République Algérienne Démocratique et Populaire

Ministère de l'enseignement supérieur et de la recherche scientifique

Université 20 Aout 1995 - SKIKDA

Faculté des Sciences

Département d'informatique

Mémoire de fin d'études en vue de l'obtention du diplôme de

Master professionnel en Informatique

Option : Réseaux et Systèmes Distribues (RSD)

Thème

Analyse et interprétation de la catégorisation des clients d'un mall

Réalisé par :

INNAL Assila

MERZOUK Abir

Encadre par :

Dr. RAMDANE Chafika

Session : Juin 2023

REMERCIEMENT

Nous Remercions en tout premier lieu ALLAH le tout puissant, qui nous 'a donné la force, la volonté et le courage pour accomplir ce modeste travail.

Nous avons abouti un travail, qui a été le résultat d'un cheminement de tout un parcours pédagogique, qui a duré tout le long de notre parcours éducatif dans l'enseignement supérieur

Nous tiendrons à remercier très chaleureusement Dr.Ramdane qui nous a permis de bénéficier de son encadrement. Ces conseils, Son œil critique, sa patience et son dévouement ont été déterminants dans la réalisation de ce modeste travail.

Un remerciement particulier à nous très chers parents, frères, sœurs, collègues et amies respectives qui nous ont encouragés, soutenu durant tout notre parcours.

Dédicace

Ce mémoire est dédié :

A mes très chers parents, que Dieu tout puissant les protège.

A mes frères et mes sœurs

A mes amis et mes copains

Résumé :

Dans ce travail de master, nous traitons le problème de la catégorisation de clients basé sur le clustering. Le clustering de données est une tâche fondamentale pour un grand nombre de domaines différents. Il s'agit d'une démarche très courante qui permet de mieux comprendre l'ensemble de données analysé. Ce problème peut être modélisé comme un problème d'optimisation. Pour sa résolution, nous optons pour l'algorithme de la recherche de coucou.

Nous appliquons l'algorithme à un ensemble de données des clients d'un mall.

Des expérimentations, des tests, des analyses et interprétations des résultats vont etre présentés pour un enrichissement du travail.

Mots-clés: catégorisation de clients, clustering de données, recherche de coucou, métaheuristique, optimisation.

Abstract:

In this master work, we deal with the problem of customer categorization based on clustering. Data Clustering is a fundamental task for a large number of different fields. This is a very common approach that helps to better understand the dataset being analyzed. This problem can be modeled as an optimization problem. For its resolution, we opted for the cuckoo search algorithm.

We apply the algorithm to a data set of mall customers.

Experiments, tests, analyzes and interpretation of the results will be presented to enrich the work.

Keywords: customer categorization, data clustering, cuckoo search, met-heuristic, optimization.

Table des matières

Chapitre 2 : la Recherche de Coucou

Tables des matières

Chapitre 3 : La conception de l'application

Chapitre 4 : Implémentation de l'approche

Tables des matières

Liste des Figures

Liste des figures

Liste des tableaux

Introduction Générale

Introduction générale :

Ces dernières années les volumes de données de toutes sortes croît de plus en plus. Avec tant de données disponibles les humains ont alors recours à la catégorisation comme l'une des activités mentales les plus primitives de l'organisation des données, afin qu'ils puissent en tirer des informations importantes.

Le clustering connu aussi sous le nom de catégorisation est un thème de recherche majeur en analyse et en fouille de données, permet de réduire les données en un ensemble de représentants moins nombreux permettant une représentation simplifiée des données initiales. Le problème de clustering a été adressé dans de nombreux contextes et par de nombreux chercheurs dans beaucoup de disciplines, cela reflète son large appel et utilité comme une des étapes dans l'analyse de données exploratoire. Ces applications sont nombreuses, en statistique, traitement d'image, intelligence artificielle, reconnaissance des formes, l'analyse du Web, le marketing, le diagnostic médical, la biologie etc.

La problématique principale traitée dans ce mémoire est celle d'analyser et diviser les clients d'un mall en catégories qui partagent des caractéristiques similaires. La catégorisation de la clientèle peut être un moyen puissant d'identifier les besoins insatisfaits des clients. Les entreprises peuvent alors surpasser la concurrence en développant des produits et services particulièrement attrayants et développer des campagnes marketing personnalisées.

Pour ce faire, nous allons opter pour la modélisation de clustering comme un problème d'optimisation et l'utilisation de l'algorithme de la recherche de coucous pour la résolution du problème de la catégorisation des clients d'un mall afin d'en extraire des informations utiles sur les différentes catégories de clients qui existent.

A la fin, nous présentons des analyses et des interprétations intensives, qui pourraient être utiles à une variété de décideurs différents.

Par conséquent, ce mémoire est reparti en quatre chapitres

Chapitre 1 :

Dans ce chapitre, nous présentons le domaine de clustering des données et ces différentes techniques

Chapitre 2 :

Dans ce chapitre, nous donnons des idées sur la méthode de la recherche de coucou, en présentant l'habitat, le comportement de l'oiseau coucou, et aussi plus principalement les étapes de l'algorithme et son déroulement.

Chapitre 3 :

Dans ce chapitre, nous passons à la conception de notre système, où nous présentons la majorité de modules de ce système.

Chapitre 4 :

Le quatrième chapitre présente le déroulement et les différentes parties de notre application et les différents outils et l'environnement de développement utilisés. Nous présentons également, les différentes représentations graphiques, ainsi qu'une analyse et interprétation détaillée des résultats.

Nous terminons par une conclusion.

Chapitre 1
Clustering des données

Chapitre 1 : Les clustering des données

1.1. Introduction :

La classification non supervisée ou plus communément connue sous le nom de clustering des données est un processus d'analyse exploratoire qui divise les données en un ensemble de groupes ou clusters homogènes. Ces clusters découverts peuvent être employés pour expliquer des caractéristiques de la distribution de données sous-jacente. Le clustering permet de réduire les données en un ensemble de représentants moins nombreux permettant une représentation simplifiée des données initiales. Il s'agit d'une démarche très courante qui permet de mieux comprendre les données analysées.

Du fait de son importance en analyse exploratoire de données, le problème de clustering de données a été traité dans de nombreux contextes et par de nombreux chercheurs dans beaucoup de disciplines

Dans ce chapitre en va voir quelques définitions de clustering, puis la notion de clustering et similarité, aussi les déférentes techniques de clustering avec quelque métaheuristique pour le clustering, finalement les techniques de validation de clustering.

1.2. Définition du clustering :

Un cluster est un ensemble de points tel que n'importe quel point dans le cluster est plus proche (ou plus similaire) de chaque autre point dans le cluster que de n'importe quel point qui n'est pas dans le cluster. Parfois un seuil est employé pour spécifier que tous les points dans un cluster doivent être suffisamment proches (ou similaires) l'un de l'autre. **[1]**

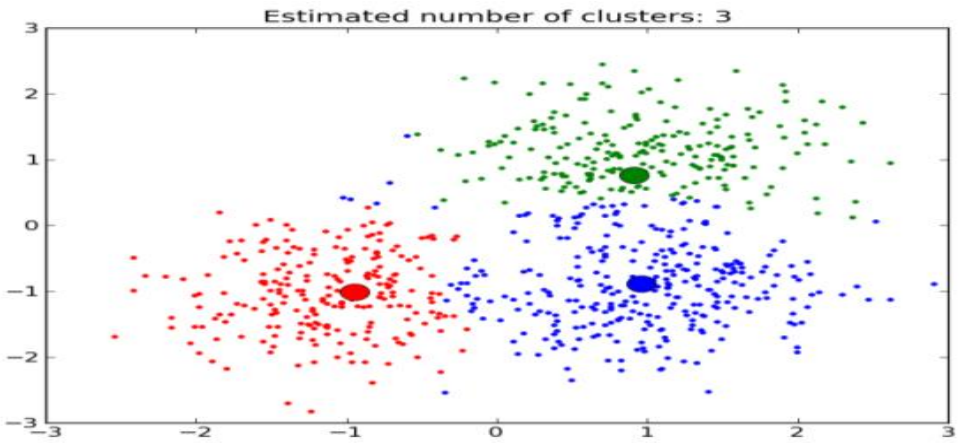


Figure 1.1. Le clustering.

1.3. Définition d'un cluster :

1.3.1. Définition d'un cluster bien séparé :

Est un cluster dans lequel les objets ont peu de similarité avec les objets des autres clusters. Autrement dit, les objets d'un cluster bien séparé sont fortement groupés ensemble et sont clairement distincts des objets dans les autres clusters. Ce concept est souvent utilisé pour évaluer la qualité d'un algorithme de clustering, car un bon algorithme doit être capable de produire des clusters bien séparé qui reflètent les structures cachées dans les données.

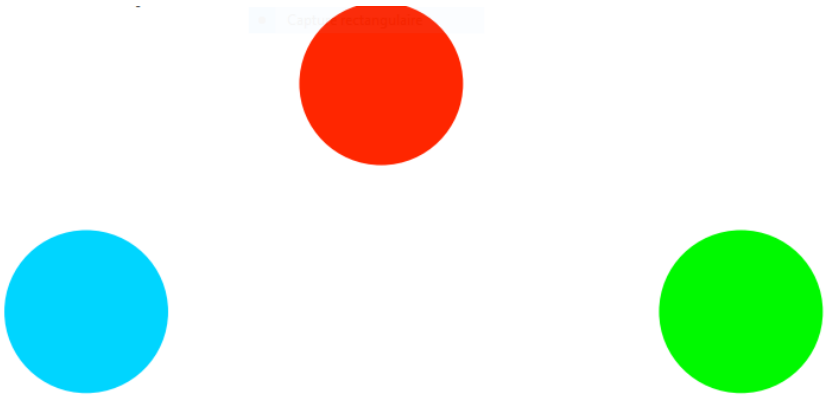


Figure 1.2. Clusters bien séparé

1.3.2. Définition d'un cluster basé sur un centre :

Est un cluster dans lequel chaque cluster est représenté par un objet central appelé center de cluster. Les autres objets du cluster sont associés au centre en fonction de leur similarité ou de

leur proximité. Les algorithmes de clustering basés sur le centre utilisent généralement une métrique de distance pour mesurer la proximité entre les objets et le centre de cluster. Le modèle final dépend de la position des centres et des associations entre. Les algorithmes de clustering basés sur le centre incluent k-means et le modèle de mélange gaussien.

Figure 1.3. Quatre clusters basés sur le centre

1.3.3. Définition de Cluster contiguë :

Un cluster contigus désigne un cluster ou les objets sont groupés de maniéré a former une forme continue ou une région contiguë dans l'espace des données. Autrement dit, les objets d'un cluster contiguous sont proches les uns des autres et forment une région cohérente est souvent utilisé dans les algorithmes de clustering qui utilisent des techniques de partitionnement de l'espace pour diviser les données en clusters .Les algorithmes qui produisent des clusters contigus sont utiles pour les applications qui impliquent la segmentation de l'image ou la géographie.

Figure 1.4. Huit clusters contigus

1.3.4. Définition basée sur la densité :

Est une approche de clustering qui regroupe les objets qui sont proches les uns des autres dans l'espace des données en formant des régions de haute densité. Cette approche considère que les clusters sont des régions ou la densité de rapport à leur environnement immédiat. Les algorithmes basés sur la densité, tels que DBSCAN et OPTICS, utilisent des méthodes pour mesurer la densité dans l'espace des données et trouver les régions de haute densité pour former les clusters. Cette approche est particulièrement utile pour les données de grande dimension et

les données complexes, car elle est capable de détecter des structures complexes dans les données.

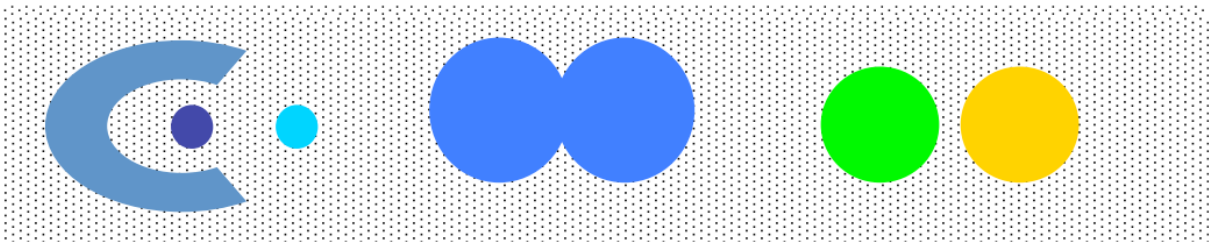


Figure 1.5. Six clusters denses

1.4. Objectifs du clustering :

Mirkin [2] a identifié une liste d'objectifs de clustering :

- Structuration : représentation des données comme un ensemble de groupes d'objets similaires (généralement, la structuration est l'objectif principal du clustering).
- Description : des clusters en fonctions des caractéristiques.
- Association : la découverte des interrelations entre différents aspects du phénomène.
- Généralisation : un relevé de données sur les propriétés du phénomène que les données ont des relations avec elles, cet objectif nécessite une analyse multi-étage des objectifs précédents.

Visualisation : représentation des structures des clusters comme une image visuelle.

1.5. Les concepts de clustering :

1.5.1. La matrice de données :

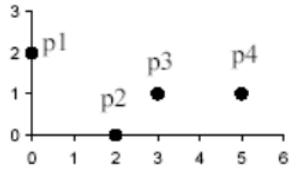
Les objets (échantillons, mesures, modèles, événements) sont habituellement représentés comme des points (vecteurs) dans un espace multidimensionnel, où chaque dimension représente un attribut distinct (variable, mesure) décrivant l'objet. Ainsi, un ensemble d'objets est représenté comme une matrice m×n, avec m lignes, une pour chaque objet et n colonnes, une pour chaque attribut. Cette matrice est appelée matrice de données ou jeu de données.

La figure1.6, ci-dessous, fournit un exemple concret de quelques points et leur matrice de données correspondantes.

1.5.2. La matrice de proximité :

Plusieurs algorithmes de clustering utilisent la matrice de données originale et beaucoup

D'autres emploient une matrice de similarité, ou une matrice de dis similarité. Pour la convenance, les deux matrices sont généralement mentionnées comme une matrice de proximité, P. Une matrice de proximité, P, est une matrice m×m contenant toutes les dis similarités ou les similarités entre les objets considérés. Si p i et p j sont le i ème et le j ème objets, respectivement, alors l'entrée à la i emme ligne et la j emme colonne de la matrice de proximité est la similarité, ou la dis similarité, entre pi et p j. La figure 1.6 montre, respectivement, quatre points, leur matrice de données et leur matrice proximité correspondante.



point	x	y
p1	0	2
p2	2	0
p3	3	1
p4	5	1

	p1	p2	p3	p4
p1	0.000	2.828	3.162	5.099
p2	2.828	0.000	1.414	3.162
p3	3.162	1.414	0.000	2.000
p4	5.099	3.162	2.000	0.000

Figure 1.6. Quatre points, leur matrice de données et leur matrice de proximité.

1.5.3. Types et échelles de données :

La mesure de proximité et le type de clustering utilisé dépendent des types et échelles des attributs de données **[3].** Les trois types d'attributs sont montrés dans le tableau 1.1, tandis que les différentes échelles de données sont montrées dans le tableau 1.2.

Tableau 1.1. Les différents types d'attributs

Binaire	Deux valeurs possibles, vraies ou fausses
Discret	Nombre de valeur finie ou d'entiers
Continu	Nombre de valeurs infinies ou de réels

Tableau 1.2. Les différentes échelles de données

Qualitative	Nominal	les valeurs sont juste des noms différents. Par exemple : les codes postaux, les couleurs, le sexe.
	Ordinal	les valeurs reflètent un ordre, rien plus. Par exemple : bon, meilleur, mieux ou couleurs ordonnées par le spectre.
Quantitative	Intervalle	la différence entre les valeurs est significative par exemple, l'intervalle de température.
	Ratio	rapport entre deux grandeurs. par exemple : les quantités monétaires, comme le salaire et le bénéfice et beaucoup de quantités physiques comme courant électrique, pression, etc.

1.5.4. Distance et similarité :

La plupart des algorithmes de clustering utilisent une mesure de la proximité des objets entre eux pour traiter. Cette notion de "proximité" est formalisée à l'aide d'une mesure de similarité et de dis similarité ou encore par une distance.

La similarité permet de mesurer la ressemblance entre objets, elle doit vérifier la positivité ($S_{i,j} \geq 0$) et la symétrie ($S_{i,j} = S_{j,i}$) tel que : $S_{i,j}$une similarité dans Rp entre deux objetsx_i, x_j .
La distance est la mesure la plus utilisée parmi les types de mesures de similarité et de dissimilarité, elle permet de vérifier les propriétés
suivantes **[4]** :

La positivité : $d_{i,j}$=d(x_i , x_j) ≥ 0.

La réflexivité : $d_{i,j}$=0 ⇔ x_i=x_j .

L'identité : $d_{i,j}$ =0

La symétrie : $d_{i,j}$=$d_{j,i}$.

L'inégalité triangulaire : $d_{i,k}$+ $d_{k,j} \geq d_{i,j}$,Tel que di, j une distance dans Rp entre deux objets

x_i , y_j

Les mesures de distance les plus courantes entre deux objets x_i et y_j sont

Distance Euclidienne $$d(x_i, y_j) = \sqrt{\sum_{i=1}^{n}\sum_{j=1}^{n}(x_i - y_j)^2} \quad (1.1)$$

Distance de Hamming $$d\,(x_i, y_j) = \sum_{i=1}^{n}\sum_{j=1}^{n} \left|x_i - y_i\right| \quad (1.2)$$

Distance de chebyschev $$d(x_i - y_i) = \max i = 1.2 \ldots n \left| x_i - y_j \right| \quad (1.3)$$

$$j = 1{,}2$$

Distance de Minkowski $$d(x_i, y_j) = \sqrt[p]{\sum_{i=1}^{n}\sum_{j=1}^{n}(x_i - y_j)^2} \quad , p > 0 \quad (1.4)$$

Distance de Canberra $$d\,(x_i, y_j) = \sum_{i=1}^{n}\sum_{j=1}^{n} \frac{|x_i - y_i|}{x_i + y_i} \quad \begin{matrix} x_i > 0 \\ et \\ x_j > 0 \end{matrix} \quad (1.5)$$

Séparation angulaire $$d(x_i, y_j) = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} x_i x_j}{[\sum_{i=1}^{n} {x_i}^2 \sum_{j=1}^{n} {y_i}^2]^{\frac{1}{2}}} \quad (1.6)$$

1.6. Techniques principale de clustering :

Plusieurs techniques de clustering sont apparues dans la littérature afin de découvrir des groupes cohésifs. Elles peuvent être classifiées aux types suivants **[1].**

1.6.1. Clustering par partitionnement :

Les techniques par partitionnement créent un partitionnement des points de données, d'un seul niveau. Si *k* est le nombre de clusters, alors les approches par partitionnement trouvent typiquement tous les *k* clusters immédiatement. Les techniques par partitionnement sont divisées en deux sous-catégories principales les algorithmes basés sur les centroïdes et les algorithmes basés sur les médoïdes. Nous allons décrire les deux algorithmes les plus connus : Kmeans et Kmedoid. Ces deux techniques sont basées sur l'idée qu'un centre représenter un cluster. Pour Kmeans on emploie la notion du centroïde qui est le point de la moyenne pour les autres points du cluster. Pour Kmedoid on utilise la notion d'un médoïde qui est le point le plus central d'un groupe de points de cluster. **[1] [5].**

A. Kmeans :

Le clustering K-means **[6.7]** est l'algorithme d largement utilisé. Il commence par choisir K points représentatifs comme centroïdes initiaux. Chaque point est ensuite attribué au centroïde le plus proche en fonction d'une mesure de proximité particulière choisie. Une foi est mise à jour. L'algorithme répète ensuite ces deux étapes de manière itérative jusqu'à ce que les centroïdes ne changent pas ou qu'un autre critère de convergence relâché alternatif soit satisfait.

Algorithme Clustering K-Means

1. Sélectionnez K points comme centroïdes initiaux.
2. Répéter

   2.1. Formez K clusters en affectant chaque point à son centroïde le plus proche.

   2.2. Recalculez le centroïde de chaque cluster.

3. Jusqu'à ce que le critère de convergence soit satisfait.

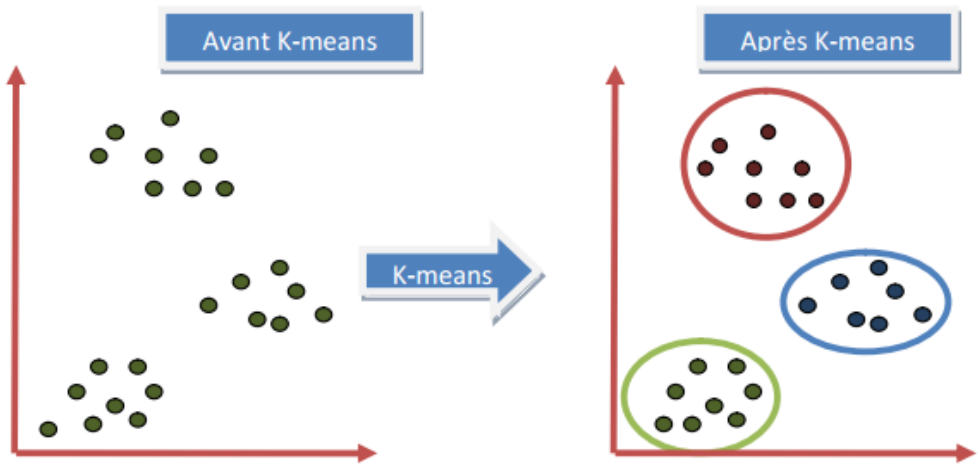


Figure 1.7. Algorithm K-means

B. Kmedoid:

Dans l'algorithme ou bien l'approche K-medoid, un cluster est représenté par un de ses points. Ce point représentatif est appelé médoïde, c'est un point qui est le plus placé au centre par rapport à l'application de quelques mesures, par exemple la distance. L'algorithme est conceptuellement simple. Il est décrit comme suit :

1. Choisir *k* points initiaux. Ces points sont les médoïdes candidats qui sont destinés à être les points les plus centraux de leurs clusters.
2. Considérer l'effet de remplacer un des points choisis (médoïdes) avec un des points non choisis. Conceptuellement, ceci est fait de la façon suivante:

   On calcule la distance entre chaque point non choisi et le médoïde candidat le plus proche et on calcule la somme de toutes les distances, cette somme représente le "coût" de la configuration actuelle. Tous les échanges possibles d'un point non choisi par un autre choisi sont considérés, et le coût de chaque configuration est calculé.
3. Choisir la configuration avec le coût le plus bas. Si c'est une nouvelle configuration, alors répéter l'étape 2.
4. Sinon, associer chaque point non choisi au point choisi le plus proche (médoïde) et arrêter.

1.6.2. Clustering hiérarchique :

Le clustering hiérarchique est une famille générale d'algorithmes de clustering qui construisent des clusters imbriqués en les fusionnant ou en les divisant successivement. Cette hiérarchie de clusters est représentée sous la forme d'un arbre (ou dendrogramme). La racine de l'arbre est l'unique cluster qui rassemble tous les échantillons, les feuilles étant les clusters avec un seul échantillon **[8]**

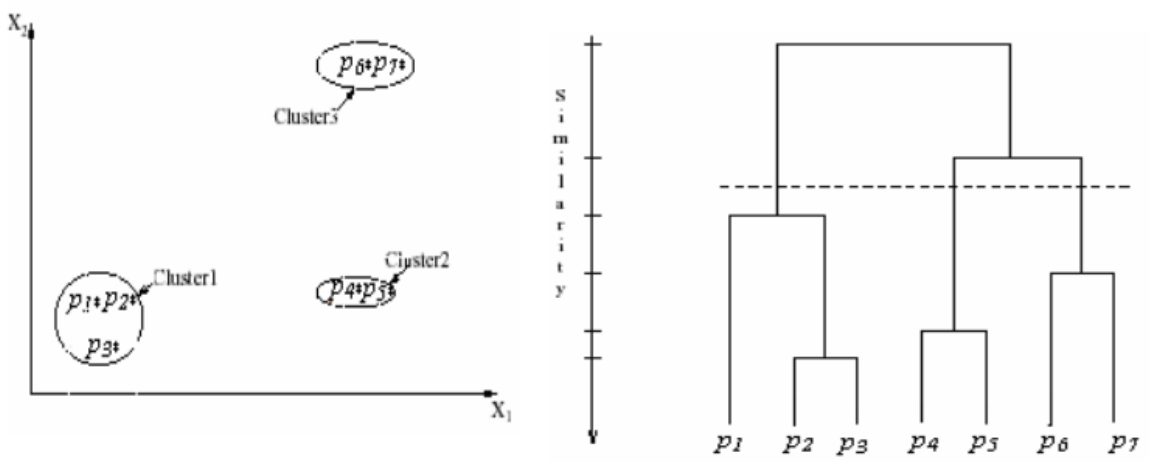


Figure 1.8. Clustering de sept points et le dendrogramme correspondant

Il y a deux approches de base pour générer un clustering hiérarchique :

A. algorithmes aggloméré :

Le clustering hiérarchique aggloméré est le type le plus courant de clustering hiérarchique utilisé pour regrouper des objets en clusters en fonction de leur similarité. C'est une approche ascendante où chaque observation commence dans son propre cluster, et les paires de clusters sont fusionnées au fur et à mesure que l'on monte dans la hiérarchie **[9].** L'algorithme de clustering hiérarchique aggloméré comprend les étapes suivantes

1- Faire de chaque point de données un cluster à un seul point
2- Prendre les deux points de données les plus proches et en faire un seul cluster
3- Répéter.
   3.1- Prendre les deux clusters les plus proches et en faire un cluster
4- Jusqu'à ce qu'il ne vous reste qu'un seul cluster.

B. Algorithmes divisifs :

Ces algorithmes commencent par la “racine” qui contient tous les objets, puis procèdent par divisions successives de chaque nœud jusqu’à obtenir des singletons (clusters d’un seulpoint). Brièvement l’algorithme est comme suit: **[10]**

1. Choisir le cluster contenant la paire d'objets la plus éloignée. C’est le cluster avec le plus grand diamètre.
2. Dans ce cluster, enlever l'objet qui a la plus grande distance moyenne des autres objets. cet objet forme un nouveau cluster de singleton.
3. Pour l'objet h dans le cluster étant dédoublé, calculez la distance moyenne entre ce dernier et le cluster courant; et la distance moyenne entre l'objet et le nouveau cluster. Si la distance au nouveau cluster est inférieure à celle au cluster courant, déplacez l'objet h à un nouveau cluster. Faites une boucle au-dessus de tous les objets dans le cluster.
4. Si aucun objet ne se déplace, et le nombre courant de clusters est plus grand que k, alors aller à l’étape 1. Sinon arrêter.

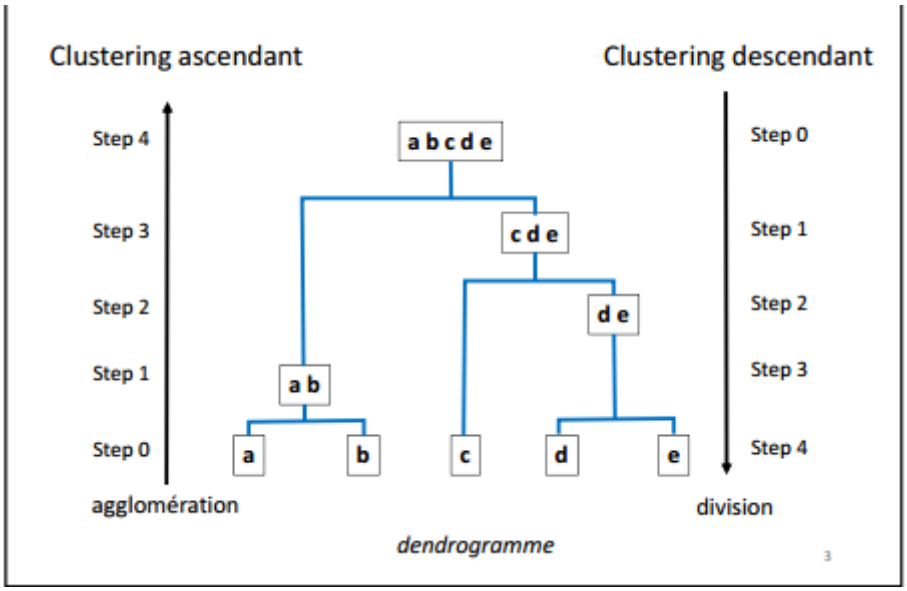


Figure 1.9. Types de clustering hiérarchique

Le dendrogramme :

Le dendrogramme est un diagramme en forme d'arbre qui montre la relation hiérarchique entre les observations. Il contient la mémoire des algorithmes de clustering hiérarchique.

Un dendrogramme peut être un graphique à colonnes (comme dans la figure ci-dessous) ou un graphique à lignes. Certains dendrogrammes sont circulaires ou ont une forme fluide, mais le logiciel produira généralement un graphique en lignes ou en importe la forme.

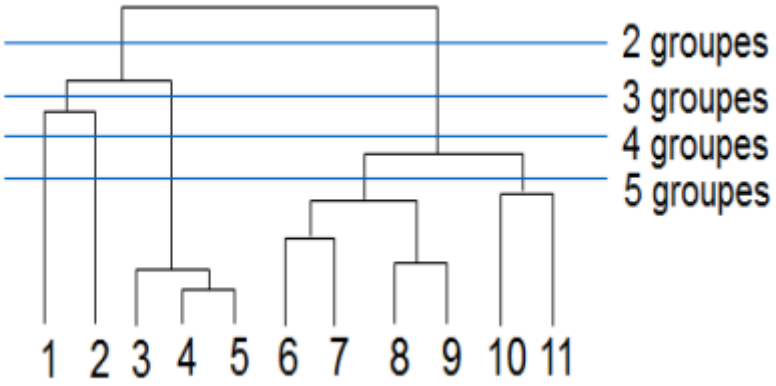


Figure 1.10. Simple exemple de dendrogramme

1.7. Métaheuristique pour le clustering :

Le problème de clustering peut être modélisé comme un problème d'optimisation où l'espace de recherche grandit exponentiellement et ne peut pas être parcouru exhaustivement même pour des problèmes de taille moyenne. En effet, le problème de clustering est connu pour être NP-difficile **[11].**

Résoudre un problème d'optimisation, c'est trouver l'optimum d'une fonction, parmi un nombre fini de choix, souvent très grand, cela se fait par les métaheuristique qui forment un ensemble de méthodes pour résoudre ces problèmes réputés difficiles. Parmi ces méthodes il existe la méthode de recherche coucou, les algorithmes génétiques, et les algorithmes évolutionnaires, les essaimes de particules et beaucoup d'autres qui ont contribué de façon remarquable pour résoudre le problème de clustering.

1.7.1. Clustering basé sur l'Algorithme génétique :

1.7.1.1. Principe de base :

La mesure de clustering qui a été adoptée est la somme des distances euclidiennes des points de leurs centres de clusters respectifs. Mathématiquement, la mesure M de clustering pour K clusters$C_1, C_2, \dots, C_K$ est donnée par

$$M(C_1, C_2, \dots, C_K) = \sum_{i=1}^{K} \sum_{x_j \epsilon c_i} \|x_j - z_i\|$$

La tâche de l'AG (Algorithme Génétique) est de rechercher les centres de clusters appropriés $z_1, z_2, \dots, z$ de sorte que la mesure de clusters M soit réduite au minimum.

1.7.1.2. Algorithme de clustering AG :

Les étapes de base d'AG, sont présentées Figure 1.11.

```
Begin
1.  t=0
2.  initialize population P(t)
3.  compute fitness P(t)
4.  t = t+1
5.  if termination criterion achieved go to step 10
6.  select P(t) from P(t-1)
7.  crossover P(t)
8.  mutate P(t)
9.  go to step 3
10.  Output best and stop
End
```

Figure 1.11. Étapes de base de l'AG

A. représentation de la chaîne de caractères :

Chaque chaîne de caractères est une séquence de nombres réels représentant les centres du K-clusters. Pour un espace N-dimensionnel, la longueur d'un chromosome est N_K^* mots, où les premières positions de N (ou, gènes) représentent les N-dimensions de du premier centre du cluster, les N-positions suivants représentent ceux du deuxième centre du cluster, et ainsi de suite.

B. Initialisation de la population :

Les centres du K clusters encodés dans chaque chromosome sont initialisés à K points choisis aléatoirement de l'ensemble de données. Ce processus est répété pour chacun des P chromosomes de la population, où P est la taille de la population.

C. Calcul du fitness :

Le processus de calcul du fitness comprend deux phases. Dans la première phase, les clusters sont formés selon les centres codés dans le chromosome considéré. Ceci est fait en assignant chaque point x_i, i=1, 2, n, à l'un des clusters C_j avec le centre $Z_j$tel que

$$\|X_i - Z_j\| < \|X_i - Z_p\| \text{ ,p=1,2 ,…..,K, et p} \neq j$$

Après la mise de clusters, les centres des clusters encodés dans le chromosome sont remplacés par les points moyens des clusters respectifs. En d'autres termes, pour le clusterC_i, le nouveau centre Z_i^* est calculé comme suit :

$$Z_i^*=\frac{1}{n_i}\sum_{x_j \epsilon c_i} x_i, \text{i =1.2,….,k.}$$

Ces Z_i^* remplacent maintenant les zi précédents dans le chromosome

D. Sélection

Un chromosome est attribué un certain nombre de copies, qui est proportionnelle à son aptitude dans la population, qui vont dans la piscine d'accouplement pour d'autres opérations génétiques. Roulette sélection est une technique commune qui met en œuvre la stratégie de sélection proportionnelle.

E. Croissement

Le croisement est un processus probabiliste qui échange de l'information entre deux chromosomes parents pour générer deux chromosomes enfants. un croisement à point unique avec une probabilité de croisement fixe de $\mu_c$est utilisé. Pour les chromosomes de longueur l, un nombre entier aléatoire, appelé point de croisement, est généré dans la plage [1, l-1]. Les parties des chromosomes situées à droite du point de croisement sont échangées pour produire deux descendants

F. Mutation

Chaque chromosome subit une mutation avec une probabilité fixe μ_m . Pour la représentation binaire des chromosomes, une position de bit (ou gène) est mutée en changeant simplement sa valeu. Puisque nous considérons la représentation en virgule flottante, nous utilisons la mutation suivante. Un nombre Δ dans l'intervalle [0, 1] est généré avec une distribution uniforme. If the value at a gene position is , after mutation it becomes

$$v \pm 2 * \delta * v\,, v \neq 0$$

$$v \pm 2 * \delta\,, v \neq 0$$

Le signe + ou - se produit avec une probabilité égale. Notez que nous aurions pu mettre en œuvre la mutation comme

$$v \pm \delta * v$$

Cependant, un problème avec cette forme est que si les valeurs à une position particulière dans tous les chromosomes d'une population deviennent positives (ou négatives), alors nous

ne serons jamais en mesure de générer un nouveau chromosome ayant une valeur négative (ou positive) à cette position. Afin de surmonter cette limitation, nous avons incorporé un facteur de 2 tout en mettant en œuvre la mutation. D'autres formes comme

$$v \pm (\delta + \varepsilon) * v$$

Où 0< ε <1 aurait également satisfait notre objectif. On peut noter dans ce contexte que des types similaires d'opérateurs de mutation pour l'encodage réel ont été utilisés principalement dans le domaine des stratégies évolutives **[12]**

Les processus de calcul de fitness, de sélection, de croisement et de mutation sont exécutés pour un nombre maximum d'itérations. La meilleure chaîne de caractères vue jusqu'à la dernière génération fournit la solution au problème de clustering. Nous avons mis en œuvre l'élitisme à chaque génération en préservant la meilleure chaîne de caractères vue jusqu'à cette génération dans un endroit en dehors de la population. Ainsi à la fin, cet emplacement contient les centres des clusters finals.

1.7.2. Clustering par fourmis artificielles :

Les mécanismes de coopération et d'apprentissage chez les fourmis réelles ont inspiré les chercheurs dans de nombreux domaines.

Cela se justifie particulièrement quand on connaît la richesse comportementale de ces insectes.

Plusieurs comportements observés chez les fourmis ont été directement mis en relation avec le problème du clustering comme le tri collectif, la constitution de cimetières, le transport coopératif et la recherche du chemin le plus court. Deneubourg apparaît comme un pionnier dans le domaine du tri d'objets par des fourmis artificielles. Dans les principes suivants ont été proposés :

Des fourmis artificielles se déplacent sur un plan. Les objets à rassembler sont répartis sur ce plan. Une fourmi ne dispose que d'une perception locale de ces objets et ne communique pas avec les autres. Au lieu de cela, la configuration des objets sur le sol va influencer leurs actions. Lorsqu'une fourmi rencontre un objet, la probabilité qu'elle en ramasse est d'autant plus grande que cet élément est isolé. Lorsqu'une fourmi transporte un objet, elle le dépose avec une probabilité d'autant plus grande que la densité d'objets du même type dans le voisinage est grande. Ces règles relativement simples font qu'il apparaît des regroupements

D'objets, et elles permettent ainsi de trier ces objets. Cet algorithme a été à l'origine présenté pour des tâches en robotique. **[1] [13]**

1.7.3. Clustering par essaim de particules :

L'optimisation par essaim de particules a été introduite en 1995 par Kennedy et Eberhart. C'est une approche d'optimisation stochastique, inspirée du comportement social des animaux comme les abeilles. En effet, tout comment ces animaux se déplacent en groupe pour trouver de la nourriture ou éviter les prédateurs, les algorithmes à essaims de particules recherchent des solutions pour un problème d'optimisation.

Dans une approche de PSO, les individus sont appelés particules et la population est appelée essaim. Le mouvement de chaque particule dans un espace de recherche multidimensionnel est dirigé par une vectrice position et une vectrice vitesse. En commençant avec une population initialisée aléatoirement ou prédéfini, chaque particule déplacer à travers l'espace de recherche et se souvient de la meilleure solution qu'elle a trouvée. Les particules de l'essaim communiquent entre elles les bonnes positions et ajustent dynamiquement leur propre position et vitesse basée sur les bonnes positions. Le réglage de la vitesse est basé sur les expériences et les comportements historiques des particules elles-mêmes ainsi que leurs voisines. La performance de chaque particule est mesurée avec une fonction objective**. [14]**

Au temps t + 1, la vectrice vitesse est calculé par :

$$vij(t+1)= Wvij(t)+ c1r1(pij(t)- xij(t))+ c2r2(pgj(t)- xij(t))$$

$c1\ et\ c2$, sont deux constantes appelées coefficients d'accélération, $r1\ et\ r2$ sont deux nombres aléatoires tirés uniformément dans l'intervalle [0, 1], et W le coefficient d'inertie.

La position de la particule i est définit par : $(t+1)= (t) + (t+1)$

L'algorithme PSO est décrit dans la figure ci-dessous :

Initialiser aléatoirement N particule : position et vitesse
Evaluer les positions des particules
Tant que le critère d'arrêt n'est pas atteint faire
Pour i allant de 1 à N faire
Déplacer les particules
Si f (xi) < f (pi) alors
pi = xi
Si f (xi) < f (pg) alors
pg = xi
Fin si
Fin si
Fin pour
Fin tant que

Figure 1.12. Structure de l'algorithme PSO

1.8 .Techniques de validation de Clustering:

L'objectif principal de la validation de clusters est d'évaluer le résultat de clustering afin de trouver le meilleur partitionnement du jeu de données. Il existe des approches de validité de cluster pour évaluer quantitativement le résultat d'un algorithme de clustering **[15**] mais dans la plupart des évaluations expérimentales des algorithmes, des jeux de données 2D sont employés pour que le lecteur soit capable de vérifier visuellement la validité des résultats (c.-à-d., à quel point l'algorithme de clustering a découvert les clusters du jeu de données). Il est clair que la visualisation du jeu de données est une vérification cruciale des résultats de clustering. Dans le cas de grands jeux de données multidimensionnels (par exemple plus de trois dimensions) la visualisation efficace du jeu de données serait difficile. De plus la perception de clusters à l'aide des outils de visualisation disponibles est une tâche difficile pour les gens qui ne sont pas habitués aux espaces d'un grand nombre de dimensions **[16].**

Il est commun de distinguer entre l'évaluation intrinsèque et extrinsèque. Les mesures de qualité externes emploient la connaissance externe. Beaucoup de ces derniers comparent le clustering à une autre partition. Les mesures de qualité internes n'emploient aucune connaissance externe, mais elles sont basées sur ce qui est disponible pour l'algorithme de clustering **[17].**

Il existe deux critères qui ont été largement considérés suffisants pour mesurer la qualité du partitionnement de données **[18].**

La compacité, les membres de chaque cluster doivent être proche l'un de l'autre autant que possible. Une mesure commune de compacité est la variance, qu'on doit minimiser. La séparation, les clusters eux-mêmes doivent être largement espacée. La distance euclidienne entre les centroïdes de clusters donne une indication de la séparation de clusters.

1.8.1 Mesures externes:

Le premier groupe de fonctions d'évaluation analytiques disponibles pour l'analyse de clusters est le groupe de fonctions conçues pour des problèmes de références pour lesquels le bon nombre de cluster et la classification correcte pour chaque point de données sont connus. L'évaluation devient beaucoup plus juste et sérieuse dans ces cas, puisque les propriétés désirées du partitionnement (qui conforme avec un certain degré à la définition de problème) peuvent être négligées, et nous pouvons seulement se concentrer sur la validité des assignements aux clusters obtenus **[17].**

Les mesures externes que nous allons présenter, appliquent directement la connaissance des étiquettes de classes. Elles évaluent les clusters générés en prenant en compte les classes d'appartenances correctes.

- **La F-mesure :**

La F-mesure est une fonction utilisée souvent dans la littérature pour évaluer les algorithmes de clustering. La F-mesure adopte les idées de la précision et du rappel de la recherche documentaire .Elle compare la qualité de clustering en tenant compte des classes correctes connues pour un jeu de données. Soit $C = (C_1, C_2, \ldots, C_k)$ un clustering donné et $R = (R_1, R_2, \ldots, R_{k'})$ les classes correctes.

Chaque classe R_i contient N_i points de données, chaque cluster C_j (généré par l'algorithme) est considéré comme l'ensemble de N_j points de données. N_{ij} Donne le nombre de points de la classe R_i dans le cluster C_j et N donne le nombre total des points du jeu de données. Pour chaque classe R_i et un clusterC_j, la précision et le rappel sont alors défini comme :

$$\text{Prec}(R_i, C_j)=\frac{N_{ij}}{N_j} \text{ et } \text{Rep}(R_i, C_j)=\frac{N_{ij}}{N_i}$$

Et la valeur de F-mesure correspondante est :

$$\text{Fmes}\ (R_i, C_j)=\frac{(b^2+1)\ .\text{Prec}(R_i,C_j).\text{Rep}\ (R_i,C_j)}{b^2\ .\text{Prec}(R_i,C_j)+\ \text{Rep}\ (R_i,C_j)}$$

Où des coefficients égaux de Prec(R_i, C_j) et Rep(R_i, C_j) sont obtenu si b=1. La valeur globale de F-mesure F pour toute la partition est calculée comme

$$\text{F(C)}=\sum_{i=1}^{k\prime}\frac{N_i}{N}\quad\max_{ci\epsilon C}(\text{Fmes}(R_i, C_j))$$

Elle est limitée à l'intervalle [0,1] et devrait être maximale.

- **La pureté:**

La pureté de cluster $C_j\ \epsilon C$ est définie comme le pourcentage du type de données prédominant selon la classe réelle connue $R_i\ \epsilon\ R$, qui est :

$$\text{Pur}(C_j)=\ \max_{R_i\epsilon \text{R}}\frac{N_{ij}}{N_i}$$

où N_j est la taille du cluster C_i et N_{ij}est le nombre des points de données de la classe $R_i$dans ce cluster. La pureté P(C) d'une partition entière est alors calculée comme la pureté moyenne de tous les clusters. Elle est limité à l'intervalle] 0,1] et devrait être maximale [**17].**

- **L'entropie:**

En outre, le degré relatif d'aspect aléatoire du partitionnement peut être évalué en utilisant la notion d'entropie de cluster. C'est une mesure plus complète que la pureté, car elle tient compte de la distribution de toutes les classes dans chaque cluster. L'entropie d'un cluster est :

$$\textit{Entr}(C_j)=-\frac{1}{\log(N)}\sum_{R_i\epsilon R}\frac{N_{ij}}{N_j}\log(\frac{N_{ij}}{N_j})$$

et, encore, l'entropie globale E(C) est calculée en faisant la moyenne des entropies de clusters. L'entropie de cluster est limitée à l'intervalle [0,1] devrait être minimale **[17].**

- **L'indice Rand:**

D'autres méthodes qui sont généralement appliquées pour comparer le résultat du clustering obtenu à la structure de cluster connue sont des indices basés sur des statistiques relative aux assignements de clusters, qui peut généralement être appliquée également pour évaluer la conformité entre deux partitions du même jeu de données. Étant donné les partitions U et V, les quantités a, b, c et d sont calculées pour toutes les paires possibles de points de données p_i et p_j et leurs assignements de cluster respectives $C_\cup(p_i)$, $C_\cup(p_j)$, $C_V(p_i)$, et $C_V(p_j)$,où

$$\text{a}=\{|p_i, p_j|C_U(p_i) = C_i(p_j)\ \cap\ C_V(p_i) = C_V(p_j)|\}$$

$$\text{b}=\{|p_i, p_j|C_U(p_i) = C_i(p_j)\ \cap\ C_V(p_i) \neq C_V(p_j)|\}$$

$$c=\{|p_i,p_j|C_U(p_i)\neq C_i(p_j)\ \cap\ C_V(p_i)=C_V(p_j)|\}$$
$$d=\{|p_i,p_j|C_U(p_i)\neq C_i(p_j)\ \cap\ C_V(p_i)\neq C_V(p_j)|\}$$

Par conséquent, *a* et *d* maintiennent les correspondances entre les deux partitions, tandis que *b* et *c* calculent des déviations claires. L'indice le plus connu est l'indice Rand **[19],** qui est défini comme:

$$R(U,V)=\frac{a+d}{a+b+c+d}$$

Évidemment, *R* est limité à l'intervalle [0,1]. Une valeur de 1 est seulement obtenue pour une correspondance parfaite entre les assignements de clusters des partitions *U* et *V*, et les plus petites valeurs montrent un grand écart entre les deux solutions. Quelques variations de cet indice existent, comme l'indice Rand ajusté **[20]**, qui présente une normalisation afin de rapporter des valeurs près de 0 pour des partitions aléatoires.

1.8.2 Mesures internes:

Si on aborde des jeux de données dont la structure réelle est inconnue, l'évaluation des résultats de clustering devient beaucoup plus compliquée.

Les mesures qui peuvent être appliquées dans ces cas essayent de capturer les deux objectifs d'analyse de cluster : la minimisation de la distance intra cluster (qui résulte aux clusters compacts) et la maximisation de la distance inter-cluster (qui résulte aux clusters bien séparés).

- **La variance intra cluster:**

La variance intra cluster optimisée implicitement par l'algorithme Kmeans, est basée sur le concept de la minimisation de la distance intra clusters. Elle est donnée par :

$$\text{Var(C)}=\sum_{ci\epsilon c}\sum_{p_j\epsilon\, c_i} d(p_{j+\mu_i})^2$$

Où $p_j$dénote un point de données, *k* dénote le nombre de clusters, μ_i représente le centroïde de cluster C_i, *d(.,.)* est la fonction de distance utilisée pour calculer la déviation entre le point de données $p_j$et le centroïde μ_i La limite inférieure de la variance qu'on peut obtenir dépend des données et le nombre de clusters employé, mais elle peut au meilleur être égal au zéro **[21].**

- **L'indice SD :**

L'indice SD **[22]** est une combinaison linéaire entre deux mesures de distance intra clusters et de distance inter clusters, il est donné par :

$$\text{SD(C)}=\alpha\ Scat(C)+Dis(C)$$

où a est un terme de pondération qui fait un compromis entre l'importance relative de la dispersion moyenne des clusters *Scat(C)* et la séparation totale entre les clusters *Dis (C)*, où

$$\text{Scat(C)} = \frac{1}{k}\sum_{i=1}^{k} N_k \frac{|\sigma(\mu_i)|}{|\sigma(C)|}\text{ , Dis(C)} = \frac{D_{max}}{D_{min}}\sum_{i=1}^{k}\left(\sum_{j=1}^{k} d(\mu_i,\mu_j)\right)^{-1}$$

La 1^{emme} dimension de $\sigma(\mu_i)$ est $\sigma_{\mu_i}^{l}$ et la 1^{emme} dimension des $\sigma(C)$ est σ_C^l ils sont donnés par les formules suivantes :

$$\sigma_{\mu_i}^{l} = \sum_{j=1}^{Ni}(p_j^l - \mu_j^{\,l})^2 \;\; / \;\; N_i \text{ , } \sigma_C^l = \sum_{j=1}^{N}(p_j^l - x^l)^2 \;\; / \;\; N$$

où x^l est la 1^{emme} dimension de X= $\frac{1}{N}\sum_{i=1}^{N} p_j$

$\frac{D_{max}}{D_{min}}$ Dénote la proportion entre le maximum et le minimum de la distance entre les centroïdes de clusters. k est le nombre de clusters , $N_i$est le nombre de points de données assignées au cluster C_i, N est le nombre totale de points de données μ_i est le centroïde de cluster C_i.

- **La famille des indices de *Dunn :***

L'indice de validité de clusters proposé par *Dunn* **[23]** essaye d'identifier des clusters compacts et bien séparés. L'indice est défini pour un nombre spécifique de clusters. Il est donné par :

$$D_k = \min_{i=1,..,k}\left\{\min_{j=i+1,..k}\left(\frac{d(C_i,C_j)}{\max_{l=1,..k} diam(C_l)}\right)\right\}$$

Où $d(C_i, C_j)$ est la fonction de dissimilarité entre deux clusters C_i et C_j défini comme :

$$\text{d}(C_i, C_j) = \min_{p\epsilon C_i,\, q\epsilon C_j)} d(p,q)$$

Et *diam(C)* est le diamètre d'un cluster, qui peut être considéré comme une mesure de dispersion des clusters. Le diamètre d'un cluster *C* peut être défini comme suit :

$$diam\ (C) = \max_{p,q\epsilon C} d(p,q)$$

Si le jeu de données contient des clusters compacts et bien séparés, la distance entre les clusters sera grande et le diamètre des clusters sera petit. Ainsi, basé sur la définition d'indice de *Dunn*, nous pouvons conclure que les grandes valeurs de l'indice indiquent la présence de clusters compacts et bien séparés.

Les implications de l'indice de *Dunn* sont le temps considérable requis pour son calcul, et la sensibilité à la présence de bruit dans les jeux de données, puisque ceux-ci vont probable augmenter les valeurs de *diam(C).*

L'indice comme *Dunn* (index like *Dunn*) proposé dans [Pal et al, 97] est plus robuste à la présence de bruit. Il est basé sur le concept du MST (minimum spanning tree), il est donné par l'équation :

$$D_k = \min_{i=1,..,k} \left\{ \min_{j=i+1,..k} \left(\frac{d(C_i, C_j)}{\max_{l=1,..k} diam_l^{MST}} \right) \right\}$$

Il existe beaucoup d'indices basés sur l'indice de *Dunn*.

- **L'indice de Davies-Bouldin :**

L'indice *DB* **[24]** est une fonction basée sur la minimisation du rapport des dispersions intra-clusters S_i et de la séparation inter-clusters introduite L_i par Davis et Bouldin. L'indice *DB* est donné par :

$$\text{DB}=\frac{1}{k}.\sum_{i=1}^{k} L_i \ , avec$$

$$L_i= \max_{\substack{j=1,...k \\ i\neq j}} L_{i,j} \text{ et } \quad L_{i,j} = \frac{\left(S(C_i)+(C_j)\right)}{d(C_i,C_j)}$$

Une mesure typique de dispersion d'un cluster $C_i$est $S(C_i) = \frac{1}{|C_i|}\sum_{p\epsilon ci}\|p - c_i\|$ et la mesure typique de distance entre clusters est la distance entre les centroïdes , *d(μi, μj)*. Donc, la proportion est petite si les clusters sont compacts et éloignés les uns des autres. Par conséquence, l'indice de Davies-Bouldin aura une petite valeur quand le clustering sera de bonne qualité.

- **La connectivité :**

La mesure de connectivité de cluster évalue le degré auquel des points de données voisins ont été placés dans le même cluster. Elle est calculée par la formule suivante :

Conn $=\sum_{i=1}^{m}\left(\sum_{1=1}^{L} P_{i,nn_i}(1)\right)$, ou $P_{r,s} = \begin{cases} \frac{1}{1} \; if \; \neg\exists c_j : r, s \; \epsilon c_j \\ 0 \; otherwise \end{cases}$

nn_i (l) est 1^{eme} le plus proche voisin du point de données pi et L est le paramètre déterminant le nombre des voisins qui contribuent à la connectivité. La connectivité devrait être minimale [**25]**

1.9. Définition de clustering de client :

Le clustering de client, également appelé segmentation de clientèle, est une méthode d'analyse de données qui consiste a diviser un ensemble de client en groupes homogènes en fonction de caractéristiques similaires. Cette approche vise a regrouper les clients qui partagent des comportements d'achat, des préférences, des besoins ou des caractéristique démographiques similaires.

Le clustering de client est largement utilisé dans le domaine du marketing et de la gestion de la relation client. Il permet aux entreprises de mieux comprendre leur base de client et de prendre des décisions plus éclairées en matière de marketing, de communication, de développement de produits et de stratégie de vente. Les méthodes de clustering de client utilisent généralement des variables telles que l'âge, le genre, le revenu les habitudes d'achat, les préférences, les comportements en ligne, etc., pour regrouper les clients. Les algorithmes de clustering, tels que le clustering k-means, le clustering hiérarchique ou le clustering bas' sur la densité, sont souvent utilisés pour effectuer cette segmentation.

Une fois les clients segmentés en clusters, l'entreprise peuvent personnaliser leurs offres, leurs campagnes de marketing et leurs stratégies de communication en fonction des caractéristiques spécifiques de chaque segment. Cela permet de fournir des expériences plus pertinentes et personnalisées aux clients, d'améliorer a satisfaction, la fidélité et les performances commerciales globales.

1.10. Avantage du clustering de client :

Le clustering de client présente plusieurs avantages pour les entreprises :

1. Compréhension approfondie de la clientèle : Le clustering de client permet de segmenter la base de clients en groupes homogènes, ce qui aide les entreprises à mieux comprendre les différentes caractéristiques, besoins, comportements et préférences de chaque segment. Cela permet de personnaliser les offres, les messages et les stratégies pour répondre de manière plus ciblée aux attentes de chaque groupe.
2. Personnalisation des stratégies marketing : En segmentant les clients en clusters, les entreprises peuvent adapter leurs stratégies marketing à chaque segment. Cela permet de créer

des campagnes de marketing plus ciblées, des offres spécifiques et des communications plus pertinentes. La personnalisation accrue des stratégies marketing peut conduire à une meilleure rétention des clients, à une augmentation des ventes et à une fidélisation accrue.

3. Détection de nouveaux marchés et opportunités : Le clustering de client peut révéler des segments de clients sous-exploités ou émergents qui n'ont pas encore été ciblés efficacement. Cela permet d'identifier de nouveaux marchés potentiels, de développer de nouveaux produits ou services adaptés à ces segments spécifiques, et d'explorer de nouvelles opportunités de croissance.
4. Réduction des coûts marketing : En comprenant mieux les différentes caractéristiques et comportements des segments de clientèle, les entreprises peuvent optimiser leurs investissements marketing en ciblant les ressources sur les segments les plus rentables et en évitant de dépenser des ressources sur des segments moins réceptifs. Cela permet d'optimiser le retour sur investissement marketing et de réduire les coûts inutiles.
5. Amélioration de la satisfaction client : En personnalisant les offres et les communications en fonction des besoins spécifiques de chaque segment de clientèle, les entreprises peuvent améliorer la satisfaction client en offrant des expériences plus pertinentes et adaptées. Cela peut conduire à une fidélisation accrue, à une recommandation positive et à des relations client plus durables.

En résumé, le clustering de client permet aux entreprises de mieux comprendre leur base de clients, d'adapter leurs stratégies marketing, de détecter de nouvelles opportunités et de créer des expériences client plus satisfaisantes. Cela peut conduire à une amélioration de la performance commerciale globale et à une relation client plus solide.

1.11. L’objectif de l’analyse de clusters dans le marketing :

L'analyse de clusters dans le marketing vise à regrouper les clients ou les prospects en segments homogènes afin de mieux comprendre leur comportement, leurs préférences et leurs besoins. L'objectif principal est de diviser une population de clients en groupes distincts, appelés clusters, qui partagent des caractéristiques similaires. Ces caractéristiques peuvent être basées sur des données démographiques, des comportements d'achat, des habitudes de consommation, des préférences de produit, des valeurs ou tout autre critère pertinent pour l'entreprise.

Les avantages de l'analyse de clusters dans le marketing sont les suivants :

1. Segmentation du marché : L'analyse de clusters permet de segmenter le marché en groupes homogènes, ce qui aide les entreprises à mieux cibler leurs efforts de marketing. Chaque cluster peut être traité comme une cible distincte avec des stratégies de marketing spécifiques adaptées à ses caractéristiques uniques.
2. Personnalisation des offres : En comprenant les besoins et les préférences spécifiques de chaque cluster, les entreprises peuvent personnaliser leurs offres de produits, de services et de communication. Cela permet d'améliorer la pertinence des messages marketing et d'accroître l'engagement des clients.
3. Optimisation des ressources : L'analyse de clusters permet de mieux allouer les ressources marketing en identifiant les segments les plus précieux et les plus rentables. Les entreprises peuvent concentrer leurs efforts et leurs investissements sur les clusters les plus prometteurs, ce qui peut améliorer l'efficacité des activités marketing.
4. Prévision des comportements futurs : En étudiant les clusters et en analysant les modèles de comportement passés, les entreprises peuvent tirer des conclusions sur les tendances et les comportements futurs. Cela peut les aider à anticiper les besoins des clients, à développer de nouvelles offres et à prendre des décisions stratégiques éclairées.

En résumé, l'analyse de clusters dans le marketing est un outil puissant pour segmenter le marché, personnaliser les stratégies de marketing, optimiser les ressources et prévoir les comportements futurs. Elle permet aux entreprises de mieux comprendre leurs clients et de développer des initiatives marketing plus efficaces.

1.12. Conclusion :

Le clustering des données est une tâche dont l'objectif est de trouver des groupes au sein d'un ensemble de données. Dans ce chapitre, nous avons vu les grands concepts du clustering, les principales techniques existantes dans le clustering. Nous avons également présenté la différente métaheuristique de clustering. Nous avons passé ensuite, aux mesures d'évaluation.

Chapitre 2
La Recherche de Coucou

2.1. Introduction :

Les algorithmes de métaheuristique sont des méthodes d'optimisation qui permettent de résoudre des problèmes complexes et difficiles a traiter par des approches classiques.

Ils sont souvent utilisés lorsque les méthodes exactes ne sont pas réalisables en raison de la taille ou de la complexité du problème

Les métaheuristique tirent leur nom du concept de méta, qui signifie « au-dessus » ou « au-delà », ce qui indique qu'elles sont capables de guider la recherche au-delà des simples mécanismes locaux d'exploration.

Ces algorithmes sont souvent inspirés de phénomènes naturels ou de comportements sociaux, et ils utilisent des stratégies d'exploration et d'exploration pour trouver des solutions de haute qualité.

L'une des caractéristique clés des métaheuristiques est leur capacité à explorer l'espace des solutions de manière non déterministe. Cela signifie qu'au lieu de suivre un cheminement prédictible et déterminé, ces algorithmes explorent l'espace des solutions de manière probabiliste, en utilisant des mécanismes aléatoires pour éviter de rester bloqués dans des optima locaux.

La recherche de coucou en anglais Cuckoo Search (CS) est l'un des derniers algorithmes métaheuristique inspirés de la nature, développés en 2009 par Xin-She Yang et Suash Deb.CS est basé sur le parasitisme de couvaison de certaines espèces de coucou. De plus, cet algorithme est renforcé par les soi-disant vols de Lévy, plutôt que par de simples randonnées aléatoires isotropes.

Dans ce chapitre, nous allons introduire la recherche de coucou en détail, en commençant par son comportement de reproduction intéressante, nous allons voir également les caractéristiques des vols de Lévy et son algorithme de base ainsi que son champ d'applications.

2.2. Le comportement et la reproduction des Coucous

Figure 2.1. Image illustrant le Coucou gris

Toutes les espèces d'oiseaux ont la même approche de prolifération. Aucun oiseau ne donne naissance à des petits vivants. Après l'accouplement, l'oiseau pond un œuf recouvert d'une coquille protectrice qui est ensuite incubée en dehors du corps. Vu la richesse en protéine des œufs d'oiseau, ils construisent un volumineux repas pour toutes sortes de prédateurs. Afin de protéger leurs œufs des menaces des différents prédateurs, les oiseaux doivent trouver un endroit sûr pour donner plus de chance d'éclosion à leurs poussins. Trouver un endroit pour placer en toute sécurité, couver leurs œufs et élever leurs poussins au point de l'indépendance, est un défi pour les oiseaux. Ces derniers l'ont résolu de nombreuses manières astucieuses en utilisant des conceptions complexes et artistiques. De nombreux oiseaux construisent des nids isolés, cachés dans la végétation pour échapper au risque d'être détecté par les prédateurs. Certains d'entre eux sont des spécialistes dans le choix du bon emplacement de leurs nids. Ils cachent bien leurs nids de façon à ce que même les yeux de l'homme qui voit tout n'aient pratiquement jamais pu les détecter **[26].** Il ya un autre type d'oiseaux qui se dispensent des responsabilités imposées par leur lien de parentalité. Ce sont des oiseaux parasites. Ils ne construisent pas leurs propres nids. En effet, ils pondent leurs propres œufs dans les nids des autres espèces en confiant la responsabilité d'incubation, de nourriture et d'élevage de leurs œufs aux parents adoptifs.

Les coucous sont un bon exemple d'oiseaux parasites de couvée. Ce sont des oiseaux fascinants, ils ont un bon chant. Ils appartiennent à la classe des cuculidés. Cette dernière compte environ 140 espèces réparties partout dans le monde. Plusieurs espèces de coucous sont des parasites. Ils ne construisent pas leur propre nid pour pondre leurs œufs. Quelques espèces

entre eux s'engagent à pondre leurs œufs dans des nids communautaires où plus d'une femelle pond dans le même nid, comme l'Ani et le Guira **[27].** En fait, ils pondent leurs œufs dans des grands nids collectifs des membres de leur propre espèce puis ils se chargent de l'élevage collectif de leurs petits poussins. D'autres espèces pondent leurs œufs dans des nids d'autres espèces **[28]** puis ils se sauvent, comme le coucou gris, le coucou koël, Le Coucou plaintif et le coucou geai. Les coucous de ce dernier type ne couvent pas et ne nourrissent jamais leurs poussins. Quelques explications tentent de traduire ce comportement par le fait que le coucou parent se nourrie d'insectes toxiques. Cette nourriture menace la vie de ses poussins ce qui pousse le coucou parent à confier la responsabilité de nourriture et d'élevage de ses petits à d'autres espèces capables de leur fournir la bonne et la saine nourriture.

De nombreuses espèces d'oiseaux apprennent à reconnaître les œufs coucous déversés dans leurs propres nids. Dans le cas ou l'oiseau hôte arrive à découvrir l'œuf coucou dans son nid, il va soit jeter l'œuf étranger hors son nid ou abandonner son propre nid. Tandis que les hôtes tentent de trouver des moyens de détection de l'œuf parasite, le coucou cherche constamment à améliorer le mimétisme du modèle des œufs de ses hôtes **[26].** Le coucou est un expert dans l'art de tromperie. Sa stratégie se base sur la furtivité, la surprise et la vitesse **[26].** Après l'accouplement, le coucou pond ses œufs dans les nids des autres espèces. L'œuf coucou risque de ne pas survivre au cas où l'oiseau hôte découvre qu'il n'est pas le sien. Dans le but d'augmenter la ressemblance entre l'œuf coucou et les œufs d'accueil, certaines femelles coucou sont très spécialisées dans le mimétisme de couleurs et de modèles des œufs de l'hôte. Chaque femelle coucou se spécialise sur une espèce hôte particulière. En outre, Les mamans coucous peuvent détruire les œufs de l'oiseau hôte afin d'hausser la possibilité d'éclosion et de nourriture de leurs œufs.

En général, les œufs du coucou se développent plus rapidement et éclosent plus tôt que ceux de l'oiseau d'accueil. Dès son éclosion, le poussin coucou dupe ses parents adoptifs. Il éjecte les œufs d'accueil par aveuglement ce qui lui offre la monopolisation des soins de ses parents adoptifs et de la nourriture destinée à toute une couvée. Il incite ses parents adoptifs à suivre le rythme de son taux de croissance élevé par son appel de mendicité rapide **[29]** qui imite l'appel des poussins d'accueil et aussi sa bouche ouverte

qui construit un fort stimulus **[30].** En effet, il grandit plus vite au détriment des poussins d'accueil.

2.2.1. L'habitat des coucous

L'habitat est le milieu de vie, de reproduction et de développement d'une population d'une espèce donnée. De même pour les coucous, il constitue une source de nourriture et un lieu de reproduction. Les Coucous se produisent dans une grande variété d'habitats. La majorité des espèces survivent dans les forêts et les bois, principalement dans les forêts tropicales à feuilles persistantes. En plus des forêts, certaines espèces de coucous occupent des environnements plus ouverts, ce qui peut inclure même les zones arides comme les déserts **[31].** A titre d'exemple, le Coucou plaintif fréquente une assez grande variété d'habitats tels que les bois ouverts, les forêts secondaires, arbustes et broussailles, les champs cultivés et aussi les jardins, aussi bien en milieu rural que citadin. On peut aussi le voir dans les herbages et les marais. Le Coucou gris fréquente les forêts de conifères ou de feuillus, et les zones boisées en général, les espaces boisés ouverts, les lisières de forêts et les clairières, les steppes arborées, les prairies, les marais et les roselières, ainsi que les zones cultivées avec des arbres et des buissons. Le Coucou geai fréquente des habitats semi-arides tels que les zones boisées ouvertes (surtout avec des acacias et des arbustes épineux), les contreforts rocheux des collines dans les savanes sèches, et les zones agricoles sèches avec des arbres et des buissons. On le voit souvent voler au-dessus des espaces découverts. Selon la distribution, et particulièrement en Europe, on peut le trouver dans les landes de bruyères avec du chêne-liège et des conifères du genre Pinus pinea. Il fréquente également les bosquets et les plantations d'oliviers.

2.2.2. La migration des coucous

La migration est un phénomène très répandu chez certains animaux y compris les oiseaux. Ce phénomène n'est pas obligatoire. En fait, il est lié à des paramètres alimentaires, physiologiques et hormonaux. En fonction de ces paramètres, l'oiseau est incité à se déplacer ou non. La date de migration n'est pas commune. Certaines espèces partent avant de manquer de nourriture, d'autres attendent quelques changements climatiques. En effet, l'oiseau risque de perdre sa vie en cas de baisse de disponibilité de nourriture ou des conditions climatiques rigoureuses. La plupart des espèces de coucous sont sédentaires, mais plusieurs espèces de coucou entreprennent des migrations saisonnières, et plusieurs autres entreprennent des migrations partielles sur une partie de leur distribution. La migration peut être diurne, comme celle du coucou 'Channel-billed' comme elle peut être nocturne, comme celle du coucou à bec noir et

du coucou à bec jaune qui se reproduisent en Amérique du nord et migrent de nuit à travers le Mexique et l'Amérique centrale pour hiverner en Amérique du Sud. En général, les coucous parasites migrent sur de longues distances et ont un vol rapide et direct. Leurs déplacements coïncident non seulement avec la saison de reproduction des espèces hôtes, mais aussi avec l'émergence des chenilles qui constituent leur nourriture favorite **[27].** Une preuve tangible a été fournie par un coucou bagué en 1939, à sa naissance, et repris en 1952, en Allemagne. Cet oiseau avait effectué 26 trajets entre l'Europe et l'Afrique, soit au minimum 150 000 km. L'équivalent de presque quatre fois le tour du globe.

2.3. La recherche Coucou (CS)

En 2009, Xin-She Yang et Suash Deb ont proposé une nouvelle métaheuristique nommée la recherche coucou (CS). La recherche coucou est une métaheuristique très récente. Elle a enrichi le nombre des métaheuristiques à base de population de solutions. C'est une des variantes de l'algorithme d'optimisation par essaim particulaire (PSO). Les pionniers de l'algorithme CS se sont inspirés du comportement de reproduction parasitaire de quelques espèces de coucous qui pondent leurs œufs dans les nids des autres espèces en confiant la responsabilité d'incubation, de nourriture et d'élevage de leurs poussins aux oiseaux hôtes. Ces derniers peuvent détecter les œufs coucous dans leurs nids, dans ce cas là l'oiseau hôte va ou bien éjecter l'œuf coucou hors son nid ou abandonner son propre nid et construire un autre dans un autre emplacement. Yang et Deb se sont basés sur ce comportement parasitaire des coucous et sur le mécanisme du vol de Lévy qui permet la modélisation mathématique des déplacements aléatoires pour proposer une nouvelle méthode d'optimisation: La recherche coucou (CS). Malgré le nombre limité des travaux d'applications du CS **[32, 33, 34]** pour la résolution des problèmes d'optimisation, les résultats obtenus sont prometteurs en termes de performance et d'efficacité de la nouvelle métaheuristique dont le principe est élucidé dans ce qui suit.

2.3.1. Vol de Lévy

Le comportement de parasitisme de couvées chez les coucous est combiné dans l'algorithme de la recherche coucou avec les vols de Lévy pour améliorer la recherche d'un nouveau nid. Les vols de Lévy (Figure 2.3), baptisées par le mathématicien français Paul Lévy, représente un modèle des marches aléatoires caractérisées par leurs longueurs de pas qui suivent une distribution de loi de puissance qui s'écrit de la forme suivante:**[35]**

$$N(s)=s^{-t}$$

Dans la nature, les animaux cherchent de la nourriture de manière aléatoire ou quasi-aléatoire. En général, la recherche de la nourriture est effectivement une marche aléatoire parce que le prochain mouvement est basé sur l'emplacement/état actuel et la probabilité de transition à l'emplacement suivant. La direction qu'ils choisissent dépend implicitement d'une probabilité qui peut être modélisée mathématiquement. Différentes études ont montré que le comportement de vol de nombreux oiseaux et insectes a les caractéristiques typiques des vols de Lévy **[36].**

D'autre part Reynolds et Frye **[37]** ont montré que les mouches de fruits "Drosophila melanogaster" explorent leur paysage en utilisant une série de trajectoires droites ponctuées par un brusque virage à 90°, ce qui conduit à un modèle de recherche intermittente sans échelle d'un style de vol de Lévy. Ce modèle est communément représenté par de petits pas aléatoires suivis à long terme par de grands sauts **[38].** Un tel comportement de vol de Lévy a été appliqué à l'optimisation et les résultats ont montré une capacité prometteuse **[36].**

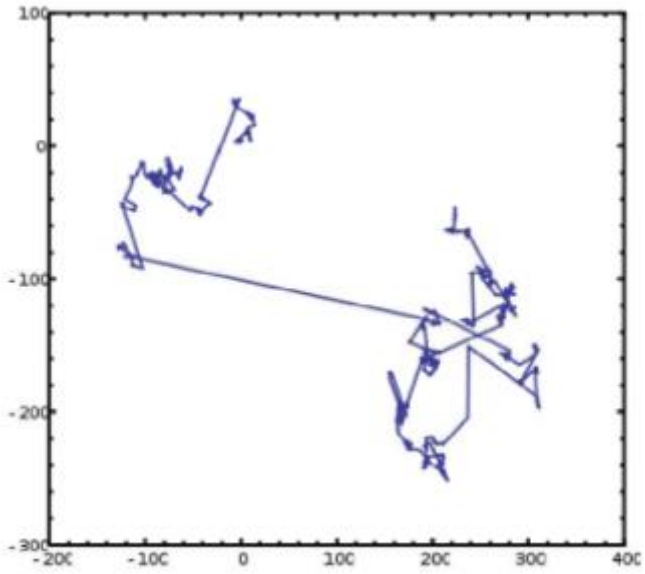


Figure 2.2. Exemple de 1000 pas par les vols de Lévy en 2 dimensions

2.4. Les principes de bases de l'algorithme de la recherche coucou (CS)

En s'inspirant du comportement des coucous dans leur reproduction, Yang et Deb se sont basés sur trois principes pour proposer leur nouvelle métaheuristique **[39].**

- Chaque coucou pond un œuf à la fois et choisit un nid aléatoirement ;
- Un bon nid de bonne qualité peut passer vers une nouvelle génération ;

- Le nombre de nids hôtes est fixé, et un œuf posé par un coucou peut être découvert par l'oiseau hôte suivant une probabilité $p_\alpha \in [0, 1]$.

Dans ce cas, l'oiseau hôte peut jeter l'œuf ou abandonner le nid et construire un nid complètement nouveau. Pour simplifier, cette dernière hypothèse peut être approchée par la probabilité pα de n nids d'être remplacés par de nouveaux nids.

L'algorithme de la recherche du coucou se résume autour des règles suivantes :

- Chaque œuf du coucou dans un nid représente une solution.
- Chaque oiseau de coucou pondra un seul œuf à la fois, et choisira sont nid de façon "aléatoire". Donc, chaque individu de la population des coucous a le droit de générer aléatoirement une seule nouvelle solution.
- Les meilleurs nids de meilleure qualité d'œufs nous mèneront vers les nouvelles générations. Ici, on a introduit implicitement la notion d'intensification ou la recherche autour des meilleures solutions.
- Certaines nouvelles solutions doivent être générées par les vols du Lévy autour de la meilleure solution obtenue jusqu'ici. Cela accélérera la recherche locale.
- Le nombre de nids hôtes est fixe, et l'œuf pondu par l'oiseau est découvert par l'hôte avec une probabilité pα ∈[0, 1]. Dans ce cas, l'oiseau hôte choisi de se débarrasser de l'œuf, ou d'abandonner le nid et de reconstruire un autre nid quelque part. Pour la simplification, cette dernière hypothèse sera approximée par la fraction pα des n nids qui sont remplacés par des nouveaux (nouvelles solutions aléatoires).
- Une fraction importante de nouvelles solutions doivent êtres générées par randomisation vers des régions lointaines et dont les emplacements doivent être assez loin de la meilleure solution actuelle, ce qui fera que le système ne sera pas pris au piège dans un optimum local.

2.5. Pseudo-code de l'algorithme de la Recherche Coucou (CS) par le vol de Lévy :

Les étapes de base de la recherche Coucou (CS) peuvent être résumées par le code pseudo illustré ci-dessous (figure 2.3) :

début

fonction objective f (x), x = (x 1, ..., X d) T

Générer la population initiale de n de nids hôtes xi (i = 1, 2, ..., n)

tant que (T <MaxGeneration) ou (critère d'arrêt)

1. Obtenez un coucou (i) aléatoirement par le vol de Lévy
2. évaluer sa qualité/ *fitness Fi*
3. Choisissez un nid parmi n (par exemple, j) aléatoirement
4. *T=T+1*

si (*Fi* > *Fj*), // les nouvelles solutions de nid j dominent ceux de nid i

Remplacer j par la nouvelle solution;

Fin si

Une fraction (P_a) des nids les plus mauvais est abandonnée et de nouveaux sont construites;

Gardez les meilleures solutions (ou des nids avec des solutions de qualité);

Triez les solutions et trouver ci le meilleur courant

fin tant que

Post-traitement des résultats et la visualisation

Fin

Figure 2.3. Pseudo-code de l'algorithme de la recherche Coucou (CS)

Un coucou i génère une nouvelle solution x^{t+1} via les vols de Lévy, selon l'équation :

$$x_i^{(t+1)} = x_i^{(t)} + \alpha \oplus \text{Lévy}(\lambda) \qquad (2.2)$$

Où α est la longueur du pas suivant la distribution des vols de Lévy montrée dans l'équation

$$\text{Lévy} \sim u = t^{*\lambda} \quad (1<\lambda\leq 3) \qquad (2.3)$$

Cette équation est stochastique (une marche aléatoire). En général, une marche aléatoire est une chaine de Markov dont l'étape suivante ne dépend que de l'étape actuelle, qui est le premier terme de l'équation, suivi par la probabilité de transition qui est le deuxième terme. Le produit $\oplus$ représente le produit matriciel. α est la longueur maximale du pas qui devrait être liée à l'échelle de l'espace de recherche du problème. Dans ce cas, α = 1. La marche aléatoire par le biais des vols de Lévy est plus efficace dans l'exploration de l'espace de recherche, que les autres marches aléatoires, vu que la taille de son pas est beaucoup plus grande à long terme.

2.6. Efficacité de la recherche coucou (comparaison avec d'autres métaheuristiques):

La faculté d'équilibrer la recherche entre l'exploitation des zones prometteuses et l'exploration de l'espace de recherche à l'échelle globale, est un indice de performance lié à toutes les métaheuristiques. La métaheuristique la plus robuste est celle qui balance parfaitement entre l'intensification et la diversification. CS a un meilleur contrôle d'équilibre de la stratégie de recherche intensive locale et une exploration plus efficace de l'ensemble de l'espace de recherche. Aussi, le nombre réduit de paramètres fait de CS un algorithme moins complexe et donc potentiellement plus générique. CS utilise un nombre réduit de paramètres de contrôle. En effet, CS utilise deux paramètres, la taille de la population n et les portions des nids abandonnés pa. En principe, n est fixé et pa essentiellement contrôle l'élitisme et l'équilibre entre la randomisation et la recherche locale. Le nombre réduit de paramètres rend l'algorithme moins complexe et donc potentiellement plus générique. Tout ceci est expliqué par la facilité de régler les deux paramètres de CS tout en gardant sa robustesse de résoudre une large gamme de problèmes d'optimisation même les problèmes qualifiés NP-difficiles. La comparaison de CS et de PSO et GA a montré que CS surpasse les performances de ces métaheuristiques. Effectivement, une étude analytique **[40]** a montré que DE, PSO et SA sont des cas particuliers de la recherche de coucou et il a été aussi montré que la recherche de coucou a une convergence globale. Des études théoriques sur l'optimisation des essaims de particules ont suggéré que la PSO peut converger rapidement vers la meilleure solution actuelle, mais pas nécessairement les meilleures solutions globales. En fait, certains analystes suggèrent que les équations de mise à jour des PSO ne satisferont pas aux conditions de convergence globale, et donc il n'y a aucune garantie pour la convergence globale. D'autre

part, il a prouvé que la recherche du coucou satisfait aux exigences de convergence globale et a ainsi garantit des propriétés. Cela implique que pour l'optimisation multimodale, la PSO peut converger prématurément vers un optimum local, alors que la recherche de coucou peut généralement converger vers l'optimalité globale. Un autre avantage de la recherche de coucou est que sa recherche globale utilise des vols Lévy, plutôt que des randonnées aléatoires standard. Comme les vols de Lévy ont une moyenne et une variance infinies, CS peut explorer l'espace de recherche plus efficacement que les algorithmes utilisant des processus gaussiens standard. Cet avantage, conjugué à la fois aux capacités locales et de recherche et à la convergence globale garantie, rend la recherche de coucou très efficace. En effet, diverses études et applications ont démontré que la recherche du coucou est très efficace **[40].**

2.7. Applications de la recherche coucou :

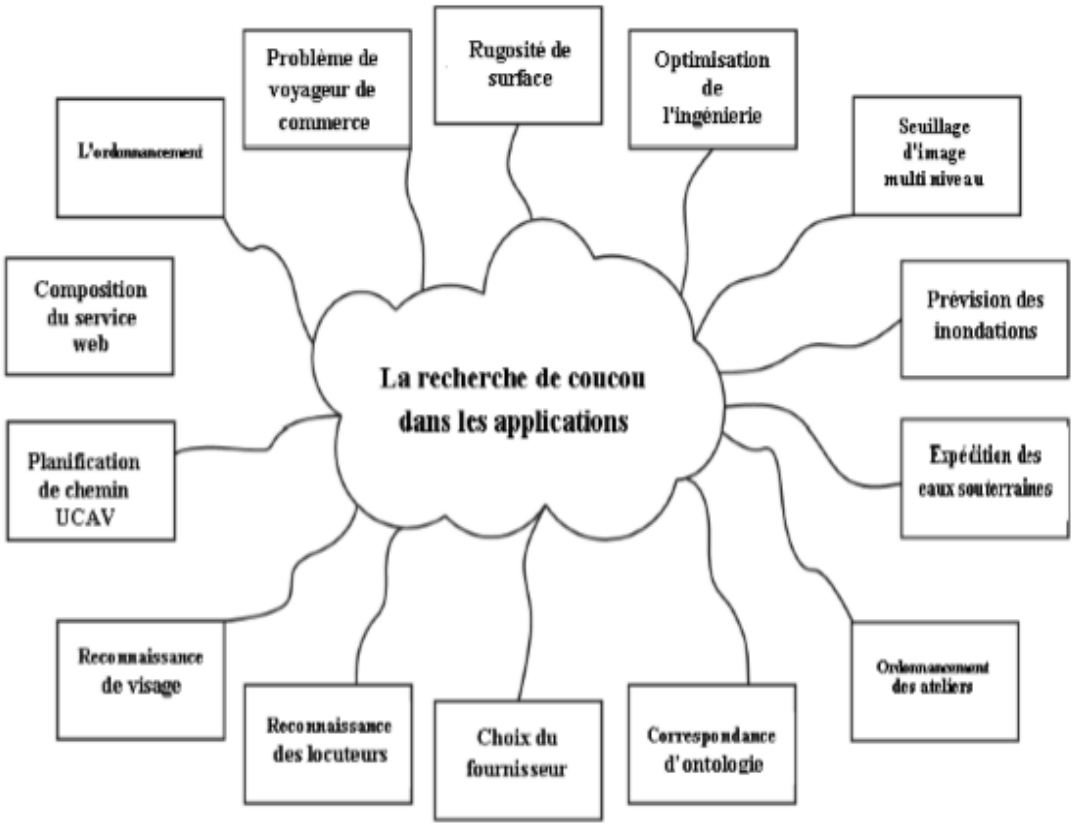


Figure 2.4. Applications de la Recherche de coucou.

Évidemment, l'optimisation de l'ingénierie fait partie des diverses applications. En fait, La recherche de coucou et ses variantes ont été appliquées dans presque tous les domaines des

sciences, de l'ingénierie et de l'industrie. Certaines des études d'application sont résumées dans le Tableau 3.1 **[41].**

Tableau 2.1. Applications de la recherche coucou

Application	Auteur	Année
Seuil d'image multi niveau	Brajevic et al.	2012
Prévision des inondations	Chaowanawatee and Heednacram	2012
Réseaux de capteurs sans fil	Dhivya and Sundarambal	2011
La fusion des données	Dhivya et al.	2011
Cluster dans les réseaux sans fil	Dhivya et al.	2011
Clustering	Goel et al	2011
Expédition des eaux souterraine	Gupta et al.	2011
Choix du fournisseur	Kanagaraj et al.	2012
Prévision de charge	Kavousi-Fard and Kavousi-Fard	2013
Rugosité de surface	Madic et al	2013
Planification de magasin de flu	Marichelvam	2012
Remplacement optimal	Mellal et al.	2012
Allocation DG dans le réseau	Moravej and Akhlaghi	2013
Optimisation du filtre Floraison	Natarajan et al.	2012
Réseau neuronal BPNN	Nawi et al.	2013
Problème Voyageur de commerce	Ouaarab et al.	2013
Composition du service Web	Pop et al.	2011
Composition du service Web	Chifu et al.	2012
Correspondance ontologique	Ritze and Paulheim	2011

Reconnaissance des locuteurs	Sood and Kaur	2013
Tests automatisés de logiciel	Srivastava et al.	2012
Optimisation de fabrication	Syberfeldt and Lidberg	2012
Reconnaissance de visag	Tiwari	2011
Formation des modèles neuronaux	Vázquez	2013
Expédition économique non convexe	Vo et al.	2013
Planification des trajets UCA	Wang et al.	2012
Optimisation des activités	Yang et al.	2012
Sélection des paramètres d'usinage	Yildiz	2013
Planification des travaux en grille	Prakash et al.	2012
Affectation quadratique	Dejam et al.	2012
Problème d'emboîtement de feuilles	Elkeran	2013
Optimisation des requêtes	Joshi and Srivastava	2013
Puzzle des N reines	Sharma and Keswani	2013
Jeux d'ordinateur	Speed	2010

2.8. Conclusion

Nous avons vu dans ce chapitre la méthode de recherche de coucou en détail, nous avons également présenté le comportement de coucou, le vol de Lévy ainsi que les principes de base de l'algorithme de coucou en spécifiant le pseudo-code de cet algorithme par le vol de Lévy.

Chapitre 3
La conception de l'application

3.1. Introduction

Dans le chapitre précédent, nous avons présenté les notions et les concepts de base de la recherche coucou, nous allons présenter dans ce chapitre son adaptation au problème de clustering.

Nous commençons par la description de cette adaptation, nous passons ensuite à la présentation de l'objectif de notre application ainsi que la mise en évidence de l'architecture générale de notre système et la description de ses modules qui constitue une étape fondamentale qui précède l'implémentation, nous détaillons, également, les différents diagrammes et scénarios à implémenter dans la phase suivante. Ceci permettra de mieux comprendre notre application.

Pour cela nous allons utiliser les diagrammes de flux de données pour la modélisation de notre application.

3.2. Clustering basé sur la Recherche Coucou

La recherche de Coucou (CS) est l'un des algorithmes métaheuristique. Cet algorithme fonctionne sur la base de la stratégie de reproduction de l'oiseau coucou. Notre travail représente un algorithme de clustering de données basé sur l'optimisation par la recherche de coucou. La recherche de coucou est générique et robuste pour de nombreux problèmes d'optimisation et dispose de fonctionnalités intéressantes.

Pour résoudre le problème de clustering de données, l'algorithme de recherche de coucou est adapté pour trouver les centroïdes des clusters. Pour faire cela, nous supposons que nous avons *n* objets (données) et chaque objet est défini par *m* attributs. Dans notre travail, l'objectif principale du CS est de trouver *k* centroïdes des clusters qui réduisent au maximum la fonction objectif dénotée par SSE qui représente une mesure de qualité interne, elle permet de calculer la distance intra cluster (la somme des distances entre les points et leur centroïde de cluster correspondant), elle est définie par l'équation (3.1). Sachant que l'ensemble de données (jeu de données) doit être représenté par une matrice (n,m), tel que *n* représente le nombre d'objets (le nombre de points de données) et *m* représente le nombre des attributs (le nombre de dimensions).

$$\text{SSE}=\sum_{i=1}^{k}\sum_{j=1}^{n} W_{ij} * \sqrt{\sum_{p=1}^{w}(O_{jp} - C_{i\,p})^2} \qquad (3.1)$$

Où $W_{ij} = 1$ si l'objet est dans le cluster et **0** sinon, **k** est nombre de cluster, **n** est le nombre d'objet, **m** est le nombre des attributs, et est c_{ip} la valeur de l'attribut numéro **p** du centroïdes de cluster numéro **i [42].**

Dans le mécanisme de recherche de coucou, les nids sont des solutions et dans notre travail une solution est un ensemble de centroïdes représenté par un vecteur de dimension **k*m** (où **k** est le nombre de centroïdes des clusters et **m** représentent le nombre d'attributs).

3.3. Les étapes principales de notre système:

Notre système est composé des étapes suivantes :

- Déterminer un jeu de données à partitionner.
- Appliquer l'algorithme CS sur un jeu de données.
- Calcules utilisant des mesures d'évaluation internes (la fonction objective SSE).
- Evaluer l'indice FE qui représente le nombre d'évaluations de la fonction objectif que l'algorithme effectue jusqu'à l'obtention de la meilleure valeur de cette fonction.
- Représenter les résultats graphiquement.

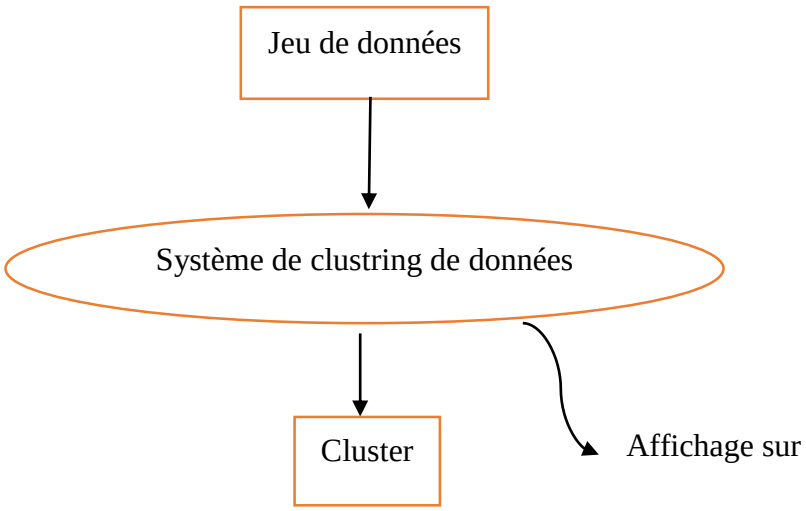


Figure 3.1. Diagramme de flux de données (DFD)-niveau 1-schéma global

3.4 L'architecture générale de notre système:

L'architecture globale de notre système et ses différents modules sont illustrés dans la figure ci-dessous :

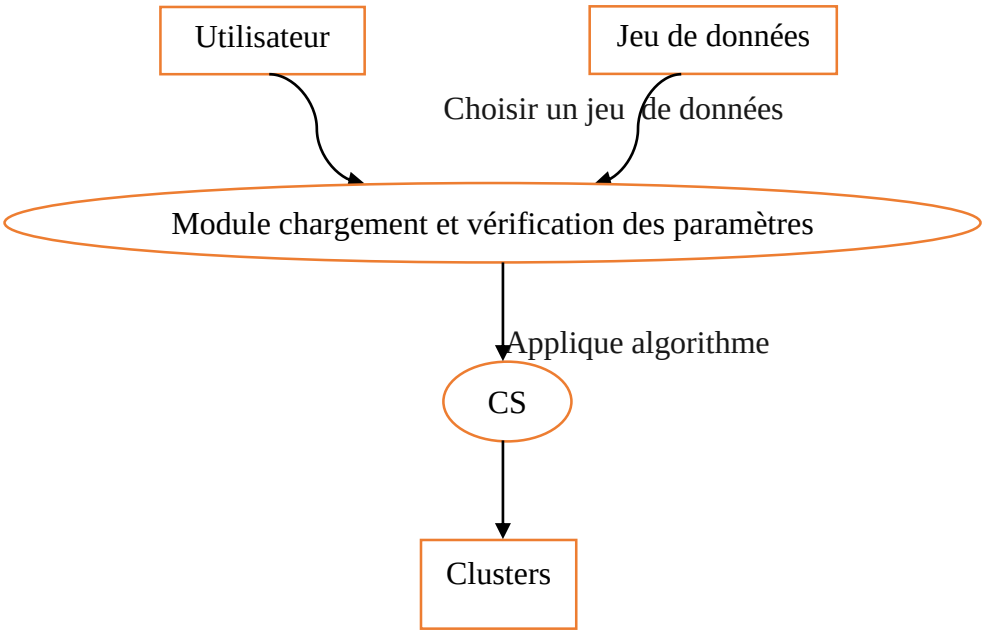


Figure 3.2. Diagramme de flux de données (DFD)-niveau 2

Le système permet à l'utilisateur de :

- Choisir un jeu de données.
- Chargement et affichage des paramètres du jeu choisi : nombre de dimensions (nd), nombre de clusters (K) et le nombre de points (nbr).
- Applique algorithme: CS.
- Affichage des résultats.

3.5 Module CS (Cuckoo Search):

Ce module permet d'appliquer l'algorithme CS à un jeu de données, cet algorithme est constitué de plusieurs étapes qui sont :

- Initialisation aléatoire des centroïdes.
- Affectation des points aux clusters.
- Evaluer les solutions en utilisant la fonction objective
- Trier les solutions actuelles et trouver la meilleure.
- Générer des solutions en utilisant les vols de Lévy
- Abandonner les mauvaises solutions avec une probabilité Pa
- Remplacer la solution courante par une meilleure et calculer l'indice FE.

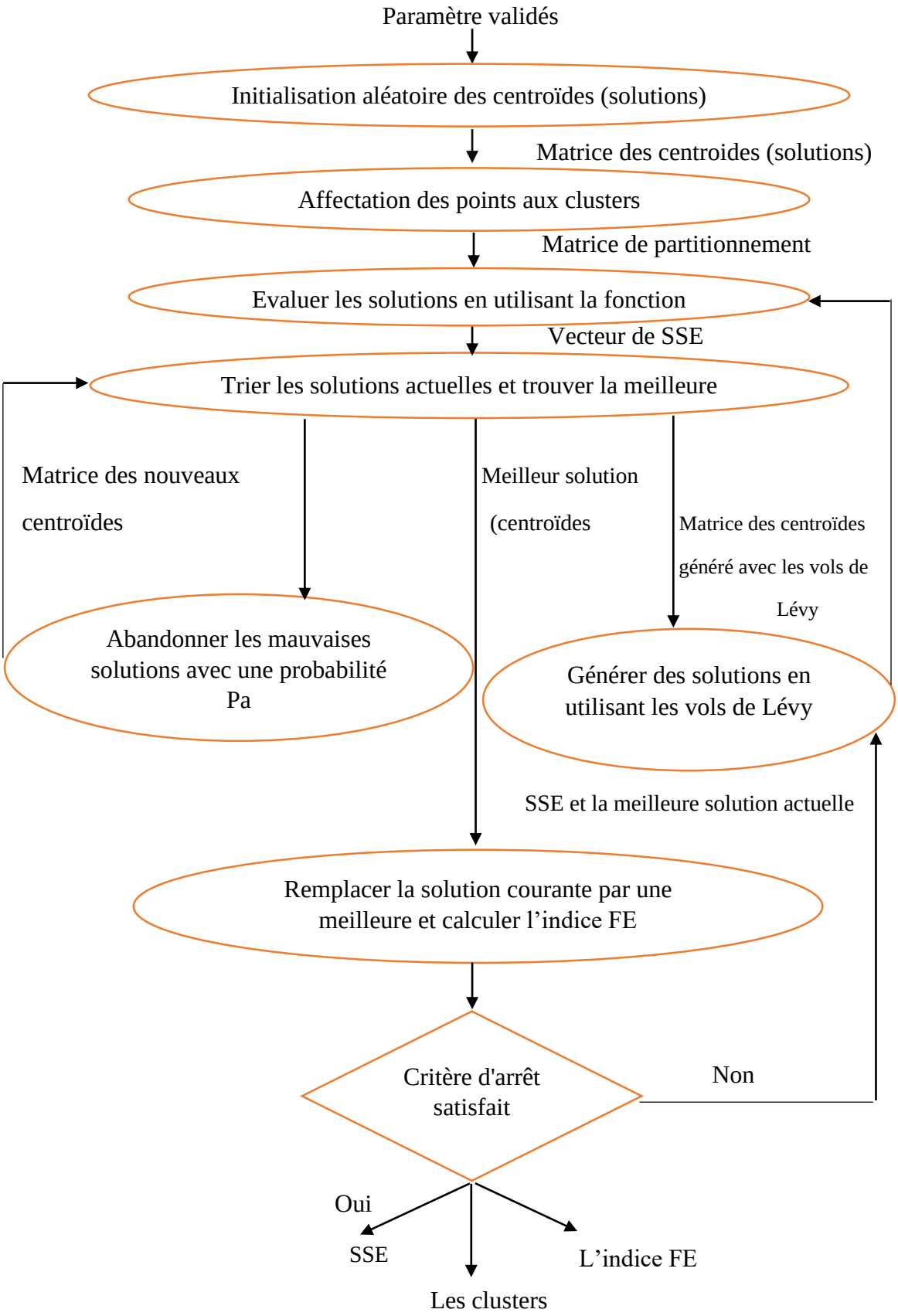


Figure 3.3. Diagramme de flux de données(DFD)- niveau 3- Algorithme CS

La description des étapes de l'algorithme CS est la suivante :

A. Initialisation aléatoire:

Cette étape permet de choisir aléatoirement les centroïdes à partir d'un jeu de données.

B. Affectation des points aux clusters :

Après avoir initialisé les centroïdes, on affecte les points aux clusters en calculant la distance entre les points et les centroïdes en utilisant la formule suivante :

$$d = \sqrt{\sum_{p=1}^{m}(O_{jp} - C_{i\,p})^2} \qquad (3.2)$$

C. Evaluer les solutions en utilisant la fonction objectif :

Cette étape permet d'évaluer les solutions (centroides) en minimisant la fonction objectif SSE représentée dans la formule (3.1).

D. Trier les solutions actuelles et trouver la meilleure :

Après avoir évalué les solutions, on effectue le tri de celle-ci de façon croissante en utilisant ses fonctions objectif, ce qui permet de trouver la meilleure solution.

E. Calcul des centroides avec le vol de Lévy :

En appliquant le vol de Lévy, cette étape permet de générer des solutions (calcul de centroïdes) en utilisant les formules suivantes

$$S = \frac{u}{|v|^{1/\beta}} \qquad (3.3)$$

$$u \sim N(O,\sigma_u^2)\ ,\ v \sim N(O,\sigma_v^2) \qquad (3.4)$$

$$\sigma_u = \left\{\frac{\Gamma(1+\beta)\sin(\pi\beta/2)}{\Gamma\left[\frac{1+\beta}{2}\right]\beta 2^{(\beta-1)/2}}\right\}^{1/\beta} \qquad (3.5)$$

Où S représente la longueur du pas par la distribution de vol de Lévy.

F. Abandonner les mauvaises solutions (centroïdes) avec une probabilité P_a :

Cette étape permet d'abandonner les mauvaises solutions avec une probabilité p_a et de construire des nouveaux.

G. Remplacer la solution courante par une meilleure et calculer l'indice FE :

Cette étape permet de remplacer la solution courante par une meilleure en utilisant un test de comparaison entre les valeurs de la fonction objectif courante et l'ancienne. Elle permet

également de calculer le nombre d'évaluations (FE) de la fonction objectif que l'algorithme CS effectue pour obtenir la meilleure valeur de la fonction objectif SSE.

A. Critère d'arrêt :

Le critère d'arrêt de cet algorithme est un nombre d'itération limité, si ce dernier atteint la fin des itérations l'algorithme s'arrête sinon il va continuer.

3.6. Conclusion:

Nous avons vu dans ce chapitre la description et l'adaptation de la méthode de recherche de coucou pour le clustering de données, suivi par déterminé l'idée de base de notre approche, ainsi que les étapes principales et l'architecture générale de notre système.

Dans le prochain chapitre nous allons présenter deux parties, la première partie implémentation de notre approche en déterminant le langage de programmation utilisé, la structure de données utilisées dans notre application ainsi que le déroulement d'une session de travail et la partie seconde l'expérimentation en montrant les tests, les résultats et leur analyse.

Chapitre 4
Implémentation de l'approche

Chapitre 4 : Implémentation de l'approche

4.1. Introduction :

Dans ce chapitre, nous présentons notre travail. Deux volets vont être abordés : le premier concerne l'implémentation de l'application. D'abord, nous commençons par les composants de l'environnement de travail et le langage utilisé, suivi par la présentation de différentes fenêtres de notre application, le déroulement d'une session de travail. Le deuxième volet concerne la visualisation graphique des résultats basée sur différents outils suivie par une phase d'analyse et interprétation de résultats.

4.2. Implémentation :

4.2.1. Composants de l'environnement de travail :

- Le système d'exploitation : Windows 10 professionnel.

- Les outils utilisés : PC (Intel®Calerons®, CPU N2840@ 2,16 GHz, 2.00 Go de RAM)

- Le langage de programmation : MATLAB R2014a

4.2.2. Représentation du langage MATLAB :

MATLAB : (MATrixLABoratory) est un langage de programmation de quatrième génération et un environnement de programmation interactif pour le calcul scientifique et la visualisation des données, il est développé par la société The Math Works

4.2.3. Les structures de données :

Pour la représentation de différents composants de notre application, nous avons utilisé la notion de tableau

4.2.3.1. La notion des tableaux :

Un tableau est représenté par une liste qui contient les éléments d'une manière contiguë et accessible.

4.2.3.2. Variables utilisés

– K : nombre de clusters.

- nd : nombre de dimensions des données.

- nbr : nombre des points de données.

- Data [nbr, nd] : Matrice des données avec nd dimensions et nbr points.

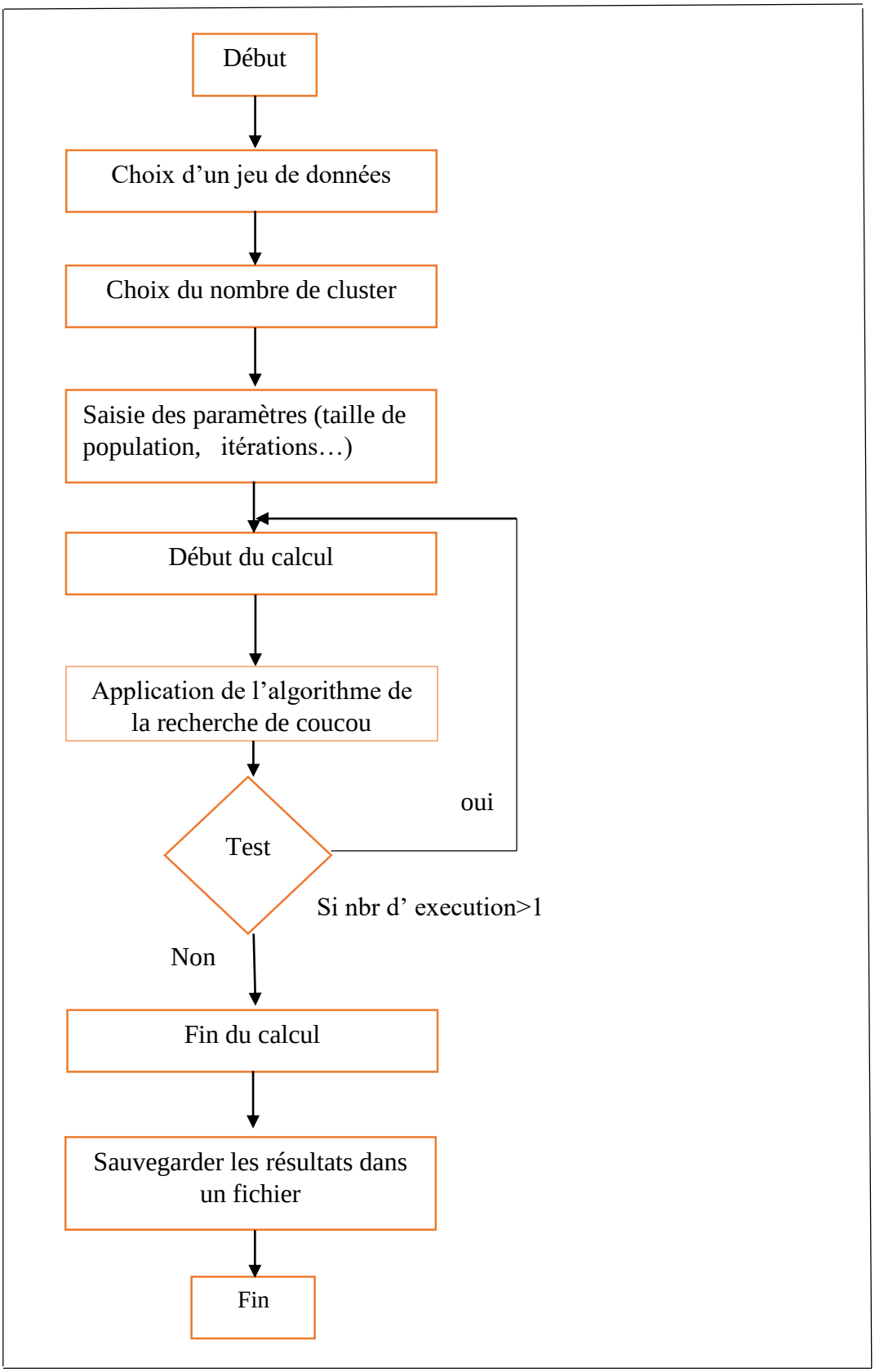


Figure 4. 1. Déroulement d'une session de travail

4.2.5. Présentation de l'application :

Notre application se compose d'une fenêtre principale (voir figure 4.2) qui est divisée en 5 sections.

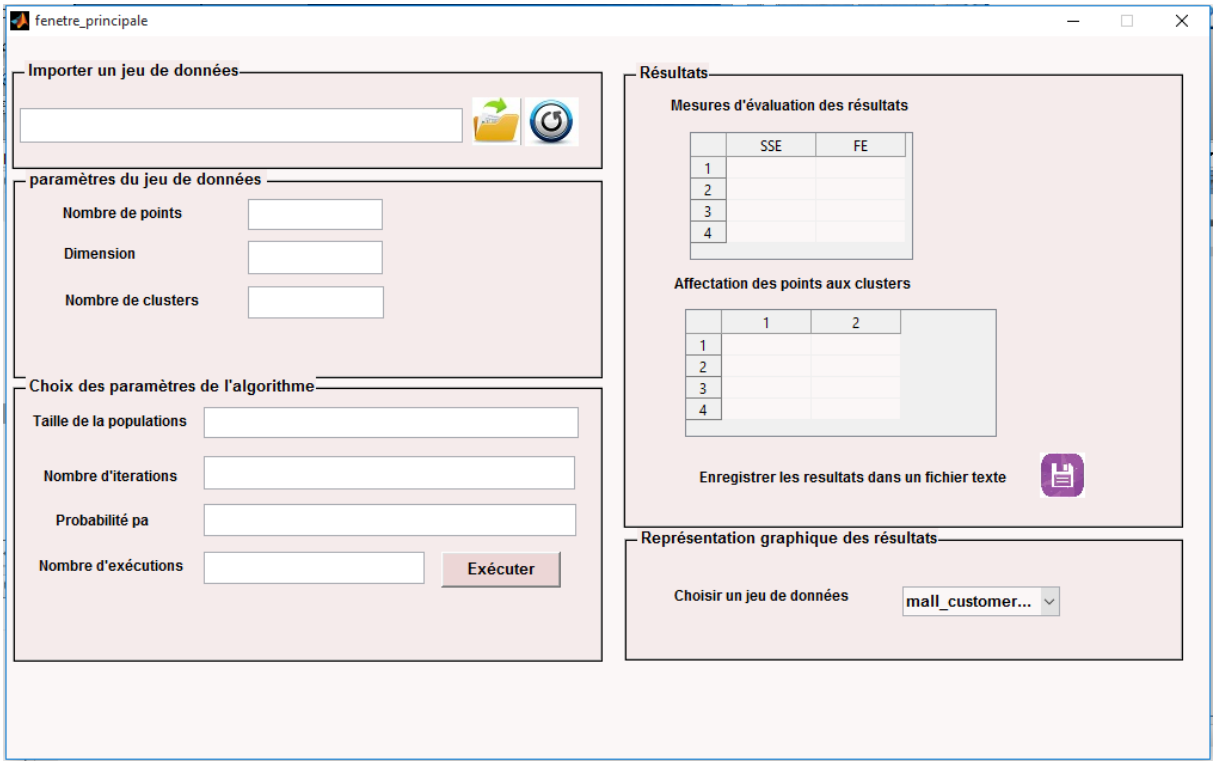


Figure 4.2. Fenêtre principal

1) Section 1 « Importer un jeu de données» : c'est là où l'utilisateur peut faire l'importation d'un jeu de donnée en cliquant sur le bouton  suivi par une boite de dialogue pour sélectionner un fichier d'une extension *.mat* parmi les jeux existant (voir figure 4.3)

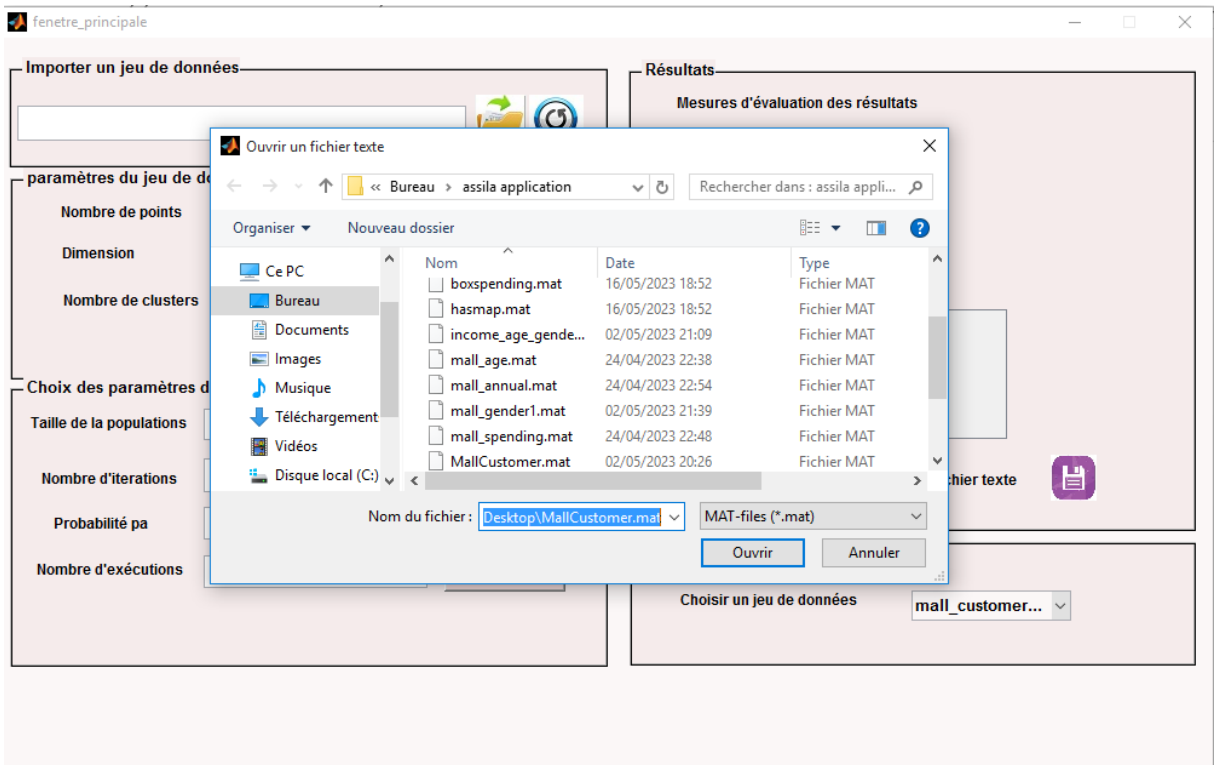


Figure 4.3. Fenêtre de la sélection du jeu de donnée

2) Section 2 « Paramètre du jeu de données » : application import les paramètres constantes du jeu donnée qui sont : Le nombre de clusters **K**, le nombre de points **nbr** et nombre de dimensions **nd.** Qui peut être modifiée par l'utilisateur sans modifier le fichier original. du jeu de donnée.  Le bouton est pour actualiser les paramètres par défaut (champs vide) (voir Figure 4.4).

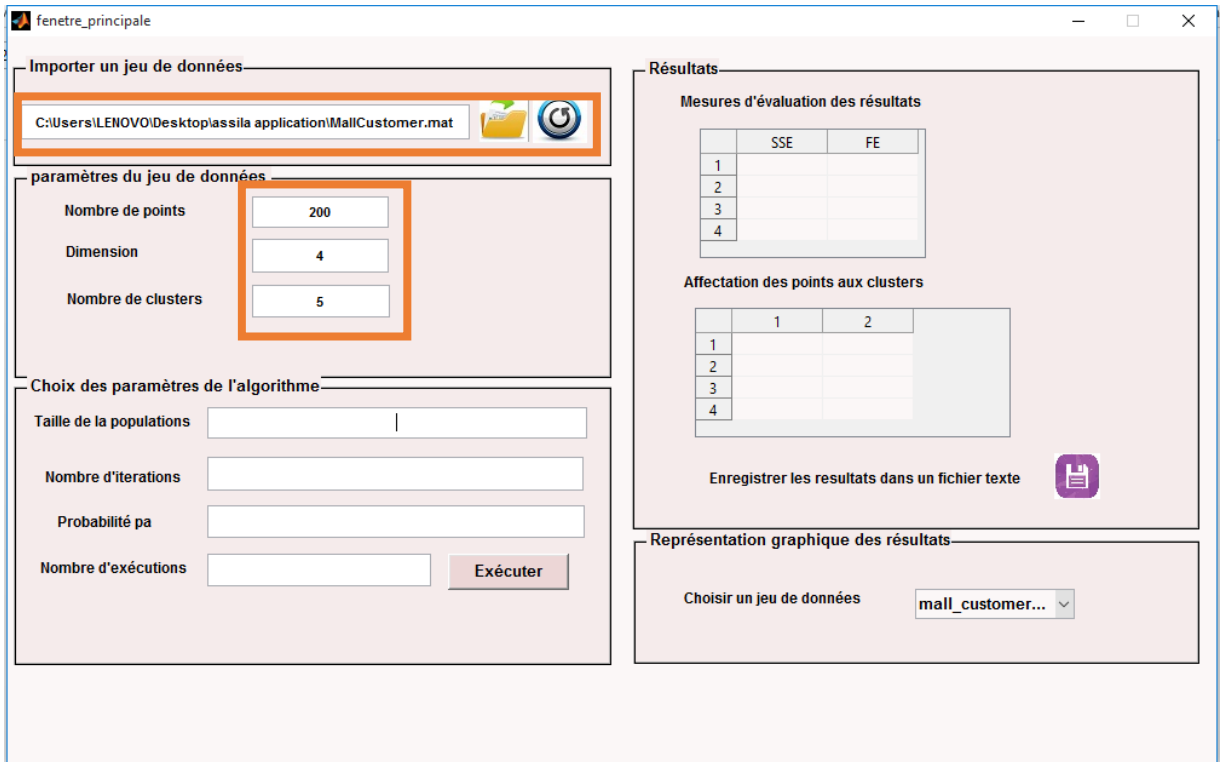


Figure 4. 4. Paramètres du jeu de données

3) Section 3 «Choix des paramètres de l'algorithme» : après la saisie de chaque paramètre tel que : la taille de la population, le nombre d'itérations, probabilité et nombre d'exécutions, l'utilisateur clique sur le bouton «Exécuter» pour démarrer l'exécution. Une barre de progression s'affiche pour chaque exécution (voir Figure 4.5)

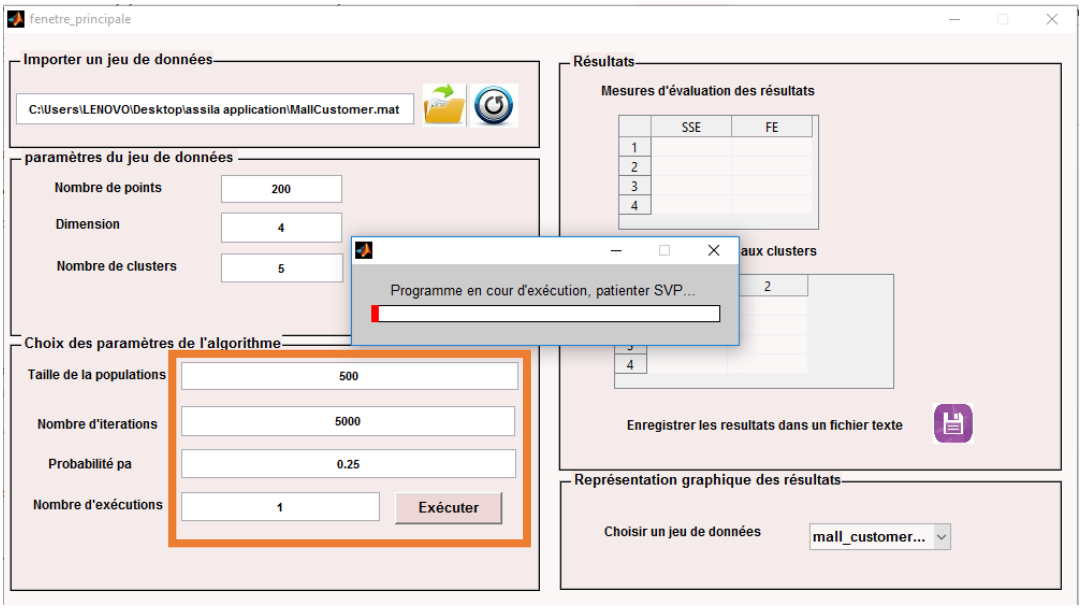


Figure 4. 5. La saisie des paramètres et barre de progression après l'exécution

4) Section 4 «Résultats» : Un foie que l'exécution termine, les résultats s'affiche dans les deux tableaux : Mesures d'évaluation des résultats et affectation des points aux clusters (voire Figure 4.6).

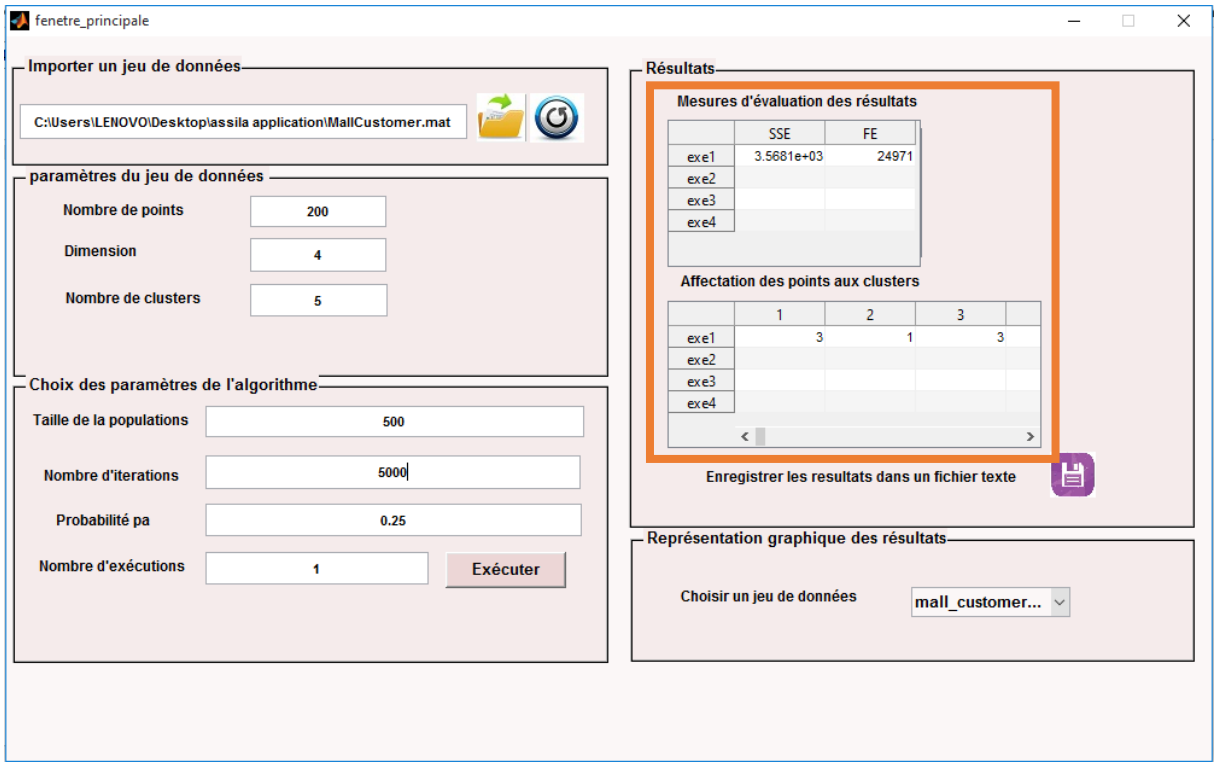


Figure 4.6. Les résultats d'exécutions

L'utilisateur peut aussi cliquer sur le bouton pour sauvegarder les résultats dans un fichier texte (voir Figure 4.7).

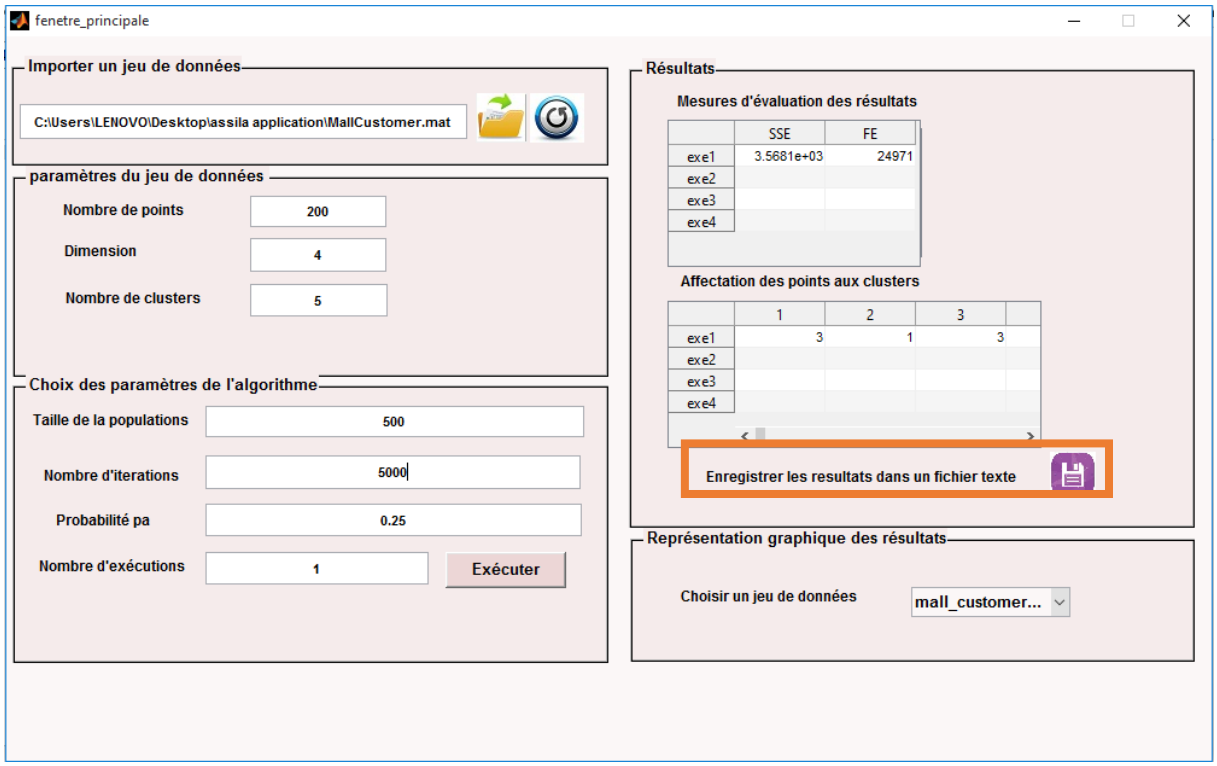


Figure 4.7. Enregistrement des résultats

5) Section 5 «La représentation graphique» : après le choix d'un jeu de données, les fenêtres de résultats s'affichent telles que : la visualisation du clustering des points, les diagrammes à barres, les diagrammes circulaires et les boxplot.

4.3. Résultats et analyse :

4.3.1. Jeu de données:

Le jeu de données utilisé ici est le jeu de données appelé Mall Customer extrait du site [kaggle datasets download –dvchoudhary / Customer-segmentation-tutorial-in-python] , il contient les propriétés de 200 clients de. Ce jeu de donnes inclut 4 attributs :

Age (âge,) Annual income (le revenu annuel), Spending (les dépenses), Gendre (le gêner) :

Tableau 4.1. Résumé du jeu de données de Mall Customer

Jeu de données	K	Dim	M
MallCustomer	5	4	200
Mall_ âge	5	1	200
Mall_annual	5	1	200
Mall_spending	5	1	200
Mall_gender1	5	1	200
Age_annual	5	2	200
Age_spending	5	2	200
Annual_gender	5	2	200
Annual_spending	5	2	200
Gender_age	5	2	200
Spending_gender	5	2	200
Income_age_gender	5	3	200
Spending_income_gender	5	3	200
Spending_age_gender	5	3	200
Spending_age_annual_3	5	3	200

Dans le tableau 4.1 le jeu de données de un seul attribute (mall_age, mall_annual, mall_spending, mall_gender1), deux attribute (age_annual, age_spending, annual_gender, annual_spending, gender_age, spending_gender), Trios attribute (income_age_gender, spending_income_gender, spending_age_gender, spending_age_annual_3), et quater attribute (mall customer).

4.3.2. L'évaluation des résultats :

L'évaluation est faite en utilisant la fonction interne SSE, elle calcule la somme de la distance entre des points de données et leurs centroïdes de clusters correspondants

$$SSE=\sum_{i=1}^{n}(y_i - y'_i)^2 \qquad (4.1)$$

Où :

y_i : La valeur observée

y'_i : La valeur estimée par la ligne de régression

Une autre evaluation est faite en calculant le nombre d'évaluations de la fonction objectif FE pour avoir une idée de la vitesse et la qualité de convergence de l'algorithme.

4.3.3. Résultats des tests :

Tests sur le jeu de données Mall Customer

Nous avons appliqué l'algorithme CS pour le clustering du jeu de données Mall Customer en fixant les paramètres suivants :

- Nombre de population : 25
- Nombre d'itérations : 1000
- Probabilité Pa : 0.25
- Nombre d'exécutions : 1

Les résultats numériques sont montrés dans les tableaux ci-dessous (Tableau 4.2, 4.3) qui montrent les valeurs de chacune des mesures SSE et FE.

Tableau 4.2.Valeurs de la fonction objectif SSE obtenus

Jeu de données	Cuckoo Search
MalCustomer	3.5310 e+03
Mall_age	464.0535
Mall_spending	821.0082
Mall_annual	970.0310
Mall_gender1	0
Spending_income_gender	2.5779 e+03
Income_age_gender	2.2322 e+03
Spending_age_gender	2.7079 e+03
Spending_age_annual_3	3.5288 e+03
Age_annual	2.2279 e+03
Age_spending	1.9256 e+03

Annual_gender	986.6834
Annual_spending	2.5753 e+03
Gender_age	484.5705
Spending_gender	837.3376

Tableau 4.3. Valeurs de FE obtenu

Jeu de données	Cuckoo Search
Mal Customer	24998
Mall_age	24977
Mall_spending	24992
Mall_annual	24979
Mall_gender1	24976
Spending_income_gender	24979
Income_age_gender	24980
Spending_age_gender	24995
Spending_age_annual_3	24987
Age_annual	24999
Age_spending	24998
Annual_gender	24988
Annual_spending	24987
Gender_age	24980
Spending_gender	24985

4.3.4. Analyse du jeu de données basée sur les boxplots :

boxplot est un outil graphique utilisé pour la représentation visuelle de la distribution d'un ensemble de données.il permet de visualiser les quartiles, la médiane, les valeurs aberrantes ainsi que l'étendu des données (voir Figure 4.8)

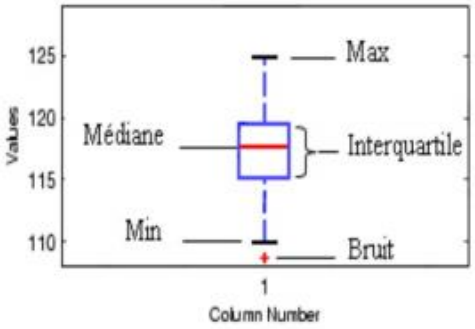


Figure 4.8. Informations données par un Boxplot

Analyse de boxplot :

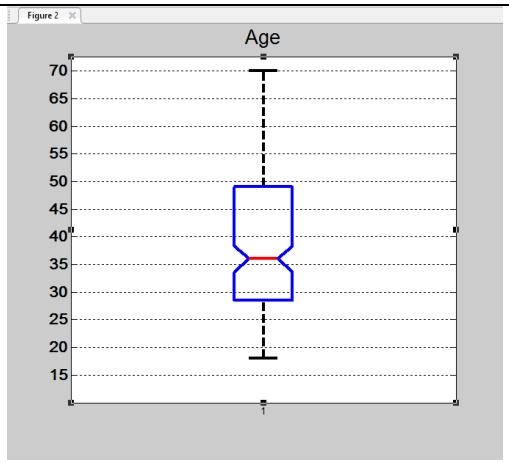


Figure 4.9. Boxplot d'âge

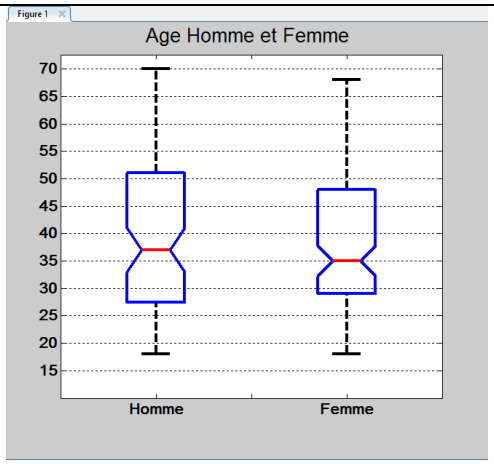


Figure 4.10. Boxplot d'âge des hommes et des femmes

Figure 4.11. Donne la représentation graphique de l'attribut âge basée sur le boxplot.

La figure 4.10 sépare entre l'âge des femmes et des hommes.

L'âge des clients s'étale entre 18 et 70 cela veut dire que les clients de ce mall sont presque de tout âge mais la majorité ont l'âge compris entre 28 et 49.

La majorité des femmes ont l'âge entre 29 et 48 tandis que les hommes ont l'âge entre 27 et 51.

Le centre de l'âge des femmes est 35. Le centre d'âge des hommes est 37.

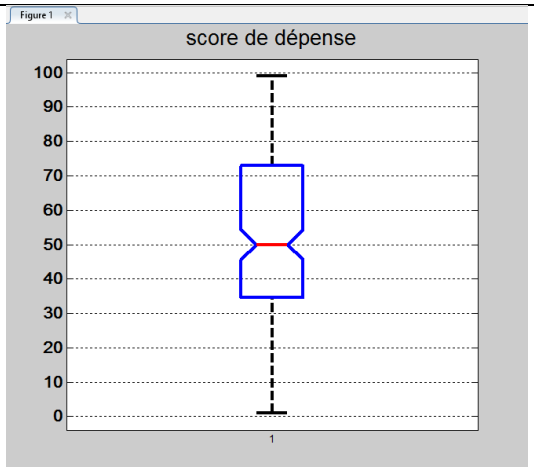


Figure 4.12.boxplot des dépenses

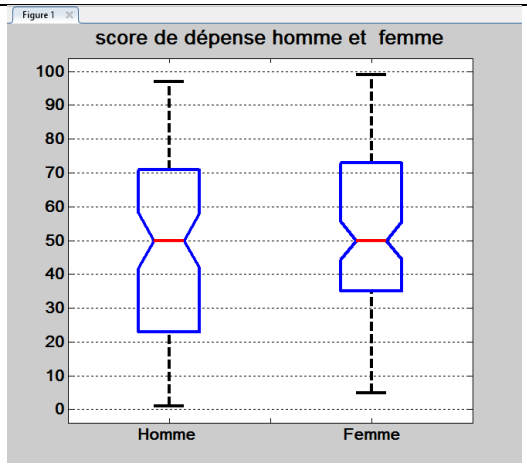


Figure 4.13.boxplots des dépenses des hommes et femmes

La figure 4.12 donne la représentation graphique des dépenses de tous les clients basés sur le boxplot.

La figure 4.13 donne la représentation graphique des dépenses des hommes et des femmes basée sur les boxplot.

Le score de dépense varie entre 1 et 99, en général, les clients ont un score de dépense entre 35 et 72 et la grande partie de ces clients ont un score supérieur à 50.

Si on compare entre les hommes et les femmes, on peut remarquer que la majorité des femmes ont un score de dépense supérieur à 50 par contre, la majorité des hommes ont un score inférieur à 50.

Les femmes ont un score de dépense allant de 35 à 73 tandis que le score de dépense des hommes varie de 23 à 71.

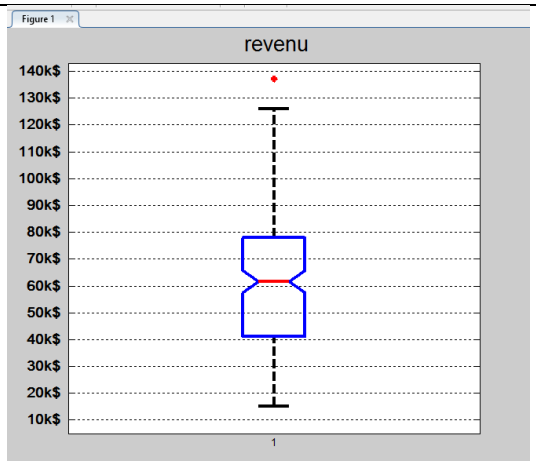


Figure 4. 14.représentation graphique de boxplot le revenu

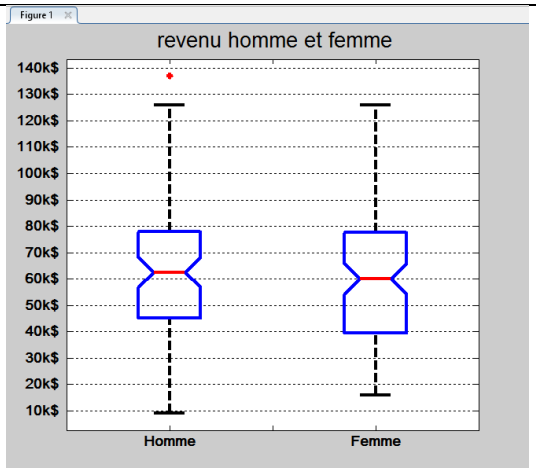


Figure 4. 15. représentation graphique de boxplot le revenu homme et femme

La figure 4.14 présente le boxplot des revenus des clients du mall.

La figure 4.15 présente le boxplot des revenus des femmes et des hommes.

Le revenu des clients s'étends sur l'intervalle [15,137] k$. La majorité des clients ont leurs revenu entre 41 k$ et 79 k$. La médiane du revenu est 61 k$.

Les hommes ont un revenu supérieur à celui des femmes. Le revenu des femmes varie entre 40 et 78 k$ par contre celui des hommes varie entre 45 et 79 k$

4.3.5. Analyse du jeu de données basée sur les graphiques à barres et les diagrammes circulaires :

Le graphique a barres, également appelé histogramme. C'est une représentation graphique utilisée pour afficher des données catégoriques ou discrètes sous forme de barres rectangulaires.

Le diagramme circulaire est utilisé pour représenter la répartition proportionnelle d'une variable en différentes catégories. Chaque catégorie est représentée par un secteur circulaire dont la taille est proportionnelle à la valeur qu'elle représente

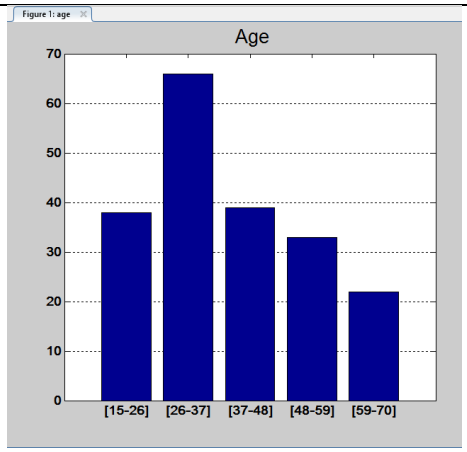


Figure 4.16. Graphique à barres d'âge

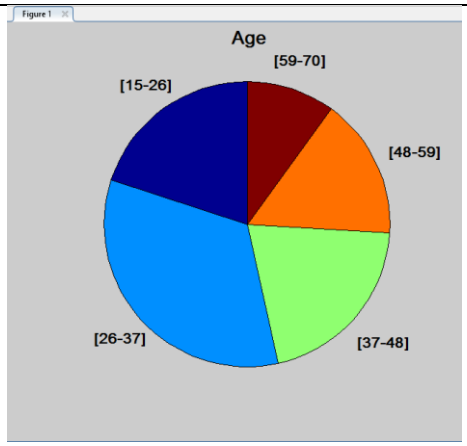


Figure 4.17. Diagramme circulaire d'âge

La figure 4.16 représente le graphique à barres d'âge. La figure 4.17représente le diagramme circulaire d'âge. Il est clair que la plus grande catégorie est celle des clients ayant l'âge entre 26 et 37 et la plus petite catégorie est la catégorie des clients les plus âgés ayant l'âge entre 59 et 70

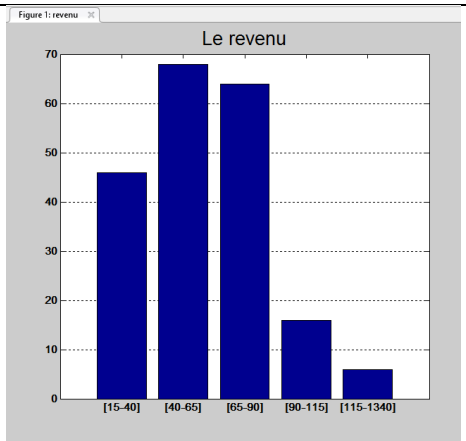


Figure 4.18. Graphique à barres de revenu annuel

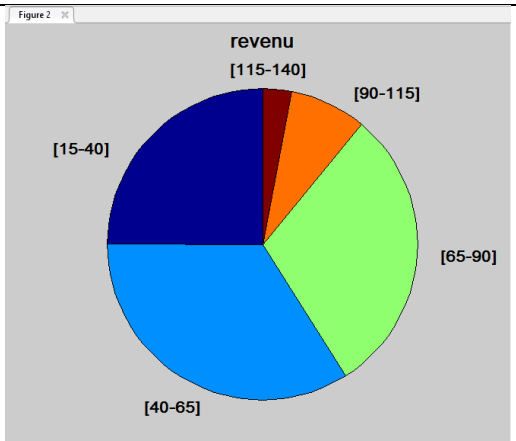


Figure 4.19. Diagramme circulaire de revenu annuel

La figure 4.18 représente le graphique à barres de revenu annuel. La figure 4.19 représente le diagramme circulaire de revenu annuel.

Les catégories les moins denses sont les catégories des clients ayant un revenu annuel élevé, c'est le cas des deux catégories [115-140] et [90-115].

La catégorie la plus dense est la catégorie ayant un revenu moyen entre 40 et 65 k$.

La majorité des clients ont un revenu annuel entre 15 et 90 k$.

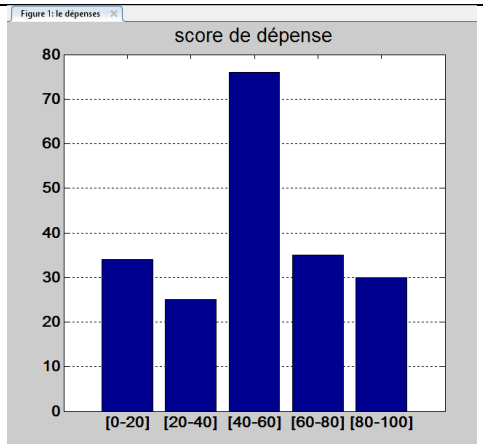


Figure 4.20. Graphique à barres de score de dépense

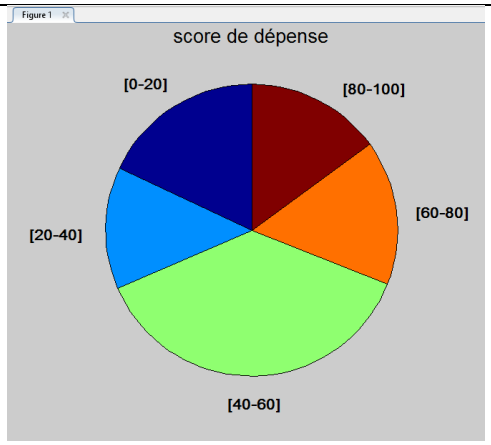


Figure 4.21. Diagramme circulaire de score de dépense

La figure 4.20 représente le graphique à barres de score de dépense. La figure 4.21 représente le diagramme circulaire de score de dépense.

On peut voir clairement que la majorité des clients ont un score de dépense moyen entre 40 et 60 k$. Il y a presque une équité entre les autres catégories de score de dépense élevé et faible

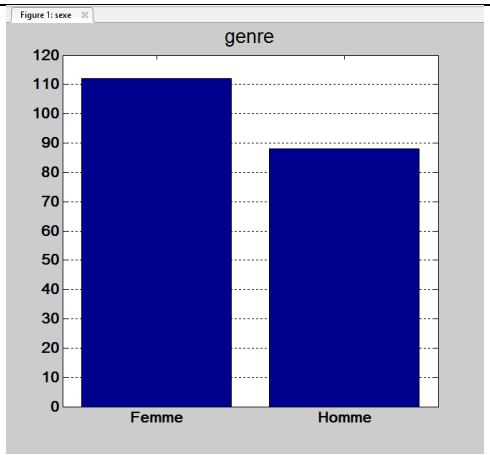


Figure 4.22. Graphique à barres du genre

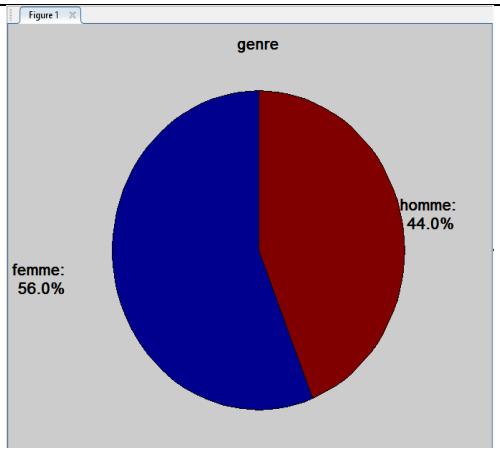


Figure 4.23.Diagramme circulaire de score de dépense

Le nombre de clientes femmes dépasse le nombre de clients hommes. Un pourcentage de 56% de femmes contre 44% d'hommes.

4.3.5. Analyse du jeu de données basée sur le cluster :

La représentation graphique du clustering du jeu de données Mall Customer est représentée dans la figure 4.24 et puisque le jeu de données est constitué de 4 dimensions, nous présentons la distribution de résultat de clustering selon les 6 paires de dimensions.

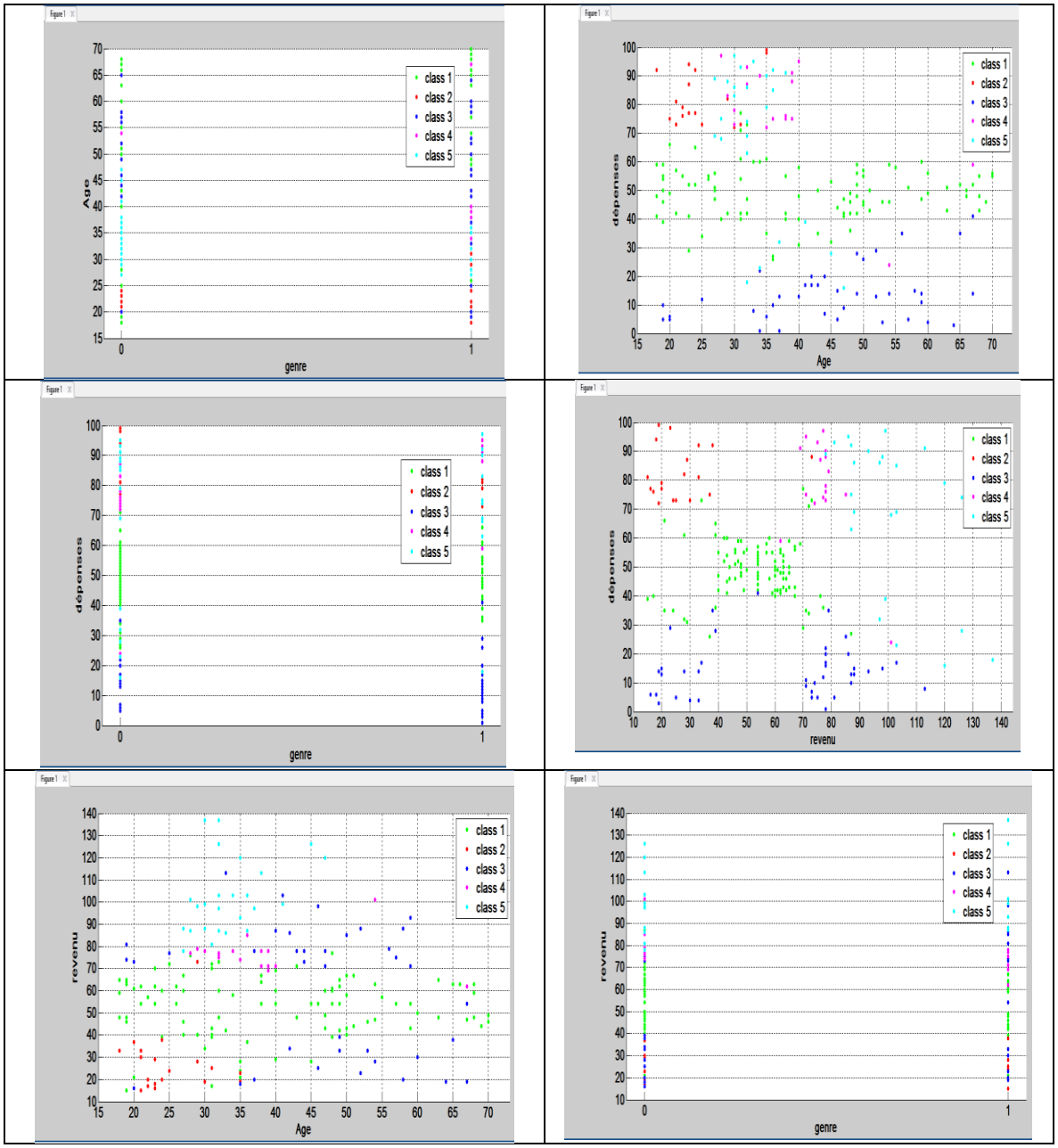


Figure 4. 24. Représentation graphique de clustering de Mall Customer.

Sur la figure 4.24, la catégorie de couleur verte représente les clients de revenu annuel moyen et un score de dépense aussi moyen. Ce sont des clients qui font un équilibre entre le revenu

et les dépenses. Ce sont des clients de différents âges, leurs âges s'étalent entre 18 et 70. Cette catégorie de clients contient des femmes plus que les hommes.

La catégorie de couleur rouge représente la catégorie de clients de revenu annuel faible et un score de dépense élevé. L'âge des clients de cette catégorie est concentré entre 20 et 25 et peut s'étaler jusqu'à 30. Les clients ayant l'âge entre 20 et 25 de cette catégorie sont beaucoup plus des femmes.

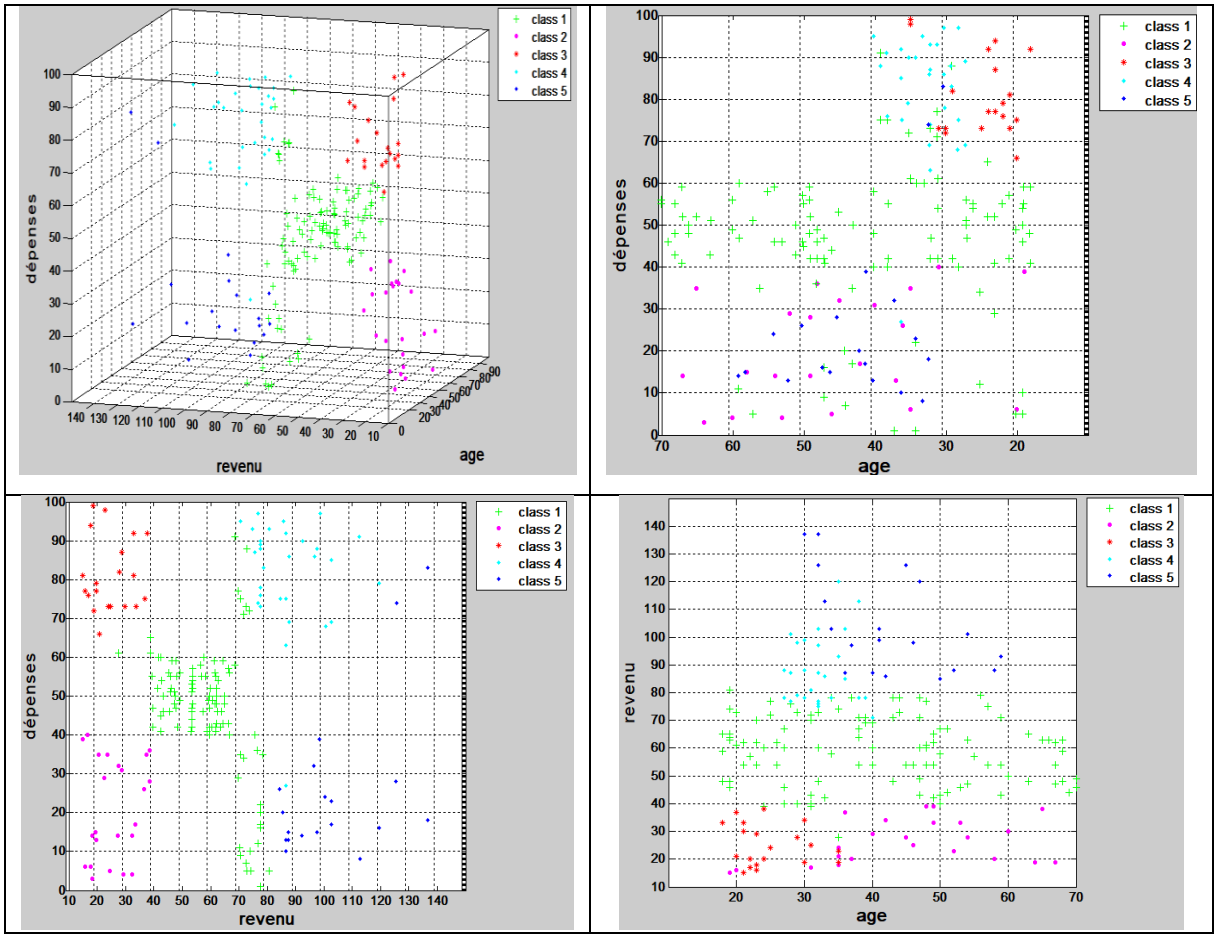


Figure 4.25. Représentation graphique en 3D de clustering du jeu de données mall Customer

La figure 4.25 représente le clustering du jeu de données mall Customer en prenant en compte 3 dimensions (score de dépense, âge, revenu).

La catégorie de couleur magenta représente les clients de faible revenu ayant un score de dépense faible. Leurs revenu annuel est inférieur à 40 k$ et leurs score de dépense est inférieur à 40.

L'âge de cette catégorie s'étale sur toute la tranche d'âge.

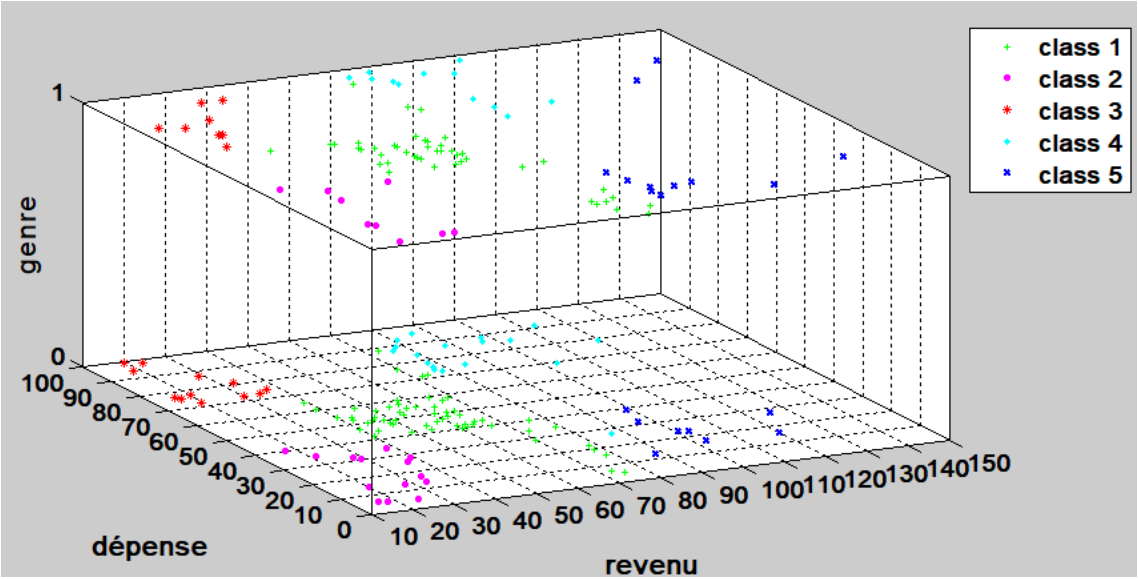


Figure 4.26. Représentation graphique en 3D de clustering du jeu de données Mall Customer (genre, dépense, revenu)

La figure 4.26 représente le clustering du jeu de données mall Customer en prenant en compte 3 dimensions selon le clustering d'âge (dépense, revenu, genre).

La catégorie de couleur bleue représente les clients ayant un revenu annuel élevé et un score de dépense faible.

Les clients ont un revenu entre 85 et 140 k$ et un score de dépense entre 10 et 40 k$.

La catégorie de couleur cyan représente les clients ayant un revenu annuel élevé et un score de dépense élevé.

Les hommes ont un revenu entre 70 et 100 k$ et les femmes ont un revenu entre 75 et 130 k$.

Les hommes ont un score de dépense entre 60 et 100 et les femmes ont un score de dépense entre 70 et 95.

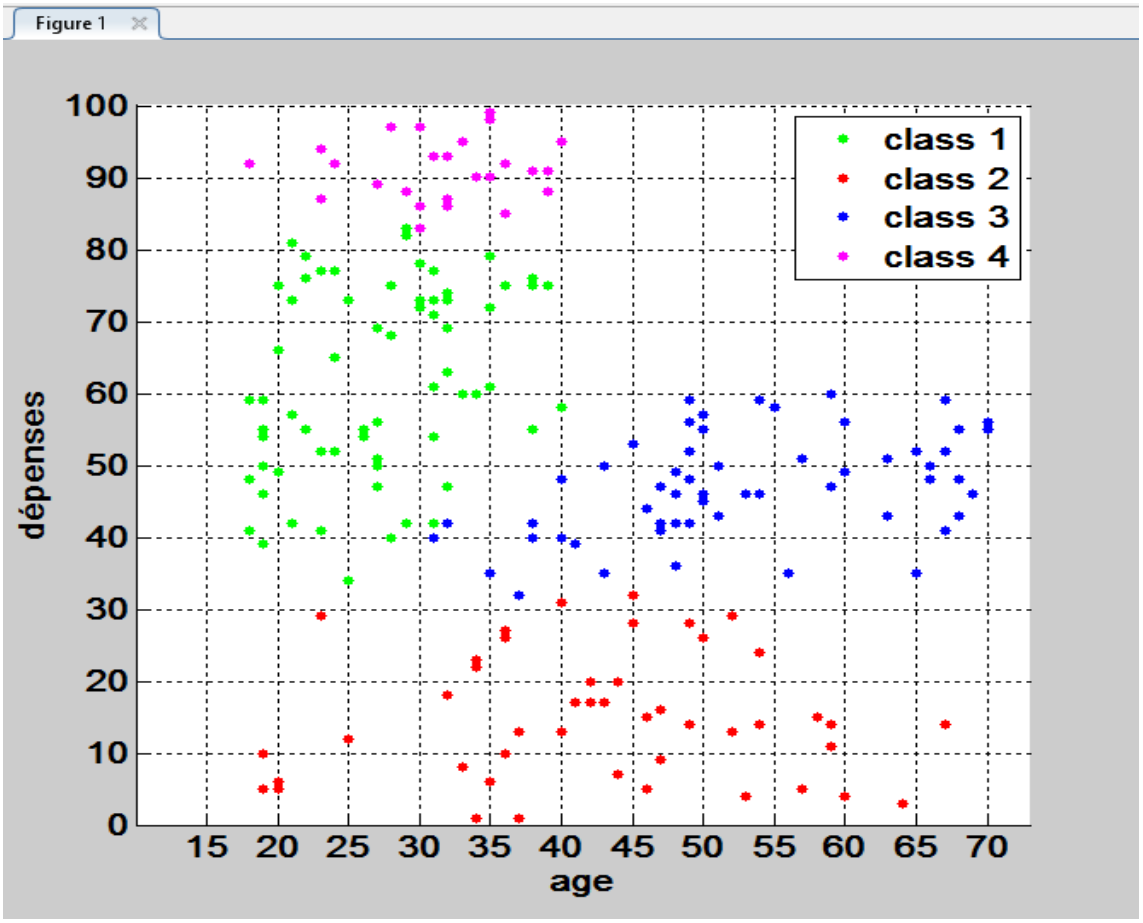


Figure 4.27. Représentation de clustering de deux dimensions (dépenses et âge)

La figure 4.27 représente le clustering du jeu de données mall Customer en prenant en compte 2 dimensions (le score de dépense et l'âge).

Les clients ayant un score de dépense élevé font partie de la tranche d'âge [20,40]. Plus de 40 ans, les clients ont soit un score moyen ou faible.

Moins de 30 ans, peu de clients qui ont un score faible, et nombreux sont qui ont un score moyen ou élevé.

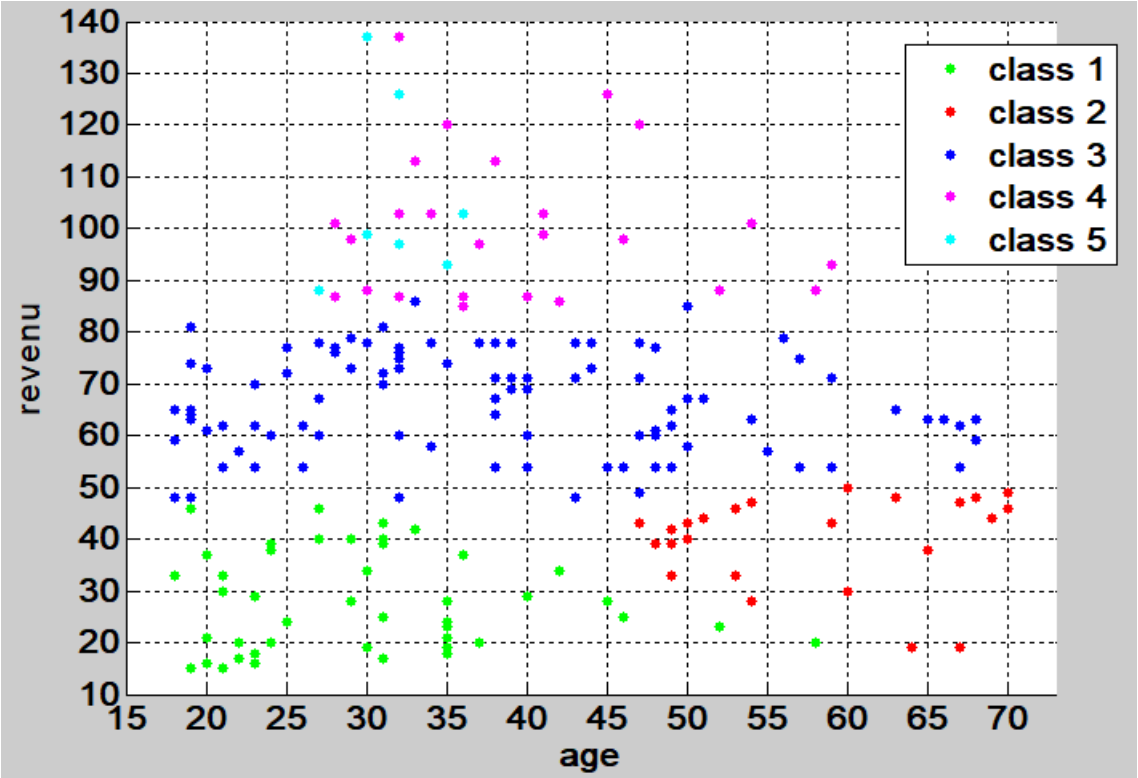


Figure 4.28. Représentation de clustering de deux dimensions (revenu et âge)

La figure 4.28 représente le clustering du jeu de données mall Customer en prenant en compte 2 dimensions (le revenu annuel et l'âge).

Uniquement pour les clients âgés entre 27 et 47, le revenu annuel peut dépasser 100k$.

Les clients âgés plus de 60 ans ont un revenu annuel inférieur à 65 k$.

4.4. Conclusion

Dans ce chapitre, nous avons présenté l'implémentation de notre application, nous avons montré le déroulement de notre application, les outils pour le développement, ainsi que la présentation détaillée des interfaces et la représentation graphique des données et leur analyse.

Conclusion Générale

Conclusion général :

Au cours de ce travail de master, nous nous sommes intéressés au problème de la catégorisation de clients, plus précisément, nous avons présenté une application de la méthode de la recherche de coucou, au clustering des clients en utilisant des données d'un mall.

Dans ce mémoire, nous avons commencé par présenter les notions et les concepts de base du domaine de clustering de données, tels que les définitions de clustering et de cluster, les techniques principales et techniques de validation. Ensuite nous avons passé à étudier les concepts, les principes et le comportement de la méthode de recherche de coucou ainsi que la description détaillée de son algorithme.

Ensuite, nous avons passé à l'adaptation de la méthode de la recherche de coucou à la catégorisation des données d'un mall en menant une série d'expérimentations, d'évaluation , d'analyse et d'interprétations basée sur des mesures internes et des représentations graphiques dans des espaces de 2 et 3 dimensions.

Bibliographie

Bibliographie :

[1] : R. CHAFIKA, Le clustering des données : une nouvelle approche évolutionnaire quantique, Constantine, Département d'Informatique, 2006, p. 140.

[2]: B. Mirkin, Clustering for Data Mining: A data Recovery Approach, National Research University Higher School of Economics, 2005

[3]: A. K. Jain et R.C. Dubes: "Algorithms for Clustering Data", Prentice Hall series de reference avancée, 1988.

[4] : Laetitia Jourdan, thèse de doctorat : "Métaheuristique pour l'extraction de connaissances application à la génomique", Novembre 2003.

[5] : R. CHAFIKA, Apports du Calcul Quantique et des Concepts Possibilistes à la Classification non supervisée Evolutionnaire, Constantine, Département d'Informatique Fondamentale et ses Applications, 2013, p. 160.]

[6]: Minho Kim and R. S. Ramakrishna. New indices for cluster validity assessment. Pattern Recognition Letters, 26(15):2353–2363, 2005.

[7]: Johann M. Kraus, Christoph Mssel, Gnther Palm, and Hans A. Kestler. Multiobjective selection for collecting cluster alternatives. Computational Statistics, 26:341–353, 2011.

[8]: https://scikit-learn.org/stable/modules/clustering.html

[9]: https://www.kdnuggets.com/2019/09/hierarchical-clustering.html

[10]: D.P.Mercer, linacre college : " Clustering large Datasets ", Octobre 2003

[11]: J.Bilmes, A.Vahdat, W.Hsu et E.J.Im. "Empirical observations of probabilistic heuristics for the clustering problem". Technical Report TR-97-018, International Computer Science Institute, University of California, Berkeley, CA, 1997

[12]: Z. Michalewicz, Genetic Algorithms + Data Structures" Evolution Programs, Springer, New York, 1992

[13]: R. CHAFIKA, Apports du Calcul Quantique et des Concepts Possibilistes à la Classification non supervisée Evolutionnaire, Constantine, Département d'Informatique Fondamentale et ses Applications, 2013, p. 160.

Bibliographie

[14] : H. S. Daoudi Meroua, Contribution au clustering de données génomiques: Approche parallèle et distribuée basée MapReduce – HADOOP, Constantine, Département de l'Informatique Fondamentale et ses Applications, 2014, p. 91

[15]: M. Halkidi, Y. Batistakis, M. Vazirgiannis, "On Clustering Validation Techniques", Intelligent Information Systems Journal, Kluwer Pulishers, vol.17, n°2-3, pp. 107-145, 2001.

[16]: M. Halkidi, Y. Batistakis, M. Vazirgiannis. "Cluster Validity Methods: Part II", SIGMOD Record, 2002.

[17]: J Handl, Ant-based methods for tasks of clustering and topographic mapping: extensions, analysis and comparison with alternative techniques. Masters Thesis, Université d' Erlangen-Nurnberg, Erlangen, Germany, 2003.

[18]: M. Halkidi ,M.Vazirgiannis. "Clustering Validity Assessment: Finding the optimal partitioning of a data set ",in the Proceedings of ICDM Conference ,California, USA, 2001.

[19]: W.Rand.Objective criteria for the evaluation of clustering methods. Journal of the American Statistical Association, Vol 66 n336, pp 846-850, 1971.

[20]: L.Hubert et P.Arabie. Comparing partitions. Journal of classification, Vol 2 ,pp 193-218, 1985.

[21]: S.Lioyd.Least squares quantization in PCM.IEEE Transactions on Information Theory,Vol 28 n2,pp 127-138, 1982.

[22]: M. Halkidi, M. Vazirgiannis, I. Batistakis. "Quality scheme assessment in the clustering process", In the Proceedings of the Fourth European Conference on Principles of data mining and Knowlegde discovery, Vol1910 of LNCS, pp 265-267.Springer-Verlag, 2000.

[23]: J.C. Dunn, A fuzzy relative of the ISODATA process and its use in detecting compact, well-separated clusters, J. Cybern. 3, pp 32–57, 1974.

[24]: Davies, D.L. and Bouldin, D.W. A Cluster Separation Measure. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1(2), 224–227, 1979

[25]: J.Handl et J.Knowles. "Exploiting the trade-off -- the benefits of multiple objectives in data clustering". In the Proceedings of the Third International Conference on Evolutionary Multi-Criterion Optimization, pp 547-560, LNCS 3410, 2005.

[26]: D. Attenborough. The Life of Birds, Parenthood.

http://www.pbs.org/lifeofbirds/home/index.html. (Récupéré le 10.05.2012).

Bibliographie

[27]: N. Bouglouan. http://www.oiseaux-birds.com. (récupéré le 10.05.2012)

[28]: P. B. Robert. The Cuckoos. Oxford University Press. 2005

[29]:S. Adams. Cuckoo chicks dupe foster parents from the moment they hatch. http://www.telegraph.co.uk/earth/wildlife/4109282/Cuckoo-chicks-dupefoster-parents-fromthe-moment-they-hatch.html. (Récupéré le 10.05.2012).

[30]: N.A. Campbell. Fixed action patterns. In: Biology, fourth ed., Benjamin Cummings, ISBN 0-8053-1957-3. New York, 1996.

[31]: R. Rajabioun. Cuckoo Optimization Algorithm. Applied Soft Computing. Vol. 11, N° 8, pp. 5508-5518, 2011.

[32]: L.H. Tein, R. Ramli. Recent advancements of nurse scheduling models and a potential path. In Proc. 6th IMT-GT Conference on Mathematics, Statistics and its Applications (ICMSA 2010). pp. 395-409, 2010.

[33]: A. Layeb. A novel quantum inspired cuckoo search for knapsack problems. Int. J. Bio-Inspired Computation. Vol. 3, N°. 5, pp 297-305, 2011.

[34]: A. Gherboudj, A. Layeb, S. Chikhi. Solving 0-1 knapsack problems by a discrete binary version of cuckoo search algorithm. International Journal of BioInspired Computation. Vol. 4, N°.4, pp 229-236. Inderscience Publishers ISSN (Print): 17580366, ISSN (online): 1758-0374. DOI: 10.1504/IJBIC.2012.048063. 2012.

[35]: Bak, P: ″How nature works″. Oxford universitypress Oxford,1997.

[36]: Yang X. S. et Deb, S : " Engineering optimisation by cuckoosearch".

International Journal of MathematicalModelling and Numerical Optimisation, 1(4):330{343, 2010.

[37]: Reynolds, A. M. et Frye, M. A. : ″ Free-flight odortracking in drosophilais consistent with an optimal intermittent scale-free search″. PloS one, 2(4):e354, 2007.

[38] : Aziz OUAARAB :Thèse de Doctorat ″Résolution de Problèmes d'Optimisation Combinatoire par des Métaheuristiques Inspirées de la Nature : Recherche du Coucou via les Vols de Lévy″ ,2015.

[39]: X-S Yang, S. Deb. Engineering optimisation by cuckoo search. Int. J. Mathematical Modelling and Numerical Optimisation. Vol. 1, N° 4, pp. 330-343, 2010.

[40]: Xin-She Yang :″CuckooSearch and FireflyAlgorithm″: Theory and Applications; Studies in ComputationalIntelligence ,Vol. 516 Springer International Publishing; 2014.

Bibliographie

[41]: Xin-She Yang :″ Cuckoo Search and Firefly Algorithm″: Theory and Applications; Studies in Computational Intelligence ,Vol. 516 Springer International Publishing; 2014.

[42]: Ishak B.S, Kamel.N et Bendjeghada.O : Article " A New Algorithm for Data Clustering Based on Cuckoo Search Optimization " Advances in Intelligent Systems and Computing 238, Springer International Publishing Switzerland 2014.