.5	0					
D	4.8	4.72	0				
E	5.57	5.55	1	0			
F	4.07	4.23	2.01	2.06	0		
G	4.68	4.84	2.06	1.81	0.61	0	
H	4.75	5.02	3.16	2.90	1.28	1.12	0

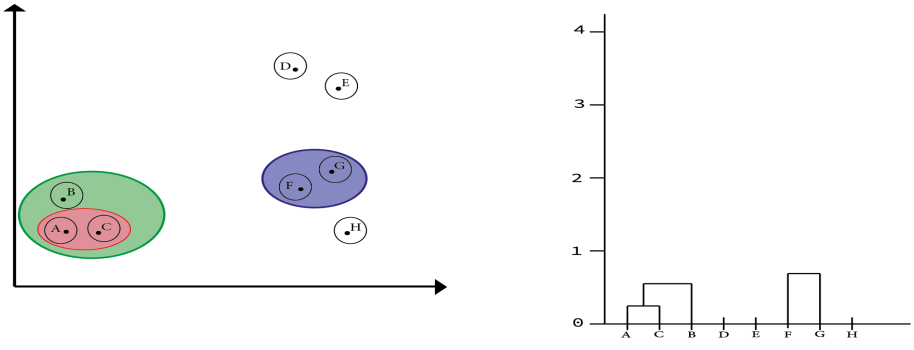
FIGURE 4.2 – Etape 1

Et on va continuer cette étape itérative encore et encore jusqu'à ce qu'il ne reste plus qu'un seul et unique groupe réunissant tous les individus.



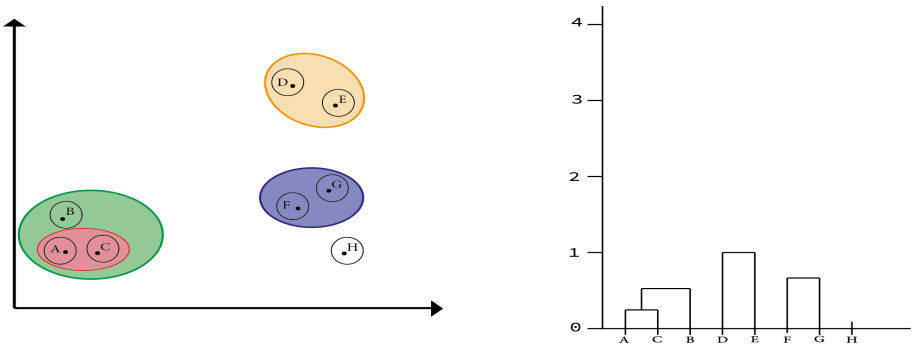
	ACB	D	E	F	G	H
ACB	0					
D	4.72	0				
E	5.55	1	0			
F	4.07	2.01	2.06	0		
G	4.68	2.06	1.81	0.61	0	
H	4.75	3.16	2.90	1.28	1.12	0

FIGURE 4.3 – Etape 2



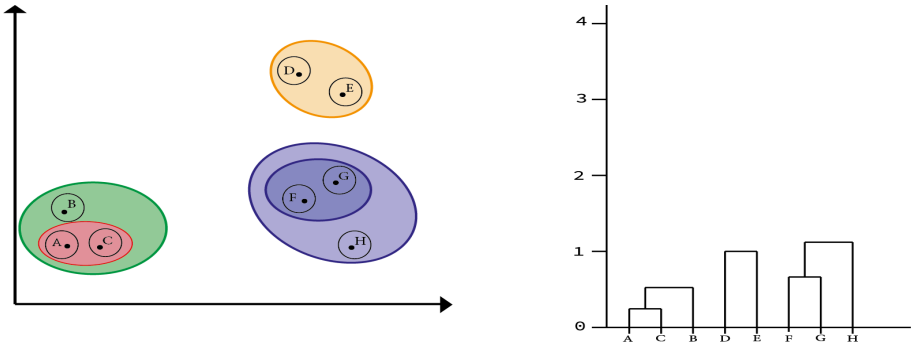
	ACB	D	E	FG	H
ACB	0				
D	4.72	0			
E	5.55	1	0		
FG	4.07	2.01	1.81	0	
H	4.75	3.16	2.90	1.12	0

FIGURE 4.4 – Etape 3



	ACB	DE	FG
DE	4.72	0	
FG	4.23	1.81	0
H	4.07	2.90	1.12

FIGURE 4.5 – Etape 4



	ACB	DE
DE	4.72	0
FGH	4.01	1.81

FIGURE 4.6 – Etape 5

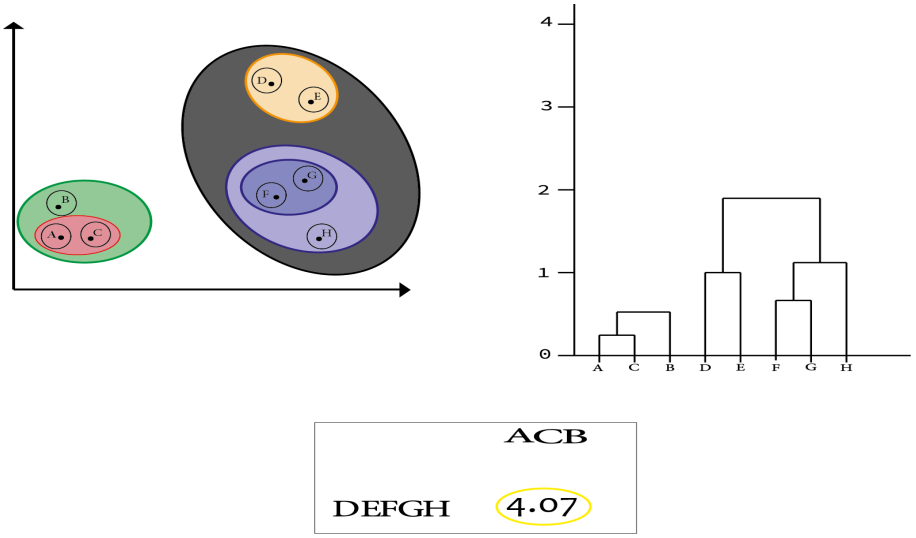


FIGURE 4.7 – Etape 6

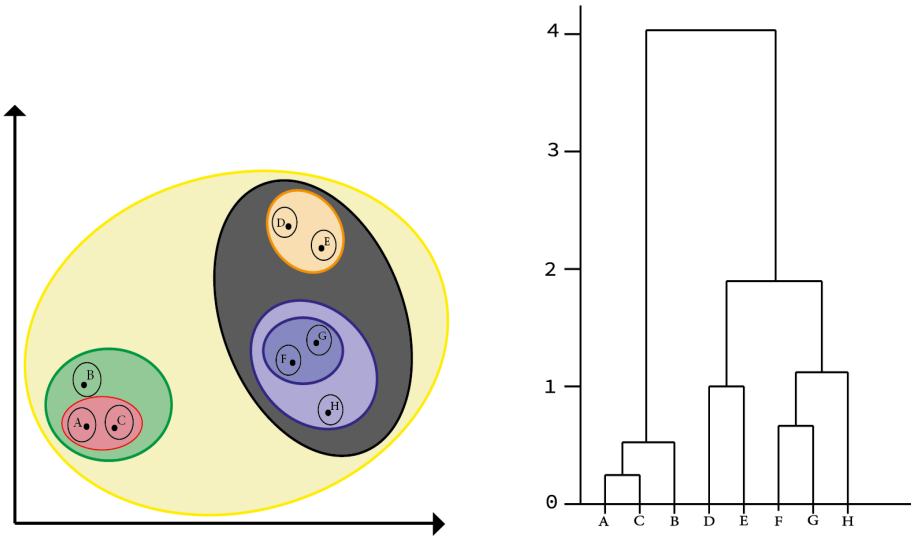


FIGURE 4.8 – Etape 7

CHAPITRE 5

CONCLUSION

En conclusion, la Classification Ascendante Hiérarchique (CAH) est une méthode de clustering puissante qui permet de regrouper des observations similaires en clusters ou groupes. Elle utilise une approche hiérarchique pour regrouper progressivement les observations en clusters de plus en plus grands, jusqu'à ce que toutes les observations soient regroupées en un seul grand cluster.

La CAH peut utiliser différentes méthodes pour mesurer la distance entre les clusters et pour déterminer comment les clusters sont combinés. La distance euclidienne est l'une des mesures les plus couramment utilisées, mais d'autres mesures peuvent également être utilisées en fonction des données et du contexte.

La CAH est largement utilisée dans de nombreux domaines, tels que la biologie, la médecine, la finance, la géologie, etc. Elle est utile pour explorer des données et identifier des groupes similaires d'observations. Cependant, la CAH a quelques limites, telles que la difficulté de déterminer le nombre optimal de clusters et la sensibilité aux valeurs aberrantes.

En somme, la CAH est une méthode de clustering utile qui peut aider à découvrir des structures dans les données et à faciliter la prise de décision. Elle nécessite cependant une compréhension adéquate des données et une sélection judicieuse des paramètres pour obtenir des résultats fiables et pertinents.