République Algérienne Démocratique et Populaire

Ministère de l'Enseignement Supérieur et de la Recherche Scientifique

UNIVERSITÉ MOHAMED KHIDER, BISKRA

Faculté des Sciences Exactes et des Sciences de la Nature et de la Vie

DÉPARTEMENT DE MATHÉMATIQUES

Mémoire présenté en vue de l'obtention du Diplôme :

MASTER en Mathématiques

Option : **Statistique**

Par

GHLAM Ikram

Titre :

Méthodes de Classification

Membres du Comité d'Examen :

Pr.	**MERAGHNI Djamel**	UMKB	Président
Pr.	**BENELMIR Imen**	UMKB	Encadreur
Dr.	**TOUBA Sonia**	UMKB	Examinateur (rice)

Juin 2024

Dédicace

À mes chers parents **Abdelaziz** et **Djamila**, pour leur amour et leur soutien inépuisable.

À mon frère et à ma sœur, pour leur complicité et leur encouragement constant.

À mon encadreur, pour ses conseils précieux et son accompagnement rigoureux.

À toutes mes amies, pour leur présence et leur soutien moral tout au long de ce parcours.

À tous ceux qui m'ont soutenu et aidé, chacun de vos gestes a contribué à la réalisation de ce travail.

REMERCIEMENTS

Avant toute chose, je tiens à remercier **ALLAH** le Tout-Puissant de m'avoir accordé courage, patience et force tout au long de ces années d'étude.

Je tiens à exprimer ma sincère gratitude envers mon encadreur, **Dr. Benelmir Imane**, pour son accompagnement exceptionnel tout au long de ce projet. Son soutien précieux, ses remarques pertinentes, ses conseils constructifs, sa patience et son expertise ont été cruciaux pour l'accomplissement de ce mémoire. La confiance qu'elle m'a témoignée, sa disponibilité constante et ses judicieux conseils ont joué un rôle déterminant dans la réussite de ce travail.

Je tiens également à adresser mes vifs remerciements aux membres du jury qui m'ont fait l'honneur d'examiner et d'évaluer mon travail, apportant la richesse de leurs compétences à cette évaluation.

Je tiens à exprimer ma gratitude à mes chers parents, **Abdelaziz** et **Djamila**, pour leur amour et leur soutien inépuisable.

Je remercie mon cher frère et ma chère sœur pour leur complicité et leur encouragement constant.

Je remercie toutes mes amies surtout **Rayane**, qui a été une partie merveilleuse de ma vie.

Table des matières

Table des figures

Liste des tableaux

Introduction

La classification est une tâche essentielle en apprentissage automatique, jouant un rôle central dans de nombreuses applications scientifiques et industrielles. À travers l'histoire, la classification a été utilisée pour ordonner et comprendre une vaste gamme de phénomènes naturels, notamment dans des domaines aussi divers que la biologie, la médecine, et plus récemment, la technologie de l'information et les sciences des données. Le cœur de cette démarche consiste à attribuer des étiquettes à des objets ou des observations sur la base de leurs caractéristiques. Cette capacité à classer précisément et efficacement a des implications profondes pour la prise de décision, la prédiction des phénomènes.

Dans mes recherches, je me suis principalement concentré sur la classification non supervisée, également connue sous le nom de clustering. Cette méthode vise à regrouper des individus ou des données présentant des similitudes, sans disposer de connaissances préalables sur les catégories.

Dans ce mémoire, j'ai abordé la problématique suivante : "Comment procéder à la méthode de classification ?". Cette question soulève plusieurs interrogations :

- Quelles sont les méthodes de classification disponibles ?
- Comment les données sont-elles représentées ?
- Quelles méthodes de classification conviennent le mieux à ces données spécifiques ?
- Quels critères doivent être pris en compte pour effectuer la classification ?

Ces questions ont constitué le cadre de mes recherches et ont guidé mon exploration des différentes approches de classification non supervisée, ainsi que des méthodes pour représenter et évaluer les données. En répondant à ces interrogations, j'ai cherché à approfondir notre compréhension de la classification non supervisée et à fournir des orientations pratiques pour son application dans divers domaines.

Le but de la classification est de regrouper des objets selon leurs caractéristiques similaires en classes homogènes tout en séparant celles qui présentent des différences. Mon modeste travail se compose de deux chapitres :

- Dans le premier, on va d'abord discuter sur le principe de la méthode de classification hiérarchique on passant d'abord par quelques rappels statistiques. Puis, on présente les deux types de classification hiérarchique : la classification ascendante hiérarchique (CAH) et la classification descendante hiérarchique (CDH).
- Le deuxième chapitre est sur la classification non hiérarchique où on va expliquer le principe de cette méthode et présenter ces différents algorithmes qui sont l'algorithme de K-means, l'algorithme des nuées dynamiques et enfin algorithme des centres mobiles.

Ce travail se conclut par une synthèse générale.

Chapitre 1

Classification hiérarchique

La classification hiérarchique se distingue comme une technique essentielle d'analyse des données, précieuse pour déchiffrer la structure profonde des ensembles des données. Son application transcende divers domaines, allant de l'analyse phylogénétique en biologie à la segmentation de la clientèle en marketing, en passant par des applications en sociologie et psychologie, partout où il est crucial de regrouper naturellement des éléments similaires.

L'un des principaux atouts de cette méthode est sa capacité à générer un dendrogramme. Ce diagramme offre une visualisation intuitive des regroupements formés à travers les différentes étapes, qu'il s'agisse de la fusion dans l'approche ascendante ou de la division dans l'approche descendante. Le dendrogramme ne se contente pas de révéler la structure des données ; il fournit également un moyen efficace de déterminer le nombre de clusters optimaux en identifiant le point de coupe idéal dans l'arbre. Elle se décline en deux approches principales : l'approche ascendante, qui est la plus répandue, et l'approche descendante.

Dans l'approche ascendante, chaque objet débute comme un cluster isolé. Ces clusters sont ensuite progressivement fusionnés en fonction de leur similarité, calculée par des mesures telles que la distance euclidienne ou la distance de Manhattan. Ce processus

se poursuit jusqu'à ce qu'un critère d'arrêt prédéfini soit atteint, aboutissant à un grand cluster unique ou à une structure préférée.

L'approche descendante, moins courante, commence avec tous les objets dans un seul cluster qui est ensuite divisé récursivement en sous-clusters plus petits, jusqu'à ce que chaque objet forme son propre cluster ou jusqu'à un certain point spécifié par le critère d'arrêt.

1.1 Principe de la classification hiérarchique

Le principe de la classification hiérarchique consiste à organiser des éléments en groupes (ou clusters) de manière que les éléments à l'intérieur de chaque groupe soient le plus similaires possible entre eux, selon un critère de proximité ou de similitude prédéfini. Cette méthode vise à créer une série de subdivisions au sein d'un ensemble d'individus, basées sur un standard de similitude qui est quantifié à travers une matrice de distance. Cette matrice exprime la distance entre chaque paire d'individus, permettant d'évaluer leur degré de ressemblance.

Les individus sont considérés comme similaires si les points qui leur sont associés dans l'espace des caractéristiques sont proches les uns des autres, i.e. si les distances qui les séparent sont faibles. La classification hiérarchique procède ensuite par étapes successives de regroupement ou de division des individus, formant ainsi un arbre de classification (ou dendrogramme) qui illustre la manière dont les groupes sont formés ou divisés à partir de l'ensemble initial.

Cette structure arborescente permet d'observer non seulement la formation des clusters à différents niveaux de similarité, mais aussi la relation hiérarchique entre les clusters, offrant une vue d'ensemble de la disposition des données et facilitant l'interprétation des groupements naturels au sein des données analysées.

1.2 Données et leurs caractéristiques

Généralement, les données sont organisées dans un format de tableau rectangulaire. Dans cette structure, les lignes sont associées à des individus ou des entités statistiques, et les colonnes représentent des variables, également nommées caractères ou attributs.

1.2.1 Tableau des données

L'ensemble des valeurs z_{ij} est présenté sous la forme d'une matrice Z, de n lignes et k colonnes, appelée tableau des données.

$$Z = (z_{ij})_{n\times k} := \begin{pmatrix} z_{11} & z_{12} & & z_{1k} \\ z_{21} & z_{22} & & z_{2k} \\ & & \ddots & \\ z_{n1} & z_{n2} & & z_{nk} \end{pmatrix} \in M_{n,k}(\mathbb{R}),$$

sachant que $i = 1, \ldots, n$ et $j = 1, \ldots, k$.

Remarque 1.2.1 **1**. *Chaque variable peut être représentée par un vecteur de dimension n, appelé vecteur variable, correspondant aux valeurs prises par cette variable sur les n individus. On la note par*

$$z_j = (z_{1j}, \ldots, z_{nj})^t \in \mathbb{R}^n.$$

2. *Chaque individu est décrit par k variables, formant un vecteur de dimension k, appelé vecteur individu. On la note par*

$$e_i = (z_{i1}, \ldots, z_{ik})^t \in \mathbb{R}^k.$$

3. *On note par z_{ij} l'observation du caractère z_j sur l'individu e_i.*

1.2.2 Poids

Le poids attribué à chaque individu représente l'importance qu'on lui accorde dans l'étude. Il est par fois d'assigner des poids variés à différents individus. Pour ce faire, on emploie une matrice diagonale, notée p suivante

$$p = \begin{pmatrix} p_1 & 0 & & 0 \\ 0 & p_2 & & 0 \\ & & \ddots & \\ 0 & & & p_n \end{pmatrix},$$

où $\sum_{i=1}^{n} p_i = 1$, avec $0 \leq p_i \leq 1$.

Remarque 1.2.2 *On note par I_n, la matrice indicatrice définie comme suit*

$$I_n = \begin{pmatrix} 1 & 0 & & 0 \\ 0 & 1 & & 0 \\ & & \ddots & \\ 0 & & & 1 \end{pmatrix} \in M_{n,n}(\mathbb{R}).$$

On général, le poids est le même pour tous les individus, donc $p_i = \frac{1}{n}$.

1.2.3 Matrice de variance-covariance et matrice de corrélation

Dans cette partie, on va définir les deux types de liaisons entres les v.as, à savoir la matrice de variance-covariance et la matrice de corrélation.

1.2.3.1 Matrice de variance-covariance

La matrice de variance-covariance (ou simplement matrice de covariance) d'un vecteur de k variables aléatoires (v.a.) $\overrightarrow{Z} = (Z_1, Z_1, \ldots, Z_k)^t$, dont chacune a une variance finie est la matrice carrée dont le terme générique est donné par

$$\boxed{\sigma_{ij} = Cov(Z_i, Z_j).}$$

Soit Z_c la matrice des données centrées définie par

$$Z_c = \begin{pmatrix} z_{11} - \overline{z}_1 & z_{12} - \overline{z}_2 & & z_{1k} - \overline{z}_k \\ z_{21} - \overline{z}_1 & z_{22} - \overline{z}_2 & & z_{2K} - \overline{z}_k \\ & & \ddots & \\ z_{n1} - \overline{z}_1 & z_{n2} - \overline{z}_2 & & z_{nk} - \overline{z}_k \end{pmatrix} \in M_{n\times k}(\mathbb{R}),$$

où $\overline{z}_j = \sum_{i=1}^{n} p_i z_{ij}$.

La matrice de covariance notée V est définie par

$$\begin{aligned} V &= Z_c^t p Z_c \\ &= \begin{pmatrix} Var(Z_1) & Cov(Z_1, Z_2) & & Cov(Z_1, Z_k) \\ Cov(Z_2, Z_1) & Var(Z_2) & & Cov(Z_2, Z_K) \\ & & \ddots & \\ Cov(Z_k, Z_1) & Cov(Z_k, Z_2) & & Var(Z_k) \end{pmatrix} \\ &= \begin{pmatrix} \sigma_1^2 & \sigma_{12} & & \sigma_{1k} \\ \sigma_{21} & \sigma_2^2 & & \sigma_{2k} \\ & & \ddots & \\ \sigma_{k1} & \sigma_{k2} & & \sigma_k^2 \end{pmatrix}, \end{aligned}$$

avec

- $\sigma_j^2 = Cov(Z_j, Z_j) = \sum_{i=1}^{n} p_i (z_{ij} - \overline{z}_j)^2$.
- $\sigma_{lj} = Cov(Z_l, Z_j) = \sum_{i=1}^{n} p_i (z_{ij} - \overline{z}_j)(z_{il} - \overline{z}_l)$.

Preuve. On montre que $Z_{cj}^t p Z_{cl} = Cov(Z_l, Z_j)$.

$$
\begin{aligned}
Z_{cj}^t p Z_{cl} &= \begin{pmatrix} z_{11}-\overline{z}_1 & & z_{n1}-\overline{z}_1 \\ & \ddots & \\ z_{1k}-\overline{z}_k & & z_{nk}-\overline{z}_k \end{pmatrix} \begin{pmatrix} p_1 & & 0 \\ & \ddots & \\ 0 & & p_n \end{pmatrix} \begin{pmatrix} z_{11}-\overline{z}_1 & & z_{1k}-\overline{z}_k \\ & \ddots & \\ z_{n1}-\overline{z}_1 & & z_{nk}-\overline{z}_k \end{pmatrix} \\
&= \begin{pmatrix} p_1(z_{11}-\overline{z}_1) & & p_n(z_{n1}-\overline{z}_1) \\ & \ddots & \\ p_1(z_{1k}-\overline{z}_k) & & p_n(z_{nk}-\overline{z}_k) \end{pmatrix} \begin{pmatrix} z_{11}-\overline{z}_1 & & z_{n1}-\overline{z}_1 \\ & \ddots & \\ z_{1k}-\overline{z}_k & & z_{nk}-\overline{z}_k \end{pmatrix} \\
&= \begin{pmatrix} \sum_{i=1}^n p_i(z_{i1}-\overline{z}_1)^2 & & \sum_{i=1}^n p_i(z_{i1}-\overline{z}_1)(z_{ik}-\overline{z}_k) \\ & \ddots & \\ \sum_{i=1}^n p_i(z_{ik}-\overline{z}_k)(z_{il}-\overline{z}_l) & & \sum_{i=1}^n p_i(z_{ik}-\overline{z}_k)^2 \end{pmatrix} \\
&= \begin{pmatrix} Var(Z_1) & Cov(Z_1,Z_2) & Cov(Z_1,Z_k) \\ Cov(Z_2,Z_1) & Var(Z_2) & Cov(Z_2,Z_K) \\ & \ddots & \\ Cov(Z_k,Z_1) & Cov(Z_k,Z_2) & Var(Z_k) \end{pmatrix} \\
&= Cov(Z_l, Z_j).
\end{aligned}
$$

■

Remarque 1.2.3 ***1.*** *La matrice est symétrique, étant donné la propriété que* $Cov(Z_l, Z_j) = Cov(Z_j, Z_l)$.

2. *Ses valeurs propres sont positives ou nulles. Lorsqu'il n'existe aucune relation affine presque sûre entre les composantes du vecteur aléatoire, la matrice* V *est à valeurs propres strictement positives (elle est définie positive).*

3. *Les éléments de sa diagonale représentent la variance de chaque variable, étant donné la propriété que* $Cov(Z, Z) = Var(Z)$.

4. *Les éléments en dehors de la diagonale représentent la covariance entre les variables* i *et* j *quand* $i \neq j$.

1.2.3.2 Matrice de corrélation

La matrice de corrélation d'un vecteur de k v.a. $\overrightarrow{Z}$, dont chacune possède une variance (finie), est la matrice carrée dont le terme générique est donné par

$$\boxed{r_{ij} = Cor(Z_i, Z_j) = \frac{Cov(Z_i, Z_j)}{\sigma_i \sigma_j}.}$$

La matrice diagonale de la racine de la variance (écart type) est définie comme suit

$$D_\sigma = \begin{pmatrix} \sigma_1 & 0 & & 0 \\ 0 & \sigma_2 & & 0 \\ & & \ddots & \\ 0 & & & \sigma_k \end{pmatrix}$$

Soit Z_r la matrice centrée réduite associée à Z définie par

$$\begin{aligned} Z_r &= Z_c D_\sigma^{-1} \\ &= \begin{pmatrix} \frac{z_{11} - \overline{z}_1}{\sigma_1} & \frac{z_{12} - \overline{z}_2}{\sigma_2} & & \frac{z_{1k} - \overline{z}_k}{\sigma_k} \\ \frac{z_{21} - \overline{z}_1}{\sigma_1} & \frac{z_{22} - \overline{z}_2}{\sigma_2} & & \frac{z_{2K} - \overline{z}_k}{\sigma_k} \\ & & \ddots & \\ \frac{z_{n1} - \overline{z}_1}{\sigma_1} & \frac{z_{n2} - \overline{z}_2}{\sigma_2} & & \frac{z_{nk} - \overline{z}_k}{\sigma_k} \end{pmatrix}, \end{aligned}$$

où D_σ^{-1}est la matrice inverse de D_σ.

La matrice de corrélation notée R est définie comme suit

$$\begin{aligned} R &= Z_r^t p Z_r \\ &= D_\sigma^{-1} V D_\sigma^{-1} \\ &= \begin{pmatrix} 1 & r_{12} & \cdots & r_{1k} \\ r_{2k} & 1 & \cdots & r_{2k} \\ & & \ddots & \\ r_{k1} & r_{k2} & \cdots & 1 \end{pmatrix}. \end{aligned}$$

Les termes diagonaux de cette matrice sont égaux à 1, elle est symétrique, semi-définie positive et ses valeurs propres sont positives ou nulles.

Preuve. On montre que $R = D_\sigma^{-1} V D_\sigma^{-1}$.

$$\begin{aligned} R &= Z_r^t p Z_r \\ &= (Z_c D_\sigma^{-1})^t p (Z_c D_\sigma^{-1}) \\ &= (D_\sigma^{-1} Z_c^t) p (Z_c D_\sigma^{-1}) \\ &= D_\sigma^{-1} (Z_c^t p Z_c) D_\sigma^{-1} \\ &= D_\sigma^{-1} V D_\sigma^{-1}. \end{aligned}$$

■

1.3 Mesures d'éloignement

Tout système destiné à analyser ou organiser automatiquement un ensemble de données ou de connaissances doit incorporer, d'une manière ou d'une autre, un mécanisme capable d'évaluer avec précision les similitudes ou les différences présentes parmi ces données. La notion de proximité (ou ressemblance) a été l'objet d'études approfondies dans une variété de domaines très larges. Pour décrire cet opérateur, des termes tels que la similarité, la dissimilarité ou la distance sont souvent employés.

1.3.1 Dissimilartié

On not l'ensemble des individus par $\Omega = \{e_1, \ldots, e_n\}$ .Une dissimilarité est une application d de $\Omega \times \Omega$ dans $\mathbb{R}^+$ vérifiant les propriétés suivantes

$$\begin{aligned} \forall e_1, e_2 \in \Omega : \quad & d(e_1, e_2) \geq 0. \\ & d(e_1, e_2) = 0 \Longleftrightarrow e_2 = e_1. \\ & d(e_1, e_2) = d(e_2, e_1). \end{aligned}$$

1.3.2 Distance

La distance est utilisée pour mesurer la dissimilarité entre deux ensembles. On appelle distance sur un ensemble Ω toute application $d : \Omega \times \Omega$ dans $\mathbb{R}^+$ telle que

$$\begin{aligned} \forall e_1, e_2 \in \Omega : \quad & d(e_1, e_2) \geq 0. \\ & d(e_1, e_2) = 0 \Longleftrightarrow e_2 = e_1. \\ \forall e_1, e_{2,} e_3 \in \Omega : \quad & d(e_1, e_2) \leq d(e_1, e_3) + d(e_3, e_2). \end{aligned}$$

Remarque 1.3.1 *1. Si Ω est un ensemble fini, la distance peut être normée.*

2. Une distance est une dissimilarité mais l'inverse n'est pas vrai.

3. On utilise souvent une distance pour évaluer une dissimilarité.

1.3.3 Mesures de distance

La classification hiérarchique utilise des mesures de distance entre les objets pour former des classes (groupes).Cette technique da classification permet de calculer de nombreux types de mesures de distances, afin de l'utiliser directement dans la procédure.

Distance Euclidienne

C'est probablement le type de distance le plus couramment utilisé. Il s'agit simplement d'une distance géométrique dans un espace multidimensionnel. Elle est donnée

par la formule ci-dessous

$$\forall i, i' = \overline{1,n} \colon d(e_i, e_{i'}) = \sqrt{\sum_{j=1}^{k} (z_{ij} - z_{i'j})^2}.$$

Distance de Manhattan (city-block)

Cette distance est simplement la somme des différences entre les dimensions. Dans la plupart des cas, cette mesure de distance produit des résultats proches de ceux obtenus par la distance euclidienne. Elle est donnée par la formule ci-dessous

$$\forall i, i' = \overline{1,n} \colon d(e_i, e_{i'}) = \sum_{j=1}^{k} |z_{ij} - z_{i'j}|.$$

Distance Tchebyshev

La distance de Tchebychev, est la distance entre deux points donnée par la différence maximale entre leurs coordonnées sur une dimension. Elle est donnée par la formule ci-dessous

$$\forall i, i' = \overline{1,n} \colon d(e_i, e_{i'}) = \max_{j=1,\ldots,k} |z_{ij} - z_{i'j}|.$$

Distance de Minkowski

La distance de Minkowski est une mesure de distance entre deux point de l'espace vectoriel normé, c'est une généralisation de la distance euclidienne et de la distance de Manhattan, elle a été proposée par [11]. Elle est donnée par la formule ci-dessous

$$\forall i, i' = \overline{1,n} \colon d(e_i, e_{i'}) = \left(\sum_{j=1}^{k} |z_{ij} - z_{i'j}|^q \right)^{1/q} ; q \geq 1.$$

Remarque 1.3.2 ***1.*** *Dans le cas où la limite atteind l'infini, on obtient la distance de Tchebyshev*

$$\forall i, i' = \overline{1,n}\colon d(e_i, e_{i'}) = \lim_{q \to +\infty} \left(\sum_{j=1}^{k} |z_{ij} - z_{i'j}|^q \right)^{1/q} = \max_{j=1,\dots,k} |z_{ij} - z_{i'j}|.$$

2. *Lorsque le paramètre $q = 1$, nous avons la distance Manhattan.*

3. *La distance de Minkowski est généralement utilisée lorsque q est égal à 2, on obtient la distance euclidienne.*

Distance Mahalanobis [13]

Corrige les données pour les différentes échelles et des corrélations dans les variables, l'angle entre deux vecteurs peuvent être utilisés comme mesure de distance quand le regroupement des donnée est de haut dimension (l'espace produit scalaire). Elle donnée par la formule ci-dessous

$$\forall i, i' = \overline{1,n}\colon d(e_i, e_{i'}) = \sqrt{(z_i - z_{i'})^t V^{-1} (z_i - z_{i'})},$$

où V^{-1} est une matrice de covariance.

[1]

Remarque 1.3.3 ***1.*** *Si la matrice de covariance est la matrice identité, cette distance est simplement la distance euclidienne.*

2. *Si la matrice de covariance est diagonale, on obtient la distance euclidienne normalisée suivant :*

$$\forall i, i' = \overline{1,n}\colon d(e_i, e_{i'}) = \sqrt{\sum_{j=1}^{k} \frac{(z_{ij} - z_{i'j})^2}{\sigma_j^2}}.$$

[1] Produit scalaire : $a \cdot b = \sum_{i=1}^{n} a_i b_i$.

Distance de Canberra

La distance de Canberra est une mesure numérique de la distance entre paires de points dans un espace vectoriel, introduite par [6] et affinée par [7]. Elle est donnée par la formule ci-dessous

$$\forall i, i' = \overline{1,n} \colon d(e_i, e_{i'}) = \sum_{j=1}^{k} \frac{z_{ij} - z_{i'j}}{z_{ij} + z_{i'j}}.$$

1.3.4 Similarité entre objets décrits par des variables binaires

Lorsque n individus sont caractérisés par la présence ou l'absence de η attributs, une multitude d'indices de similarité ont été suggérés. Ces indices intègrent de différentes manières les quatre valeurs associées à chaque paire d'individus.

$y \diagdown x$	1	0	somme
1	a	c	$a+c$
0	b	d	$b+d$
somme	$a+b$	$c+d$	η

,

avec $\eta = a + b + c + d$.

La somme $(a + d)$ indique la concordance entre les v.a. et la somme $(b + c)$ indique la discordance.

- a le nombre de variables pour $x = y = 1$.
- b le nombre de v.a. pour $x = 1$ et $y = 0$.
- c le nombre de v.a. pour $x = 0$ et $y = 1$.
- d le nombre de v.a. pour $x = y = 0$.

Les coefficients d'associations les plus populaires sont :

- Coefficient de Jaccard [8]

$$d(x,y) = \frac{a}{a+b+c}.$$

- Coefficient d'appariement (matching) simple

$$d(x,y) = \frac{b+c}{a+b+c+d}.$$

- Coefficient de Russel et Rao [15]

$$d(x,y) = \frac{a}{a+b+c+d}.$$

- Coefficient de Sokal et Sneath [16]

$$d(x,y) = \frac{a}{a+2(b+c)}.$$

- Coefficient de Dice [3]

$$d(x,y) = \frac{2a}{2a+b+c}.$$

Exemple 1.3.1 *Soit deux objets x et y tels que*

$x = (0,1,0,0,1,1)$ *et* $y = (0,0,1,1,0,1)$ *avec* $a = 1, b = 2, c = 2, d = 1$, *alors*

- *Coefficient de Jaccard :* $d(x,y) = 1/(1+2+2) = 1/5$.

- *Coefficient d'appariement (matching) simple :* $d(x,y) = (2+2)/6 = 2/3$.

- *Coefficient de Russel et Rao :* $d(x,y) = 1/6$.

- *Coefficient de Sokal et Sneath :* $d(x,y) = 1/(1+2(2+2)) = 1/9$.

1.4 Méthodes hiérarchiques

Les fondements des méthodes de classification hiérarchique ont été établis par [9] et [10]. Cette approche de regroupement procède par des fusions successives, créant des clusters (éléments) de plus en plus larges en combinant, à chaque niveau, les éléments ou les ensembles d'éléments qui se trouvent être les plus similaires entre eux. Ce processus produit une hiérarchie ou un arbre de clusters [1]. Pour évaluer la similarité et guider le regroupement, cette méthode s'appuie sur une mesure de similarité, permettant ainsi de capturer l'homogénéité ou l'hétérogénéité au sein des groupes formés.

1.4.1 Hiérarchie

Une hiérarchie est constituée d'une suite de partitions imbriquées. Elle est typiquement représentée sous la forme d'un arbre hiérarchique, nommé dendrogramme, où les éléments individuels apparaissent à la base et la totalité de l'ensemble est placée au sommet.

Définition 1.4.1 *(**Hiérarchie**) On appelle hiérarchie H de Ω tout ensemble de parties de Ω vérifiant les propriétés suivantes :*

- $\emptyset \notin H$.
- $\Omega \in H$.
- $\forall e \in \Omega : e \in H$, *car la hiérarchie contient tous les singletons.*
- $\forall C_1, C_2 \in H : C_1 \cap C_2 = \emptyset$ *ou* $C_1 \subset C_2$ *et inversement car deux classes de la hiérarchie sont soit disjointes soit contenues l'une dans l'autre.*

Définition 1.4.2 (Hiérarchie de partie indicée) *C'est une hiérarchie de parties H à laquelle est associée une échelle d'indices qui satisfait les propriétés suivantes : Il existe une application i positive définie sur H telle que*

- $\forall e \in \Omega : i(e) = 0$.

- $A \subseteq B \Longrightarrow i(A) \leq i(B)$.

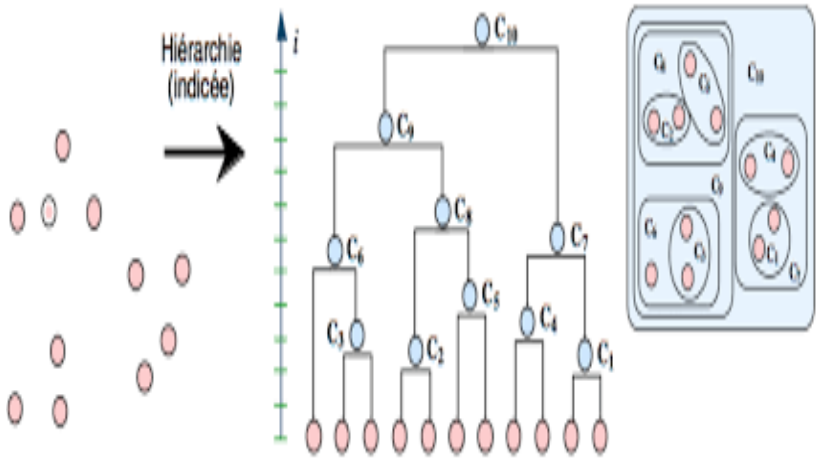

FIG. 1.1 – Classification hiérarchique de la partie indicée [5].

1.4.2 Classification ascendante hiérarchique (CAH)

La Classification ascendante Hiérarchique notée CAH constitue une approche statistique de regroupement d'éléments en clusters ou groupes, fondée sur la similitude entre ces éléments. Cette technique, comme d'autres méthodes, offre des bénéfices et présente des limites qui affectent son utilisation en fonction du contexte spécifique et des buts poursuivis par l'analyse.

1.4.2.1 Principe de la CAH

La CAH est une méthode de regroupement basée sur la similitude, construisant progressivement une structure hiérarchique de clusters à partir des données. Le processus commence par considérer chaque élément comme un cluster individuel, pour

ensuite mesurer et comparer les distances entre les clusters, en utilisant diverses méthodes de calcul de distance. Les étapes clés incluent l'identification et la fusion des paires de clusters les plus proches, la mise à jour des distances après chaque fusion, et la répétition de ces étapes jusqu'à la formation d'un unique grand cluster. Ce processus est visualisé à travers un dendrogramme, qui offre une représentation graphique des fusions et de la distance à laquelle elles se produisent, permettant de choisir le nombre de clusters pertinents pour l'analyse. La CAH est appréciée pour son approche exploratoire, révélant la structure interne des données et facilitant la compréhension des relations hiérarchiques entre les éléments.

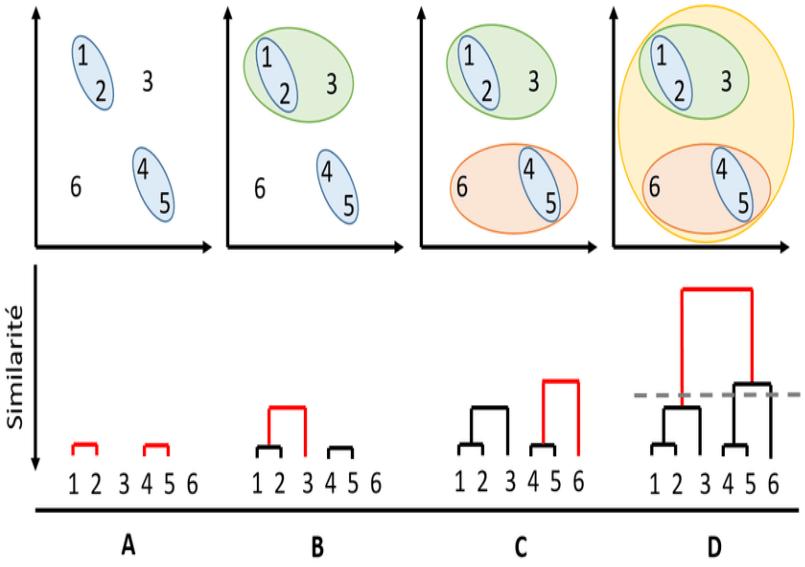

FIG. 1.2 – Classification ascendante.

1.4.2.2 Matrice de distance

Pour calculer la proximité entre différentes paires d'objets, on peut recourir tant aux distances qu'aux dissimilarités, regroupées dans ce qu'on appelle une matrice de distances ou de dissimilarités, notée D. Cette matrice est carrée et symétrique, de dimension n, définie positive et avec des zéros sur sa diagonale. Pour un ensemble

de n objets, il ya $n(n-1)/2$ distances à déterminer.

$$D = \begin{pmatrix} 0 & d(e_1,e_2) & \cdots & d(e_1,e_n) \\ d(e_2,e_1) & 0 & & d(e_2,e_n) \\ & & \ddots & \\ d(e_n,e_1) & d(e_n,e_2) & \cdots & 0 \end{pmatrix} \in M_{(n,n)}(\mathbb{R}).$$

Remarque 1.4.1 *Généralement pour cette matrice on l'utilise la distance euclidienne.*

1.4.3 Critères d'agrégation

Une fois les proximités entre objets évaluées, on peut décider de la manière dont ces objets seront regroupés. Pour ce faire, on pratique des méthodes de liaison qui associent les objets proches en paires au sein du même groupe binaire (la fusion des objets s'effectuant toujours par paires). Les groupes ainsi formés sont ensuite regroupés en de nouveaux ensembles, jusqu'à l'achèvement de l'arbre hiérarchique. Différentes méthodes de liaison sont citées ci-dessous.

1. Méthode de la liaison simple ou critère du saut minimal :

   L'agrégation D entre deux groupes C_1 et C_2 correspond au minimum des distances entre un élément du groupe C_1 et un élément du groupe C_2.

$$\boxed{D(C_1,C_2) = \min(\{d(z,y)\}; z \in C_1, y \in C_2).}$$

2. Méthode de la liaison complète ou critère du saut maximal :

   L'agrégation D entre deux groupes C_1 et C_2 correspond au maximum des distances entre un élément du groupe C_1 et un élément du groupe C_2.

$$\boxed{D(C_1,C_2) = \max(\{d(z,y)\}; z \in C_1, y \in C_2).}$$

3. Méthode de la liaison moyenne (Average Linkage) :

   Cette méthode définie la distance entre les groupes comme, la distance moyenne de tous les éléments d'un groupe à tous les points de l'autre groupe.

$$\boxed{D(C_1, C_2) = \frac{\sum_{z \in C_1} \sum_{y \in C_2} d(z, y)}{n_1 n_2},}$$

   où n_1 et n_2 est le nombre d'objets dans chacun des deux groupes C_1 et C_2 respectivement.

   Remarque 1.4.2 *Cette approche vise à regrouper les ensembles présentant de petites variances et favorise la création des groupes ayant des variances similaires ou égales.*

4. Méthode de centre de gravité (Centroid Method) :

   La distance entre deux classes C_1 et C_2 est définie par la distance entre leurs centres de gravité.

$$\boxed{D(C_1, C_2) = d(g_1, g_2),}$$

   où g_1 et g_2 sont les centres de gravité de C_1 et C_2 respectivement.

   Sachant que $g_j = \frac{1}{q_l} \sum_{e_i \in C_j} p_i e_i$ et $q_l = \sum_{e_i \in C_j} p_i$ (q_l est le poids de C_j).

5. Méthode de Ward :

   Cette méthode est basée sur la perte d'inertie expliquée résultant de l'agrégation des classes C_1 et C_2 voir.

$$\boxed{D(C_1, C_2) = \frac{n_1 n_2}{n_1 + n_2} d^2(g_1, g_2).}$$

Le graphe ci-dessous représente les differents critères d'agrégations.

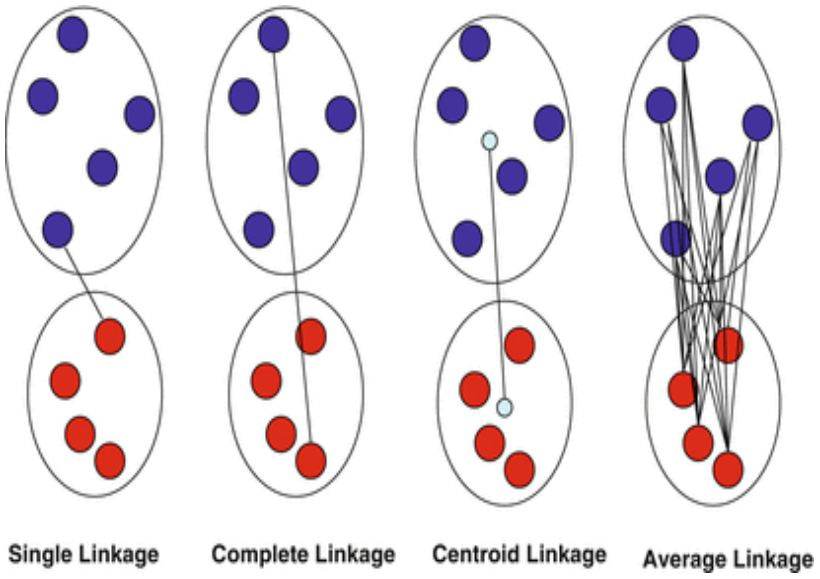

FIG. 1.3 – Différentes méthodes de liaison [14].

1.4.4 Inertie

En statistique, l'inertie quantifie la dispersion des données dans un espace multidimensionnel, de façon similaire à la variance qui mesure cette dispersion autour d'une moyenne dans un espace unidimensionnel. Utilisée principalement dans l'analyse multidimensionnelle, comme dans l'analyse en composantes principales (ACP), l'inertie se calcule en sommant les carrés des distances pondérées entre chaque point de donnée et le centre de gravité de l'ensemble, reflétant ainsi la distribution spatiale des données.

1. Inertie inter-classe : L'inertie inter-classe de la partition P_k notée $\mathfrak{I}_{inter}$ est l'inertie des centres de gravité des classes pondérées par q_l. Elle est donnée par la formule ci-dessous

$$\boxed{\mathfrak{I}_{inter} = \sum_{i=1}^{n} q_l d_M^2(g_i, g).}$$

2. Inertie intra-classe : L'inertie intra-classe de la partition P_k notée $\mathfrak{I}_{intra}$ est la somme des inerties des classes. Elle est donnée par la formule ci-dessous

$$\boxed{\mathfrak{I}_{intra} = \sum_{j=1}^{k} \sum_{e_i \in C_j} P_i d^2(e_i, g_j).}$$

3. Inertie totale : L'inertie totale du nuage des n individus notée $\mathfrak{I}_{tot}$ est donnée par la formule ci-dessous

$$\boxed{\mathfrak{I}_{tot} = \sum_{i=1}^{n} p_i d^2(e_i, g),} \ .$$

où g est le centre de gravité du nuage des individus.

On peut exprimer l'inertie totale nommée aussi **décomposition de Huygens** de la maniére ci-dessous

$$\boxed{\mathfrak{I}_{tot} = \mathfrak{I}_{inter} + \mathfrak{I}_{intra}.}$$

On constate que minimiser l'inertie intra-classe tout en maximisant l'inertie inter-classe est souvent l'objectif des méthodes de clustering comme l'algorithme K-means.

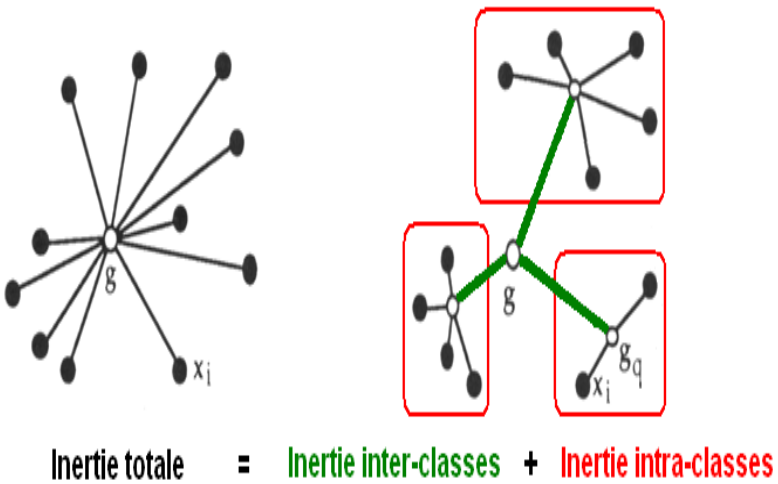

FIG. 1.4 – Décomposition de Huygens.

Remarque 1.4.3 *La méthode de Ward est la méthode la plus souvent utilisée car elle a été conçue dans un objectif de maximisation de l'inertie inter-classe et de minimisation de l'inertie intra-classe.*

1.4.5 Avantages et inconvénients de CAH

Les avantages et inconvénients de CAH sont cités ci-dessous :

Avantages :

- Interprétabilité des résultats.
- Pas besoin de spécifier le nombre de clusters.
- Flexibilité des critères de liaison.
- Applicable à divers types de données.
- Détecte les structures hiérarchiques.

Inconvénients :

- Sensibilité aux outliers.
- Coût computationnel élevé.
- Choix du critère de liaison.
- Difficulté de choix du nombre de clusters.
- Irreversibilité des étapes.

L'algorithme de la (CAH)

L'algorithme de la (CAH) est une technique de regroupement qui vise à assembler des objets similaires en clusters, en se basant sur la mesure de leur proximité ou dissimilarité. Voici les grandes lignes de cet algorithme :

1. Initialisation : On débute avec N objets, chacun formant un cluster individuel, ce qui nous donne N clusters initiaux.
2. Calcul des distances : On établit une matrice des distances entre tous les clusters, en utilisant diverses méthodes (euclidienne, Manhattan, etc.) selon le type de données traitées.
3. Identification des clusters proches : On repère la paire de clusters la plus proche, c'est-à-dire celle ayant la distance minimale entre eux.
4. Fusion : Cette paire de clusters est fusionnée pour n'en former qu'un, réduisant ainsi le nombre total de clusters de un.
5. Mise à jour des distances : La matrice des distances est actualisée pour intégrer les changements suite à la fusion, en recalculant les distances entre le nouveau cluster et les autres.
6. Répétition : Les étapes 3 à 5 sont répétées jusqu'à obtenir un seul cluster englobant tous les objets, ou jusqu'à atteindre le nombre de clusters désiré.
7. Résultat : Le processus est souvent visualisé à l'aide d'un dendrogramme, qui montre les étapes de fusion et aide à déterminer le nombre optimal de clusters.

La CAH est précieuse pour l'exploration des données, permettant de découvrir des structures cachées parmi les objets analysés.

1.4.3 Classification descendante hiérarchique (CDH)

La Classification Descendante Hiérarchique notée CDH est une méthode analytique conçue pour structurer un ensemble de données en groupes basés sur leur similarité, adoptant une démarche hiérarchique allant du général au spécifique. Initialement, tous les éléments sont considérés comme appartenant à un unique cluster global, qui est progressivement divisé en sous-groupes plus petits en fonction de leur dissimilarité, mesurée par des critères comme la distance euclidienne ou de Manhattan. Ce

processus de division se poursuit jusqu'à atteindre un critère d'arrêt prédéfini, tel qu'un nombre spécifique de clusters ou un seuil de dissimilarité. La méthode permet une visualisation claire de la structure des données via un dendrogramme, facilitant l'interprétation des relations entre les éléments. La CDH est particulièrement utile pour l'exploration des données, offrant une approche méthodique pour révéler la structure cachée au sein d'un ensemble de données sans présuppositions sur le nombre de groupes existants.

Principe de la CDH

La CDH, mettant en évidence son approche unique pour le clustering des données. En détaillant les phases d'initialisation, de division récursive et de critères d'arrêt, on a bien expliqué comment la CDH procède pour diviser un grand ensemble des données en clusters plus petits et plus significatifs. La mention des critères de séparation et des critères d'arrêt souligne l'importance de la mesure de dissimilarité et la flexibilité dans la détermination de la fin du processus de clustering. De plus, la référence à l'utilisation des dendrogrammes pour la visualisation et l'interprétation des résultats aide à comprendre comment la CDH facilite l'analyse des structures des données complexes.

Avantages et inconvénients de la CDH

Les avantages et inconvénients de CDH sont cités ci-dessous :

Avantages :

Contrairement à de nombreux algorithmes de classification automatique, l'algorithme de Classification Hiérarchique ne requiert pas la définition d'un seuil arbitraire pour la création des classes, évitant ainsi le risque de guider la recherche de partition vers une orientation peu réaliste.

Inconvénients :

Les résulats sont en générale grossiers, les niveaux des noeuds de la hiérarchie ne sont plus définis que par l'ordre dans lequel ils apparaissent.

1.4.4 Classification des variables

La classification des variables est un élément fondamental en statistique et en analyse des données, permettant de catégoriser les différents types des données selon leurs caractéristiques et la manière dont elles peuvent être analysées ou interprétées. Les variables peuvent être classées en plusieurs grandes catégories, notamment :

1.4.4.1 Variables qualitatives (catégorielles)

Ces variables décrivent des attributs ou des catégories qui ne peuvent pas être mesurées par des nombres de manière significative. Il existe deux types de variables quantitatives qui sont définies comme suit

- **Nominales** : Elles représentent des catégories sans ordre naturel entre elles, par exemple le genre (masculin, féminin), la couleur (rouge, bleu, vert).
- **Ordinales** : Elles décrivent des catégories avec un ordre ou un classement implicite., par exemple le niveau d'éducation (primaire, secondaire, universitaire), le classement (haut, moyen, bas).

1.4.4.2 Variables quantitatives

Ces variables représentent des quantités mesurables qui peuvent être exprimées numériquement.

- **Discrètes** : Elles représentent des comptages ou des valeurs qui peuvent être énumérées et qui sont souvent des nombres entiers, par exemple le nombre

d'enfants dans une famille, le nombre de voitures vendues par une concession.

- **Continues** : Elles peuvent prendre n'importe quelle valeur dans un intervalle donné, incluant des nombres décimaux ou irrationnels, par exemples le poids, la taille, la température, ect.

Chapitre 2

Classification non hiérarchique

La classification non hiérarchique, également connue sous le nom de classification partitionnelle, est une technique de clustering qui divise un ensemble de données en groupes basés sur leur similarité, sans créer de structure arborescente comme dans le cas de la classification hiérarchique. Cette méthode diffère de la classification hiérarchique en partitionnant directement l'ensemble des données en un nombre prédéterminé de clusters distinct. Son objectif principal est de regrouper les données de manière à ce que les éléments d'un même cluster soient plus similaires entre eux qu'aux éléments des autres clusters.Elle est efficace pour le traitement rapide de grands ensembles de données grâce à des méthodes comme l'algorithme K-means. Sa simplicité et sa facilité d'implémentation la rendent accessible.

Une méthode de classification non hiérarchique divise un groupe d'individus en plusieurs catégories en utilisant une approche d'optimisation itérative. Cette approche commence par créer une partition initiale et vise ensuite à l'améliorer en transférant les données d'une catégorie à une autre. Il serait peu pratique de considérer toutes les partitions possibles, donc ces algorithmes visent à trouver des optimums locaux en améliorant une fonction objectif. Cette fonction objectif reflète l'idée que les individus au sein d'une même catégorie devraient être similaires entre eux et différents

des individus d'autres catégories. Les catégories finales obtenues ne se chevauchent pas et sont distinctement séparées, chacune représentée par un noyau spécifique.

La classification non hiérarchique d'un ensemble d'individus peut être divisée en trois grandes sous-familles : K-means [12], centres mobiles [4] et Nuées dynamiques [2].

2.1 Principe de la classification non hiérarchique

La méthode de classification non hiérarchique vise à diviser un ensemble d'individus en plusieurs groupes distincts, sans établir de liens hiérarchiques entre ces groupes. Chaque groupe, ou classe, doit contenir au moins un individu, et chaque individu doit être assigné à une classe unique. Le critère principal pour évaluer l'efficacité de cette division repose sur la proximité des individus au sein d'une même classe et leur éloignement par rapport aux membres des autres classes. Cette approche repose sur l'utilisation des mesures de distance ou de similarité pour évaluer la proximité entre les éléments à classer.

Définition 2.1.1 (Partition) *Une partition notée P_k de Ω en k classes est un ensemble $\{C_1, \ldots, C_k\}$ de classes non vides, vérifiant :*

- $\bigcup_{i=1}^{k} C_i = \Omega$.
- $\forall i, j = \overline{1,k}; i \neq j : C_i \cap C_j = \emptyset$.

Exemple 2.1.1 *$\Omega = \{1, \ldots, 6\}$ et $P_3 = \{C_1, C_2, C_3\}$ avec $C_1 = \{1, 2\}, C_2 = \{3\}$ et $C_3 = \{4, 5, 6\}$ est une partition en trois classes.*

2.2 Différentes algorithmes

Les algorithmes de classification non hiérarchique sont catégorisés en trois grandes sous-familles.

2.1.1.1 Méthode de K-means (MacQueen)

La méthode de K-means, souvent associée à James MacQueen qui l'a introduit en 1967 [12], L'algorithme K-means est reconnu pour être l'un des algorithmes de classification automatique les plus élémentaires et populaires. Il fonctionne en représentant chaque classe par son centre, défini comme la moyenne de tous les éléments qui la composent. Cette technique propose une méthode simple et efficace pour diviser un jeu de données en un nombre k spécifié de groupes (classes).

La méthode repose sur un principe assez simple, centré sur le concept de centre de gravité. Elle consiste à associer chaque individu à une classe si celui-ci se trouve à proximité de son centre de gravité. Ce centre est défini comme étant la moyenne de la position de tous les individus appartenant à la classe.

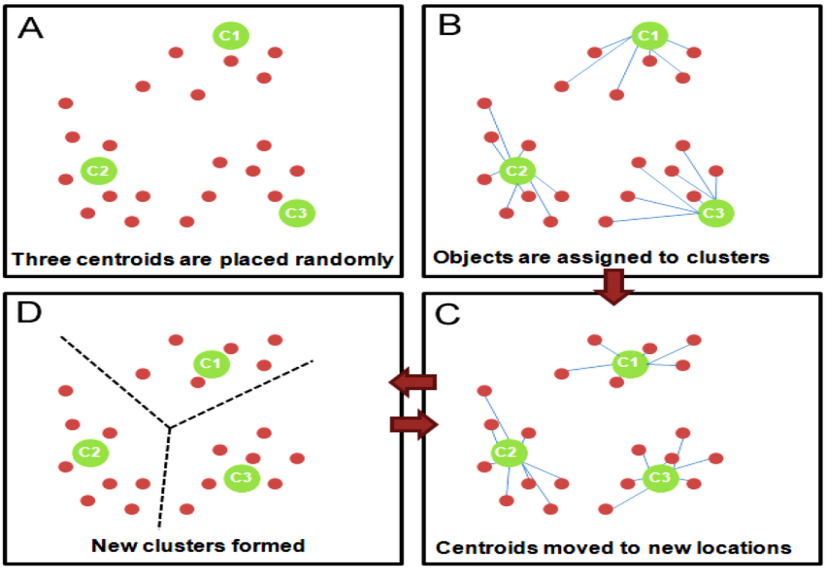

FIG. 2.1 – Méthode de K-means.

Algorithme :

La méthode de K-means est un processus de classification qui commence par la sélection aléatoire de k individus parmi la population à classer, ces individus servent de centres initiaux sans remplacement. À chaque fois qu'un individu est assigné à

une classe, la répartition des classes est ajustée. L'algorithme procède par ces étapes jusqu'à stabilisation des classes.

L'algorithme K-means se déroule selon les étapes suivantes :

Etape 1 **:** Choisir k individus au hasard (comme centre des classes initiales).

Etape 2 **:** Affecter chaque individu au centre le plus proche.

Etape 3 **:** Recalculer le centre de chacune de ces classes.

Etape 4 **:** Répéter l'étape 2 et 3 jusqu'à stabilité des centres.

Etape 5 **:** Editer la partition obtenue.

Avantages et inconvénients :

Avantages

- Simplicité et rapidité : K-means est facile à comprendre et à mettre en œuvre. Il est efficace sur de grands jeux de données, ce qui le rend pratique pour des applications réelles.
- Flexibilité : Il peut être utilisé avec divers types de données et dans différents domaines d'application, comme le marketing, la recherche scientifique, et l'analyse des données.
- Adaptabilité : Il fonctionne bien dans des situations où les clusters sont bien séparés et globalement sphériques.
- Optimisation : K-means vise à minimiser la variance au sein des clusters, ce qui peut conduire à des partitions de haute qualité lorsque les clusters sont compacts et bien séparés.

Inconvénients

- Spécifiez les valeurs k : le nombre de clusters k doit être spécifié à l'avance, ce qui peut être difficile sans une connaissance préalable des données.

- Traiter les données numériques : l'algorithme K-moyennes ne peut être exécuté que dans des données numériques.
- Problèmes de prédiction : Il est difficile de prévoir les valeurs k ou le nombre de clusters. Il est également difficile de comparer la qualité des clusters produites.

2.1.1.2 Méthode de Nuées dynamiques (Diday)

La méthode des Nuées Dynamiques, introduite par Edwin Diday en 1980 [2], est une technique de classification automatique qui appartient à la famille des méthodes de clustering, est effectivement centrée sur le partage d'une population en différentes catégories ou classes. Ce qui distingue cette méthode, c'est son utilisation de la notion de "noyau" associé à chaque classe. Ce noyau peut être compris comme le cœur ou le centre d'un cluster.

La méthode repose sur un principe assez simple, on débute par sélectionner aléatoirement k noyaux au sein d'un ensemble de noyaux, chaque noyau englobant un groupe d'individus spécifique. Ensuite, chaque élément de l'ensemble des données est attribué au noyau le plus proche, formant ainsi une division en k groupes pour lesquels les noyaux sont recalculés. Ce cycle de sélection de nouveaux noyaux et de réattribution des points continue jusqu'à ce que l'on observe plus d'amélioration significative dans la qualité de la division des données en groupes.

Algorithme :

Etape 1 **:** Sélectionner k points comme centres initiaux.

Etape 2 **:** Utiliser ces centres pour diviser l'ensemble en partitions.

Etape 3 **:** Mettre à jour chaque centre pour qu'il corresponde au barycentre de sa partition.

Etape 4 **:** Répéter l'étape 2 avec les nouveaux centres.

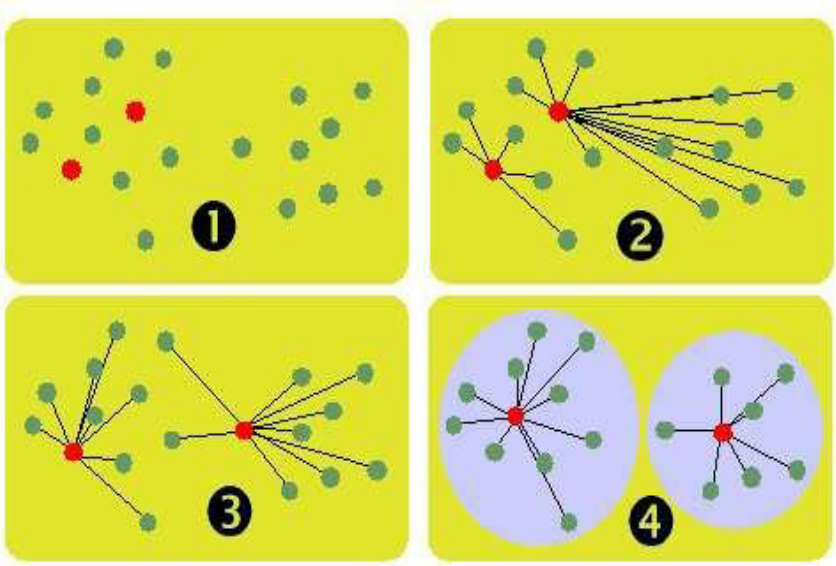

FIG. 2.2 – Méthode des nuées dynamiques.

Avantages et inconvénients :

Avantages

- Améliore progressivement la qualité des groupes formés.
- Adapté pour traiter de larges ensembles de données.
- Gérer efficacement de vastes groupes d'individus en peu de temps.

Inconvénients

- Le nombre de groupes k est fixé à l'avance.
- La partition finale est influencée par les choix initiaux, ce qui exclut la garantie d'une solution optimale globale.
- Performe principalement avec des données de forme sphérique.

2.1.1.3 Méthode du centre mobile (Forgy)

La méthode des centres mobiles, développée par Forgy [4], est l'une des approches les plus traditionnelles et les plus largement adoptées pour la classification. Cette technique est pertinente lorsque le nombre de catégories désirées est connu à l'avance,

un nombre que nous désignerons par k. Les individus sont représentés par des points dans un espace, leurs positions définies par diverses mesures (dans un espace de dimension 2 ou plus). L'objectif est de regrouper les points similaires aussi étroitement que possible tout en maintenant une séparation claire entre les différents groupes. La sélection des groupes se fait de manière automatique, basée uniquement sur les données, en analysant les similitudes et les différences entre les points.

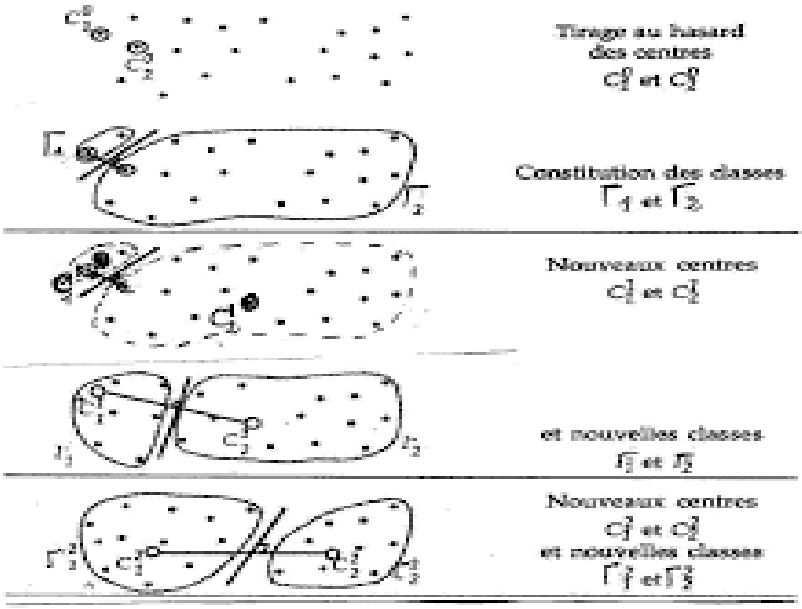

FIG. 2.3 – Méthode du centre mobile.

Algorithme :

Etape 1 **:** Sélectionner aléatoirement k individus qui seront utilisés comme centres initiaux des classes.

Etape 2 **:** Assigner chaque individu au centre le plus proche, formant ainsi une partition initiale en k classes notée $P_0 = \{C_1, C_2, ..., C_k\}$.

Etape 3 **:** Calculer de nouveaux centres pour chaque classe constituée dans P_0.

Etape 4 **:** Répéter les étapes 2 et 3 jusqu'à obtenir une partition identique lors de deux itérations consécutives.

Avantages et inconvénients :

Avantages

- Simplicité d'implémentation.
- Efficacité pour les grands jeux de données.
- Adaptabilité.
- Bonne performance sur des données bien séparées.

Inconvénients

- Sensibilité aux valeurs initiales.
- Difficulté avec des formes de clusters non sphériques.
- Nombre de clusters à spécifier à l'avance.

Conclusion

Dans cette étude, on a exploré deux méthodes d'analyse statistique exploratoire multidimensionnelle à savoir la classification hiérarchique et la classification non hiérarchique. La classification hiérarchique est l'une des principales approches de la classification automatique, tout comme la classification non hiérarchique. La principale différence entre ces deux méthodes réside dans le fait que la classification hiérarchique offre la possibilité de visualiser un éventail de nombres possibles de groupes et de sélectionner celui qui répond le mieux aux besoins de l'étude. En revanche, la classification non hiérarchique commence avec un nombre prédéterminé de groupes et se concentre sur leur répartition optimale. L'objectif de ces techniques de classification est de fournir une représentation schématique simple, telle qu'un dendrogramme, du tableau initial en répartissant les individus en groupes homogènes.

Bibliographie

[1] Celeux, G., Diday, E., Govaert, G., Lechevallier, Y., Ralambondrainy, H. (1989). *Classification automatique des données, environnement statistique et informatique*. DUNOD informatique.

[2] Diday, E. (1971). *Une nouvelle méthode en classification automatique et reconnaissance des formes la méthode des nuées dynamiques*. Revue de statistique appliquée, vol. 19, No 2, p. 19-33.

[3] Dice, L., R. (1945). *Measures of the amount of ecologic association between species*. Ecology, vol. 26, No. 3, p. 297-302.

[4] Forgy, E., W. (1965). *Cluster analysis of multivariate data : effciency versus interpretability of classifications*. Biometrics, vol. 21, p. 768-769.

[5] Gasso, G., Leray, P. Clustering. https ://www.iro.umontreal.ca/~mignotte/IFT6150/ComplementCours/clustering.pdf. INSA Rouen - Département ASI, Laboratoire LITIS.

[6] Godfrey, N., L., William, W., T. (1966). *Computer programs for hierarchical polythetic classification (similarity analysis)*. Computer Journal, vol. 9, No. 1, p. 60-64.

[7] Godfrey, N., L., William, W., T. (1967). *Mixed-data classificatory programs (Agglomerative Systems)*. Australian Computer Journal, vol. 1, no. 1, p. 15-20.

[8] Jaccard, P. (1908). *Nouvelles recherches sur la florale distribution*. Bul. Soc. Science Vaudoise Naturales, vol. 44, p. 223-270.

[9] Johnson, S., C. (1967). *Hierarchical Clustering Schemes*. Psychometrika, No 2, p 241-254.

[10] Lance, G.N., Williams, W.T. (1967). *A general theory of classificatory sorting strategies (Hierarchical systems)*. Computer Journal, No 9, p 373-380.

[11] Minkowski, H. (1896). *Sur les propriétés des nombres entiers qui sont dérivées de l'intuition de l'espace*. Nouvelles annales de mathématiques. journal des candidats aux écoles polytechnique et normale, Vol. 15, p. 393-403.

[12] Mac-Queen, J., B. (1967). *Some Methods for classification and Analysis of Multivariate Observations*. Proceedings of the fifth Berkeley symposium on mathematical statistics and probability, Vol. 1, No. 14, p. 281-297.

[13] Mahalanobis, P., C. (1936). *On the generalised distance in statistics*. Proceedings of the National Institute of Sciences of India, Vol. 2, p. 49-55.

[14] Peng.L., Yaqing S.(2014). *Cluster Analysis of RNA-Sequencing Data.*

[15] Russel, P., F., RAO, T., R. (1940). On habitat and association of speciesof anophelinae larvae in south-eastern Madras. J. Malaria Inst. India, Vol. 3, No. 1, p. 153-178.

[16] Sneath, P., H., Sokal, R., R. (1973). *The principles and practice of numerical classification*. Numerical taxonomy.

Annexe A : Logiciel R

Qu'est-ce-que le langage R ?

– Le langage R est un langage de programmation et un environnement mathématique utilisés pour le traitement de données. Il permet de faire des analyses statistiques aussi bien simples que complexes comme des modèles linéaires ou non-linéaires, des tests d'hypothèse, de la modélisation de séries chronologiques, de la classification, etc. Il dispose également de nombreuses fonctions graphiques très utiles et de qualité professionnelle.

– R a été créé par Ross Ihaka et Robert Gentleman en 1993 à l'Université d'Auckland, Nouvelle Zélande, et est maintenant développé par la R Developement Core Team. L'origine du nom du langage provient, d'une part, des initiales des prénoms des deux auteurs (Ross Ihaka et robert Gentleman) et, d'autre part, d'un jeu de mots sur le nom du langage S auquel il est apparenté.

Annexe B : Abréviations et Notations

Les différentes abréviations et notations utilisées tout au long de ce mémoire sont expliquées ci-dessous :

$\mathbb{R}$: Nombre réel.

Ω : Ensemble des individus.

n : Nombre de lignes.

k : Nombre de colonnes.

Z : Matrice de n lignes et k colonnes.

z_j : Vecteur variable.

e_i : Vecteur individu.

z_{ij} : Observation de z_j sur e_i.

p : Pois.

I_n : Matrice indicatrice.

Z_c : Matrice des données centrées.

Z_r : Matrice des données centrées réduites.

V : Matrice de variance-covariance.

$\overline{z}$: Moyenne arithmétique.

Cov : Covariance.

Var : Variance.

a, b, c, d : Attributs

σ : Ecart type.

V^{-1} : Inverse de la matrice de variance-covariance.

D_σ : Matrice diagonale de la racine de la variance.

D_σ^{-1} : Inverse de matrice diagonale de la racine de la variance.

Cor : Corrélation.

R : Matrice de corrélation.

d : Distance.

η : Nombre totale des attributs.

H : Hiérarchie.

$v.a.$: variables aléatoires.

D : Matrice de distance.

C_1, C_2 : Classes de hiérarchie.

P_k : Partition.

$\mathfrak{I}_{inter}$: Inertie inter-classe.

$\mathfrak{I}_{intra}$: Inertie intra-classe.

$\mathfrak{I}_{tot}$: Inertie totale.

CAH : Classification ascendante hiérarchique.

CDH : Classification descendante hiérarchique.

$i.e.$: c'est-à-dire

Résumé

La méthode de classification est une approche statistique exploratoire utilisée pour classer des individus en fonction de leurs caractéristiques. Elle se divise en deux types principaux : la classification hiérarchique et la classification non hiérarchique. Dans la classification hiérarchique, qui peut se faire de manière ascendante ou descendante, on regroupe progressivement des objets ou groupes d'objets similaires en classes de granularité décroissante, sans spécifier un nombre d'objets à l'avance. En revanche, la classification non hiérarchique nécessite la détermination du nombre de groupes dès le début et se distingue par l'emploi de nombreux algorithmes différents.

Mots clés : Classification; Hiérarchie; Similarité; Dissimilarité; Distance euclidienne; Dendrogramme; Objet; Centre mobile; Centre de gravité.

Abstract

The classification method is an exploratory statistical approach used to classify individuals based on their characteristics. It is divided into two main types: hierarchical classification and non-hierarchical classification. In hierarchical classification, which can be done in an ascending or descending manner, we progressively group similar objects or groups of objects into classes of decreasing granularity, without specifying a number of objects in advance. In contrast, non-hierarchical classification requires determining the number of groups from the start and is characterized by the use of many different algorithms.

Key words: Classification; Hierarchy; Similarity; Dissimilarity; Euclidean distance; Dendrogram; Object; Mobile center; Center of gravity.

ملخص

طريقة التصنيف هي طريقة إحصائية استكشافية تستخدم لتصنيف الأفراد على أساس خصائصهم. وينقسم إلى نوعين رئيسيين: التصنيف الهرمي والتصنيف غير الهرمي. في التصنيف الهرمي، والذي يمكن إجراؤه بطريقة تصاعدية أو تنازلية، نقوم تدريجيًا بتجميع الكائنات المتشابهة أو مجموعات الكائنات في فئات ذات دقة متناقصة، دون تحديد عدد من الكائنات مسبقًا. في المقابل، يتطلب التصنيف غير الهرمي تحديد عدد المجموعات من البداية ويتميز باستخدام العديد من الخوارزميات المختلفة.

الكلمات المفتاحية: التصنيف، التسلسل الهرمي، التشابه، الاختلاف، المسافة الإقليدية، مخطط شجري، العنصر، مركز متنقل، مركز الثقل .