الجمهورية الجزائرية الديمقراطية الشعبية

وزارة التعليم العالي والبحث العلمي

UNIVERSITE BADJI MOKHTAR - ANNABA
BADJI MOKHTAR – ANNABA UNIVERSITY

جامعة باجي مختار – عنابـــة

Faculté : Des Sciences

Département : De Mathématiques.

Domaine : Maths-Informatique

Filière : Mathématiques

Spécialité : Probabilités-Statistiques

Mémoire

Présenté en vue de l'obtention du Diplôme de Master

Thème :

Implémentation de l'algorithme Fuzzy C-Means sur des ensembles de données réelles

Présenté par : *Krim Amina*

Jury de Soutenance :

Dr. Chadli A.	Prof	U. BADJI Mokhtar Annaba	Président
Dr. Amira O.	MCB	U. BADJI Mokhtar Annaba	Encadrant
Dr. Talhi H.	MCA	U. BADJI Mokhtar Annaba	Examinateur
Dr. Chouia S.	MCA	U. BADJI Mokhtar Annaba	Examinateur
Dr. Redjil A.	MCB	U. BADJI Mokhtar Annaba	Examinateur

Année Universitaire : 2022/2023

Remerciements

En tout premier lieu, je remercie **ALLAH**, tout puissant, de m'avoir donné la force, ainsi que l'audace pour dépasser toutes les difficultés et me permis de mener à bien ce travail.

Je tiens à remercier tout particulièrement mon encadreur **Dr. Amira Ouafa** pour l'honneur qu'elle me fait en m'encadrant et de la confiance qu'elle m'a accordée en acceptant de diriger ce travail. Je la remercie pour ses conseils et pour tout le temps qu'elle a consacré à ce mémoire.

J'exprime mes plus sincères remerciements à madame **Pr. Chadli**, je l'assure ma profonde reconnaissance d'avoir acceptée présider mon jury.

Je tient à exprimer ma gratitude et ma haute considération aux **Dr. Talhi**, **Dr. Chouia** et **Dr. Redjil**, de m'avoir honorée par leur présences dans mon jury et pour l'intérêt qu'elles ont portées à ce travail.

Je remercie également les professeurs de l'université Badji Mokhtar Annaba et précisément les enseignants du département des mathématiques.

Dédicace

بسم الله الرحمن الرحيم

(قُلْ إِنَّ صَلاتِي وَنُسُكِي وَمَحْيَايَ وَمَمَاتِي لِلَّهِ رَبِّ الْعَالَمِينَ * لاَ شَرِيكَ لَهُ وَبِذَلِكَ أُمِرْتُ وَأَنَاْ أَوَّلُ الْمُسْلِمِينَ)

سورة الأنعام {162-163}

***Je** dédie ce travail*

***A** mes très chers parents : Krim Mohammed et Sebbar Fahima, sources de vie, d'amour, et d'affection, que Dieu les protèges.*

***A** mes chers frères Abdenacer et Khaled, Sources de joie et de bonheur.*

A mes chères sœurs Samira (et ses filles), Hanene, et Imene sources de joie et de bonheur.

A mon cher mari Chakour Chouaib, source d'amour et de joie et de bonheur, que Dieu le protège.

A mes très chers filles Meriem et Janna, sources d'amour, de fierté, de joie et de bonheur, que Dieu les protèges.

***A** toute mon adorable famille*

ملخص:

التنقيب في البيانات هو عملية تهدف إلى استخراج معلومات ذات دلالة من البيانات المتاحة. يركز هذا العمل على التنقيب في البيانات لأغراض التجميع باستخدام آليات التجميع الضبابية. الهدف هو تحديد وتصور مجموعات من العناصر المتشابهة بناءً على معايير محددة. تشمل طرق تجميع البيانات المدروسة في هذا الأطروحة طريقة K-means وطريقة C-means الضبابية. تُستخدم طريقة C- means الضبابية على وجه التحديد لتجزئة ضبابية للصور بالدرجات الرمادية. على عكس طريقة K-means التي تُعين بشكل صارم كل نقطة بيانات إلى مجموعة واحدة فقط، تسمح طريقة C-means الضبابية بوجود درجة معينة من "الضبابية" من خلال تعيين درجات انتماء جزئية لكل نقطة لكل مجموعة. يتيح ذلك تمثيلًا أدق للانتقالات بين الكائنات في الصورة وتحسين الجودة العامة للتجزئة. كما يتم تسليط الضوء على مزايا وعيوب هذه الطريقة.

الكلمات المفتاحية:

التجميع، خوارزمية k-Means، طريقة FCM، معالجة الصور

Abstract

Data mining is a process aimed at extracting meaningful information from available data. This work focuses on data mining for clustering purposes using fuzzy clustering mechanisms. The objective is to identify and visualize sets of similar elements based on defined criteria. The data clustering methods examined in this dissertation include the K-means method and the fuzzy C-means method. The fuzzy C-means method is specifically used for fuzzy segmentation of grayscale images. Unlike the K-means method, which rigidly assigns each data point to a single cluster, the fuzzy C-means method allows for a certain degree of "fuzziness" by assigning partial membership degrees to each point for each cluster. This enables a finer representation of transitions between objects in an image and enhances the overall quality of segmentation. The advantages and disadvantages of this method are also highlighted.

Keywords : Clustering ; K-means ; FCM algorithm, Image Processing.

Résumé

L'exploration de données est un processus qui vise à extraire des informations fines à partir des données disponibles. Ce travail porte sur l'exploration de données à des fins de clustering en utilisant des mécanismes de regroupement flous. L'objectif est d'identifier et visualiser des ensembles d'éléments similaires en fonction de critères définis. Les méthodes de regroupement de données examinées dans ce mémoire comprennent la méthode des K-means et la méthode de C-means floue. La méthode de C-means floue est spécifiquement utilisée pour la segmentation floue des images en niveaux de gris. Contrairement à la méthode des K-means qui assigne de manière rigide chaque point de données à un seul cluster, la méthode de C-means floue permet une certaine "floueté" en attribuant des degrés d'appartenance partiels à chaque point pour chaque cluster. Cela permet une représentation plus fine des transitions entre les objets dans une image et améliore la qualité globale de la segmentation. Les avantages et les inconvénients de cette méthode sont également mises en évidence.

Mots-clés : Clustering ; K-moyenne ; FCM algorithme, Traitement d'image.

Table des figures

1.1 Intelligence Artificiel. 3
1.2 Apprentissage non supervisé vs apprentissage supervisé. 4
1.3 Représentation d'une image numérique dans le plan cartésien. 6
1.4 Caractéristique d'une image. 7
1.5 RGB au niveau de gris. 7
1.6 Codage des couleurs. 8

2.1 Idée de base du clustring. 9
2.2 Distance intra-cluster et inter-cluster. 10
2.3 Architecture Générale de Clustering. 11
2.4 Clustering basée sur le centroïde. 13
2.5 Clustering hiérarchique : agglomératif et divisif. 13
2.6 Clustering basé sur le Modèle de Mélange Gaussien. 13
2.7 Clustering basé sur la densité. 14
2.8 Exemple de HMEANS clustring (ISODATA). 17
2.9 Étapes de K-means clustring. 19

3.1 Degrés d'appartenance des ensembles flous. 22

4.1 La structure MATLAB 30
4.2 Environement Matlab 31
4.3 Analyse de l'image 33
4.4 Exemple d'image en couleur segmentée 33
4.5 Image en niveaux de gris. 35
4.6 Image segmentée par FCM (c = 3). 35
4.7 Fonctions d'appartenances pour c = 3. 36
4.8 Image segmentée par FCM (c = 4). 37
4.9 Fonctions d'appartenances pour c = 4. 38
4.10 Image en niveaux de gris. 39
4.11 Image segmentée par FCM (c = 2). 39
4.12 Fonctions d'appartenances pour c = 2. 40
4.13 Image segmentée par FCM (c = 3). 41
4.14 Fonctions d'appartenances pour c = 3. 41

Table des matières

1 Technologie et tendances **3**
- 1.1 Intelligence Artificielle . . . 3
  - 1.1.1 Apprentissage Automatique (Machine Learning) . . . 3
  - 1.1.2 Reconnaissance de formes . . . 4
  - 1.1.3 Les données . . . 4
- 1.2 Traitement d'image . . . 5
  - 1.2.1 Définition de l'image . . . 5
  - 1.2.2 Image numérique . . . 5
  - 1.2.3 Rendu des couleurs par le codage RGB/Niveaux de gris . . . 6
  - 1.2.4 Type des images . . . 7

2 Principe de base du Clustering de Données **9**
- 2.1 Introduction . . . 9
- 2.2 Qu'est-ce que le clustering ? . . . 9
- 2.3 Comment fonctionnent les méthodes de clustering ? . . . 10
- 2.4 Qualité d'une méthode de clustering . . . 12
- 2.5 Types de Clustering . . . 12
- 2.6 Domaines d'application du clustring . . . 12
- 2.7 Avantages des algorithmes de clustering . . . 15
- 2.8 Limitations des algorithmes du clustring . . . 15
- 2.9 Algorithme de K-Means . . . 16
  - 2.9.1 Principe du clustering . . . 16
  - 2.9.2 Fonctionnement de l'algorithme K-Means . . . 17
- 2.10 Avantages et Inconvénients du K-means . . . 18
  - 2.10.1 Avantages . . . 18
  - 2.10.2 Inconvénients . . . 19
- 2.11 Conclusion . . . 20

3 La méthode Fuzzy C-Means (FCM) **21**
- 3.1 Introduction . . . 21
- 3.2 La logique floue . . . 21
  - 3.2.1 Remarques . . . 21
  - 3.2.2 Théorie des ensembles flous . . . 22
  - 3.2.3 Avantages et Inconvénients des systèmes flous . . . 23
- 3.3 Fuzzy C-Means Clustering . . . 24
  - 3.3.1 Modèle d'optimisation FCM . . . 24
  - 3.3.2 Conditions pour l'optimalité . . . 25
  - 3.3.3 Algorithme FCM . . . 26
  - 3.3.4 Notes sur le FCM clustering . . . 26
  - 3.3.5 Les avantages et les inconvénients du FCM . . . 27

3.4 Conclusion 28

4 Réalisation et expérimentation 29
4.1 Introduction 29
4.2 Le Software (Matlab) 29
4.3 Différente Fonctionnalités de Matlab 30
4.3.1 Développement d'algorithmes et d'applications 30
4.3.2 Accès aux données et analyse 31
4.3.3 Visualisation de données 32
4.3.4 Calcul numérique 32
4.4 Segmentation d'image 33
4.4.1 Définition de la segmentation 33
4.4.2 Types d'approches pour la segmentation 34
4.4.3 Le choix d'une technique de segmentation 34
4.4.4 Objectifs de la segmentation 34
4.4.5 Implémentation et résultats 34
4.5 Conclusion 42

Bibliographie 44

Introduction

Le développement des technologies de l'information a entraîné une croissance phénoménale des bases de données et d'énormes ensembles de données dans divers domaines, et les gens tombent progressivement dans la situation de "données riches et connaissances pauvres". Ce qui rend l'information un élément très présent autour de nous et depuis la venue de l'informatique, non seulement le volume des données stockées sous forme numérique est en croissance, mais aussi le type de ces données est devenu très varié. Le domaine de la fouille de données se focalise sur le traitement de ces données afin d'extraire et de mettre en évidence des informations utiles [1]. Il existe de nombreuses approches de traitement de données permettant une restitution fine de l'information extraite des données. Cependant, la complexité des résultats produits peut être un réel obstacle pour l'interprétation. Il est alors nécessaire que l'utilisateur soit guidé afin d'exploiter les résultats obtenus. Dans ce cas, la mise en place d'outils d'analyse exploratoire s'avère nécessaire.

Pour mieux comprendre le monde réel, les humains doivent faire la distinction entre différents objets et percevoir les similitudes entre les choses. Le clustering est le processus de distinction et de classification en fonction des similitudes entre les objets. L'analyse de clustering fait appel à des méthodes mathématiques pour étudier et traiter la classification d'un objet particulier. Le clustering Fuzzy C-Means (FCM) [2] [3], il est considéré comme l'une des méthodes les plus populaires et les plus utilisées en exploration de données. Il s'agit d'une extension floue des k-means conventionnels [4] et a pris plusieurs noms avant d'être appelé FCM tels que les isodata floues et les k-means flous. La formulation spécifique du FCM a été établie par Dunn [2], mais sa généralisation et son cadre actuel sont attribués à Bezdek [3]. La fonction objectif du FCM et de ses généralisations est un modèle de techniques de clustering floues qui suscite un très grand intérêt de la part des chercheurs dans la reconnaissance de modèles.

La méthode de clustering FCM permet de donner une partition floue à l'ensemble de données d'entrée. Cependant, il existe une infinité de partitions potentiellement floues. Cette méthode cherche à minimiser la variance intra-cluster en utilisant une fonction objectif donnée. Pour y parvenir, une fonction d'amélioration ou une fonction objectif doit être conçue pour rechercher la partition optimale en fonction de la fonction objectif choisie. Fuzzy C-Means est une technique de clustering basée sur une fonction objectif, dont l'optimisation est résolue par l'algorithme itératif localement optimal appelé FCM. Plusieurs techniques ont été proposées dans la littérature pour résoudre la fonction objectif du Fuzzy C-Means. Parmi ces techniques, on peut citer l'optimisation par recuit simulé décrite dans [5], l'utilisation d'une technique d'optimisation combinatoire appelée Tabu Search [6], une reformulation de la fonction objectif et des méthodes d'optimisation générales proposées par Hathaway et Bezdek dans [7]. Une approche d'optimisation génétique a également été utilisée avec succès par Hall et al. [8], tandis qu'une nouvelle implémentation de l'approximation et du clustering de fonction a été proposée dans [9] comme un cadre d'optimisation alternatif pour résoudre la fonction objectif.

La segmentation d'image joue un rôle essentiel et crucial lors de l'analyse d'images. En effet, une segmentation d'image de qualité permet d'obtenir une analyse précise, car elle per-

met d'extraire des paramètres caractéristiques utilisés pour la classification et l'interprétation. L'objectif de la segmentation est de créer une représentation concise et significative du contenu informatif de l'image, qui peut être utilisée de multiples façons.

Le contenu principal de notre mémoire est organisé comme suit :

Le premier chapitre commence par traiter des origines de l'intelligence artificielle et de son impact sur la gestion des données et notamment qui sont en relation avec le domaine de traitement d'image.

Le deuxième chapitre aborde les techniques de traitement de données par classification à caractère non supervisé, en mettant l'accent sur la méthode K-means et ses applications.

Le troisième chapitre présente l'idée fondamentale du clustering qui est en relation avec la méthode Fuzzy C-Means, dont les concepts clés de la logique floue, de la théorie des ensembles flous et des fonctions d'appartenance sont discutés.

Le quatrième chapitre concerne la mise en application de la méthode FCM pour la segmentation des images en niveau de gris sur l'environnement Software Matlab.

Chapitre 1

Technologie et tendances

1.1 Intelligence Artificielle

La terminologie de l'intelligence Artificielle (IA) est née dans un atelier en 1956, alors que sa définition vient depuis 1950, où Turing a introduit un test, pour tester l'intelligence d'une machine. Dans les années 90 et au début du 21e siècle, les ordinateurs ont commencé à disposer suffisamment de puissance de calcul et de données pour que l'IA puisse être utilisée dans la santé et la pharmacie, l'exploration de données, la logistique et d'autres domaines (Figure 1.1).

La classification de l'IA peut être regroupée en fonction de la proximité avec laquelle l'IA peut imiter la pensée humaine. Ainsi, nous pouvons différencier la façon dont le système se compare aux humains en termes de polyvalence et de performances. Sur la base de ces critères, nous classons les systèmes d'IA en fonction de leur similitude avec l'esprit humain et de leur capacité à penser.

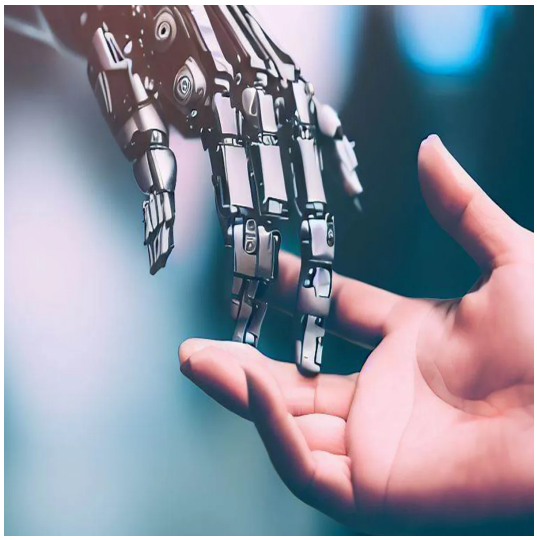

Figure 1.1 – Intelligence Artificiel.

1.1.1 Apprentissage Automatique (Machine Learning)

L'apprentissage artificiel ou automatique est devenu une branche majeure des mathématiques appliquées, à l'intersection des statistiques et de l'intelligence artificielle. Son objectif

est de réaliser des modèles qui apprennent « par l'exemple » : il s'appuie sur des données numériques (résultats de mesures ou de simulations), contrairement aux modèles « de connaissances » qui s'appuient sur des équations issues des premiers principes de la physique, de la chimie, de la biologie, et de l'économie, etc. L'apprentissage statistique est d'une grande utilité lorsque l'on cherche à modéliser des processus complexes, souvent non linéaires, pour lesquels les connaissances théoriques sont trop imprécises pour permettre des prédictions précises. Ses domaines d'applications sont multiples : fouille de données, bio-informatique, génie des procédés, aide au diagnostic médical, télécommunications, vision par ordinateur, interface cerveau-machines, et bien d'autres. les principaux types d'apprentissage automatique sont : l'apprentissage supervisé et l'apprentissage non supervisé (voir Figure 1.2). L'apprentissage automatique supervisé repose sur des données de formation d'entrée et de sortie étiquetées, tandis que l'apprentissage non supervisé traite des données non étiquetées ou brutes.

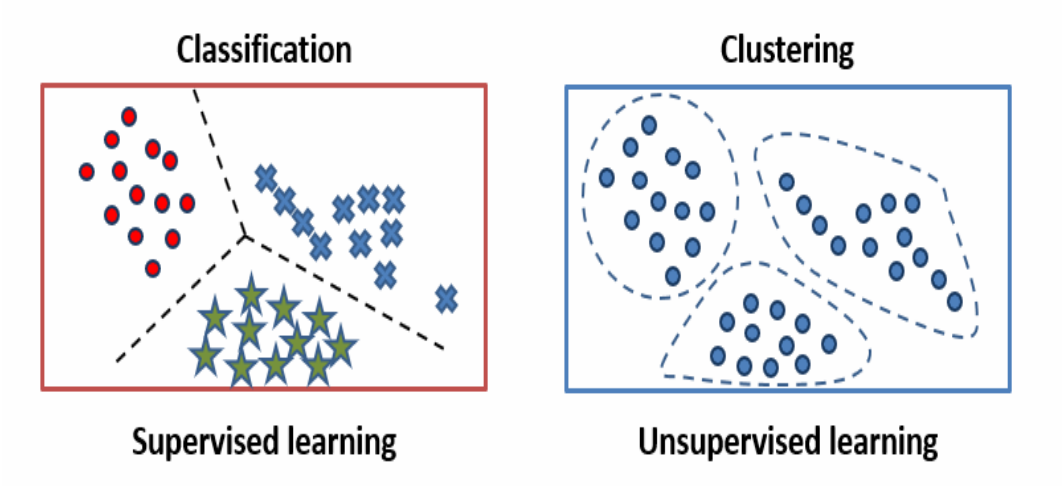

FIGURE 1.2 – Apprentissage non supervisé vs apprentissage supervisé.

1.1.2 Reconnaissance de formes

La reconnaissance de modèles est une méthode d'analyse de données qui utilise des algorithmes d'apprentissage automatique pour reconnaître automatiquement les modèles et les régularités dans les données. D'une manière simple, cela peut être défini comme la recherche de structure dans les données. Ces données peuvent être n'importe quoi, du texte et des images aux sons ou à d'autres qualités définissables. Les systémes de reconnaissance de formes peuvent reconnaître rapidement et avec précision des formes familières. Il a des applications dans l'analyse de données statistiques, la compression de données, l'infographie, le traitement du signal, l'analyse d'images, la recherche d'informations, la bioinformatique.

1.1.3 Les données

Il existe différents types de données telles que qualitatives, quantitatives, numériques, picturales, texturales et linguistiques ou, dans certains cas, peuvent être des combinaisons

indifférentes de celles-ci. Des exemples de sources de données sont les dossiers médicaux, les photos aériennes, les tendances du marché, les catalogues de bibliothèques, les positions galactiques, les empreintes digitales, les profils psychologiques, les flux de trésorerie, les constituants chimiques, les caractéristiques démographiques, les options d'achat d'actions, les décisions militaires. La technique ou la méthode du modèle de recherche est applicable à n'importe lequel de ces types et sources de données. L'option de recherche est utilisée pour connaître les techniques de traitement des données.

1.2 Traitement d'image

De nos jours, l'utilisation de l'image est l'un des moyens de communication les plus essentiels pour les individus. Elle représente un langage universel dont la richesse permet une compréhension mutuelle entre personnes de tous âges et de toutes cultures. De plus, l'image est le moyen le plus efficace de communication, car chacun peut l'analyser à sa manière, en déduire une impression et en extraire des informations précises.

Le terme **technique de traitement d'images** englobe toutes les méthodes visant à modifier les caractéristiques chromatiques des pixels des images bitmap. Le traitement d'images est souvent associé à l'amélioration visuelle des images, dans le but d'optimiser leur apparence et d'extraire des informations pertinentes qui seront utilisées dans diverses applications telles que la reconnaissance ou la segmentation.

1.2.1 Définition de l'image

L'image est une représentation visuelle d'une personne ou d'un objet créé à travers des médiums tels que la peinture, la sculpture, le dessin, la photographie, le film, et ainsi de suite. Elle constitue également un ensemble organisé d'informations qui, une fois affichées sur un écran, ont une signification pour l'oeil humain.

L'image est donc constituée par un ensemble régulier d'éléments appelés « pixels » (contraction du terme anglo-saxon « picture elements ») et est elle-même généralement appelée image « bitmap » (contraction du terme anglo-saxon « bits mapped »).

Mathématiquement parlant, une image peut être considérée comme une fonction à deux variables. Par exemple, nous pouvons représenter une image comme $I(x, y)$, où I est une fonction d'amplitude (ou d'intensité) dépendant de deux variables réelles de position (x, y) dans le plan cartésien (voir Figure 1.3).

1.2.2 Image numérique

Le terme "image numérique" fait référence, dans son sens le plus général, à toute image qui a été acquise, traitée et enregistrée sous une forme codée représentable par des nombres (valeurs numériques).

Une image numérique couleur est un fichier informatique, qui peut être ouvert avec un programme de visualisation d'images. Une fois l'image ouverte en taille réelle, elle se présente sous la forme d'un rectangle constitué par un ensemble de points colorés, les pixels.

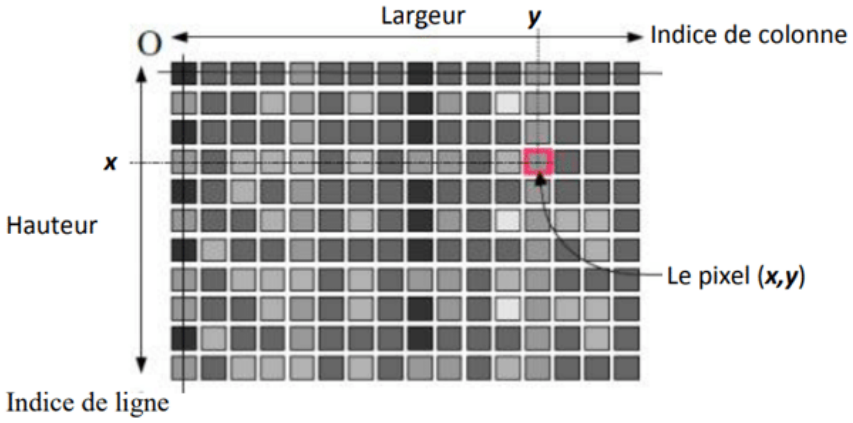

FIGURE 1.3 – Représentation d'une image numérique dans le plan cartésien.

On parle d'image matricielle pour désigner une image numérisée « classique ». En effet, en Mathématiques, une matrice désigne un tableau de nombres : chaque élément de la matrice (nombre / groupe de nombres) est associé à un pixel de l'image. La définition de l'image est le nombre de pixels qui la composent. Elle est donnée en indiquant le nombre de pixels sur une ligne, et le nombre sur une colonne.

La numérisation est le processus permettant de passer d'une image physique (par exemple, une image optique) caractérisée par un signal continu (avec une infinité de valeurs possibles, comme l'intensité lumineuse), à une image numérique caractérisée par un aspect discret (où l'intensité lumineuse ne peut prendre que des valeurs quantifiées en un nombre fini de points distincts). C'est cette forme numérique qui permet une exploitation ultérieure par des outils logiciels sur un ordinateur.

1.2.3 Rendu des couleurs par le codage RGB/Niveaux de gris

L'œil perçoit les couleurs via des cellules photoréceptrices du nom de cônes. Il en existe trois types chez l'Homme : certains sont sensibles au rouge, d'autres au vert, et d'autres encore au bleu. L'oeil perçoit ainsi toute couleur comme une combinaison des 3 couleurs primaires rouge, vert, bleu (voir Figure 1.4).

Pour le rendu des couleurs, chaque pixel est composé de 3 pastilles (aussi nommées sous-pixels), une rouge, une verte, et une bleue. En modulant la luminosité de chacune, le pixel peut reproduire toute couleur souhaitée (figure 1.5).

Dans le cadre du **codage RGB** 24 bits, l'image numérisée couleur obéit à ce principe-là. Elle peut être vue comme une matrice elle-même divisée en trois sous-matrices de même taille, une par couleur (rouge, vert, bleu). A chaque pixel, il correspond donc un triplet de trois nombres, codés en binaire. Chacun indique la luminosité du sous-pixel associé.

Il existe également des images **en niveaux de gris**. Dans ce cas là, il n'est plus nécessaire d'avoir les 3 sous-matrices du mode couleur 24 bits, une seule suffit. Pour un pixel donné, la valeur donnée par le tableau provoque l'allumage à l'identique des 3 sous-pixels. Cela induit

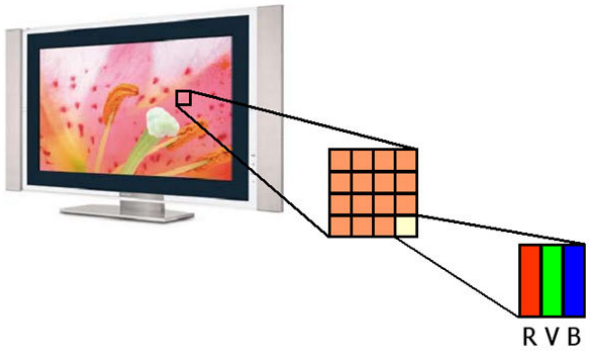

FIGURE 1.4 – Caractéristique d'une image.

un niveau de gris, ou éventuellement du blanc ou du noir. A qualité égale, la taille du fichier image en niveaux de gris est trois fois plus petite que celle de l'image couleur correspondante.

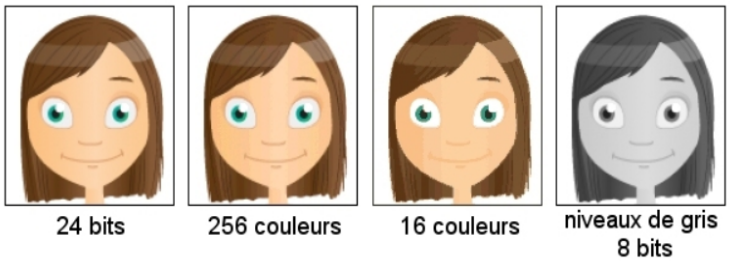

FIGURE 1.5 – RGB au niveau de gris.

1.2.4 Type des images

En informatique, il existe deux types d'images : les images vectorielles et les images bitmap.

1.2.4.1 Les images vectorielles

L'image vectorielle est dépourvue de matrice. Elle est créée à partir d'équations mathématiques, et chaque forme dépend des paramètres hauteur, largeur et rayon donnés à des vecteurs. À l'inverse de l'image matricielle composée de pixels, l'image vectorielle peut être redimensionnée sans pour autant perdre en qualité.

L'image vectorielle est surtout utilisée pour la création de cartes, cliparts ou encore pour des animations sur Internet. Elle est par contre peu appropriée pour réaliser des images photoréalistes, qui nécessitent plus de précision dans la nuance des couleurs.

Une image vectorielle est une image créée sur un ordinateur à partir de formules mathématiques. À la différence d'une image matricielle composée de pixels, une image vectorielle est faite de formes (polygones, lignes, ellipses) possédant divers paramètres tels que hauteur,

longueur, rayon, couleur, etc. Ces formes et lignes sont des combinaisons de courbes de Bézier composées de vecteurs. Chaque vecteur possède une norme, une direction et un sens.

L'avantage majeur d'une image vectorielle est sa capacité à être redimensionnée sans perte de qualité, les formules mathématiques qui la composent étant constamment recalculées. Une même image peut ainsi être utilisée de différentes façons, que ce soit pour une carte de visite, un tee-shirt ou un panneau publicitaire. Elle est aussi contenue dans un fichier beaucoup plus léger qu'une image pixelisée, indépendamment de sa taille et de sa résolution.

1.2.4.2 Une image matricielle (ou bitmap)

Une image matricielle, également appelée image bitmap, est composée d'une grille de points appelés pixels. Chaque pixel contient des informations sur sa position et sa couleur. Les photos numériques et les images scannées sont de ce type (voir Figure 1.6).

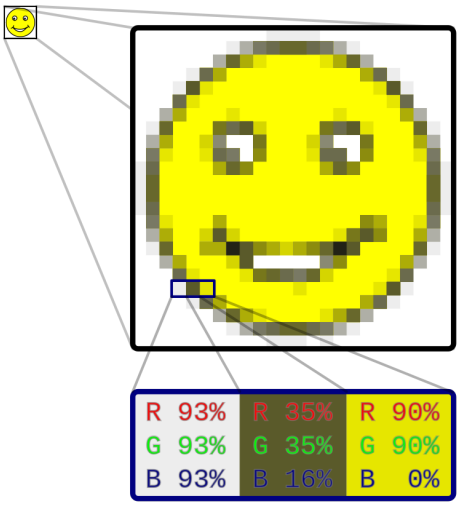

FIGURE 1.6 – Codage des couleurs.

Chapitre 2

Principe de base du Clustering de Données

2.1 Introduction

L'analyse en cluster peut se révéler puissante dans l'exploration de données. C'est une méthode statistique de traitement des données qui organise les éléments étudiés en groupes en fonction de leur degré de similitude (voir figure 2.1). Son objectif est d'identifier et visualiser des ensembles d'éléments similaires en fonction de critères définis. Le clustering sert principalement à segmenter ou classifier une base de données (par exemple trier des données clients type âge, profession exercée, lieu de résidence, etc.) ou extraire des connaissances pour tenter de relever des sous-ensembles de données difficiles à identifier à l'œil nu. L'amélioration des algorithmes de clustering ouvre alors de nouvelles perspectives en termes de partition et gestion des données.

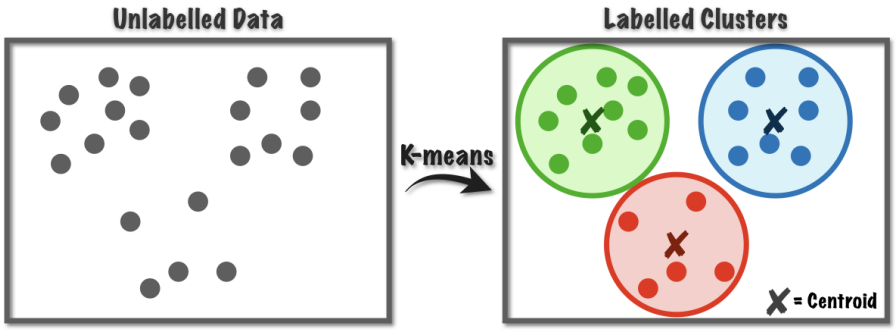

Figure 2.1 – Idée de base du clustring.

2.2 Qu'est-ce que le clustering ?

Le mot **clustering** signifie, en français, le **partitionnement** des données, le **regroupement** des données ou encore la **segmentation** d'une base de données. En anglais, on peut voir utilisée l'expressions **data clustering**.

Le clustering est une technique d'apprentissage automatique permettant de regrouper des chaînes de données par distance ou par similarité [10]. Il s'agit d'une méthode non supervisée et populaire pour une analyse des données. Ce qui signifie qu'il n'y a pas d'étiquettes

préexistantes et avec un minimum de supervision humaine. Il est alors possible d'appliquer des algorithmes de classification afin de gérer ces données individuelles dans chaque groupe spécifique.

Le clustering organise les données brutes en groups homogènes. Le problème de clustering consiste à trouver une structure de regroupement au sein d'un certain nombre d'objets, les objets de chaque cluster (groupe) doivent partager des caractéristiques communes. C'est un concept intuitif mais difficile à réaliser dans la pratique. L'outil de regroupement des données est un algorithme qui mesure la proximité entre chaque élément selon des critères prédéfinis. Pour hiérarchiser ou distribuer les données, l'algorithme tente de réduire la dissimilarité (variation) au sein d'un cluster (intra-cluster) et de l'augmenter entre les clusters (inter-cluster) (Figure 2.2). Puisque les clusters aident les gens à analyser et à décrire le monde, la recherche sur ce sujet a commencé à grande échelle. Elle doit alors être sélectionnée avec prudence selon le résultat attendu et la manipulation des données.

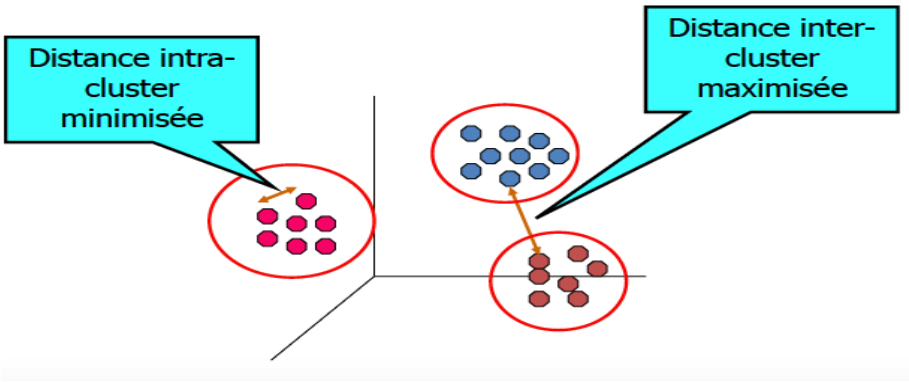

FIGURE 2.2 – Distance intra-cluster et inter-cluster.

En règle générale, les données collectées doivent être répertoriées comme une entrée de l'algorithme de clustering de calcul spécifié Figure 2.3. Ensuite, l'algorithme de clustering fournit une description de la structure détectée dans les objets, tels que des modèles cachés ou des ensembles de données, sans nécessiter d'intervention humaine. Cette description contient généralement une liste de chaque objet ainsi que le groupe qui lui est affecté. Sa capacité de savoir les similitudes et les dissemblances dans les informations en fait la solution idéale dans de nombreux domaines tels que l'analyse exploratoire de données, les stratégies de ventes croisées, la segmentation d'images, la reconnaissance d'images, etc.

2.3 Comment fonctionnent les méthodes de clustering ?

Les méthodes de clustering fonctionnent en regroupant des objets similaires en clusters ou groupes, tout en maximisant la différence entre ces groupes et en minimisant la différence à l'intérieur de chaque groupe. Il existe plusieurs méthodes de clustering, chacune utilisant une approche différente pour diviser les données en groupes. Les étapes de base de la plupart des méthodes de clustering sont ainsi :

1. **Sélection des variables pertinentes :**

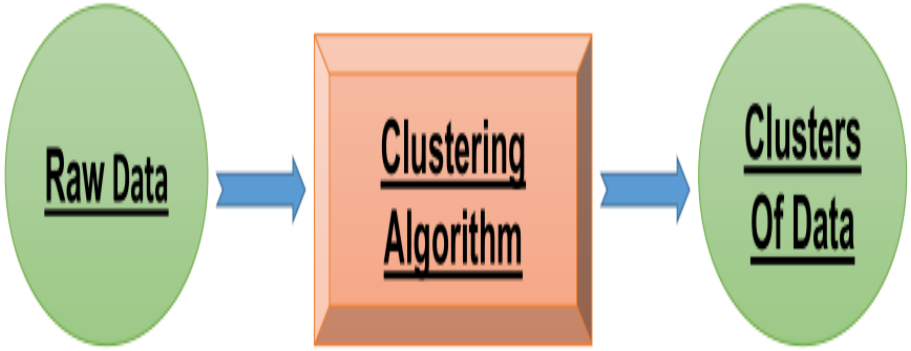

FIGURE 2.3 – Architecture Générale de Clustering.

Les variables à inclure dans l'analyse de clustering sont sélectionnées en fonction du contexte de l'analyse et des objectifs de l'analyse.

2. **Choix de la méthode de clustering :**
   Il existe plusieurs méthodes de clustering, notamment le clustering hiérarchique, le clustering k-means et le clustering basé sur la densité. Chaque méthode a ses avantages et ses inconvénients, et il est important de choisir la méthode appropriée pour les données et l'objectif de l'analyse.
3. **Calcul de la similarité :**
   Les données sont comparées les unes aux autres pour évaluer leur similarité. La mesure de similarité la plus couramment utilisée est la distance euclidienne, mais d'autres mesures peuvent également être utilisées en fonction des données et de la méthode de clustering.
4. **Formation des clusters :**
   Les objets similaires sont regroupés en clusters en fonction de leur similarité. Dans le clustering k-means, les clusters sont formés en définissant de manière aléatoire k centres initiaux et en attribuant chaque point de données au centre le plus proche, puis en calculant les nouveaux centres et il faut répéter le processus pour obtenir un bon résultat.
5. **Évaluation des clusters** :
   Les clusters sont évalués en fonction de leur homogénéité, et de leur séparation.
6. **Interprétation des résultats :**
   Les résultats du clustering sont interprétés en fonction de l'objectif de l'analyse et du contexte. Les résultats peuvent être utilisés pour prendre des décisions éclairées, pour identifier des modèles ou des relations cachées dans les données, pour segmenter les clients ou pour détecter des anomalies ou des valeurs aberrantes.

2.4 Qualité d'une méthode de clustering

Evaluer les performances d'un algorithme de clustering (classification non supervisée) n'est pas facile. Pour cela, quelques indicateurs de qualité communément utilisés dans la liturature.

— Une des questions difficiles du clustering : à quel point les clusters trouvés sont bons ?
— Capacité à traiter différents types d'attributs.
— Découverte de clusters de formes arbitraires.
— Connaissances minimales du domaines requises pour définir les paramètres.
— Capacité à traiter les données bruitées.
— Insensibilité à l'ordre des objets du jeu de données.
— Résultat interprétable et utilisable.

2.5 Types de Clustering

Dans cette partie, nous donnerons un aperçu rapide des principaux types de clustering. Ces types sont répertoriés comme suit :

1. **Méthode basée sur le centroïde :**
   Elle organise l'ensemble de données en clusters non hiérarchiques, et le centre est considéré comme un représentant de chaque cluster. L'affectation des objets aux clusters repose sur la proximité, c'est-à-dire la distance minimale entre les objets et les centres (Figure 2.4).
2. **Clustering hiérarchique (HC) :**
   Le clustering hiérarchique est une méthode de clustering qui regroupe les données en fonction de leur similarité en formant des groupes hiérarchiques ou des arbres. Cette méthode commence par former des clusters individuels pour chaque objet, puis fusionne progressivement les clusters en groupes plus larges en fonction de leur similarité. Cette méthode peut être divisée en deux sous-types : le clustering hiérarchique agglomératif et le clustering hiérarchique divisif (Figure 2.5).
3. **Clustering basé sur la distribution :**
   Dans ce type, les objets de données sont supposés être générés par un mélange de distributions de probabilité. Chaque distribution a ses paramètres qui définissent les clusters. Si cette distribution de probabilité est gaussienne, cela signifie que nous avons affaire aux célèbres Modèles de Mélange Gaussien (Figure 2.6).
4. **Clustering basé sur la densité :**
   Les méthodes de clustering basées sur la densité considèrent les régions à forte densité comme des clusters. Alors que les objets dans les zones clairsemées qui séparent les clusters sont considérés comme des points de bruit ou de frontière (Figure 2.7).

2.6 Domaines d'application du clustring

En raison de l'explosion des informations sensorielles et textuelles, le clustering est devenu un outil très essentiel pour la découverte de connaissances [11] [12], l'analyse intelligente

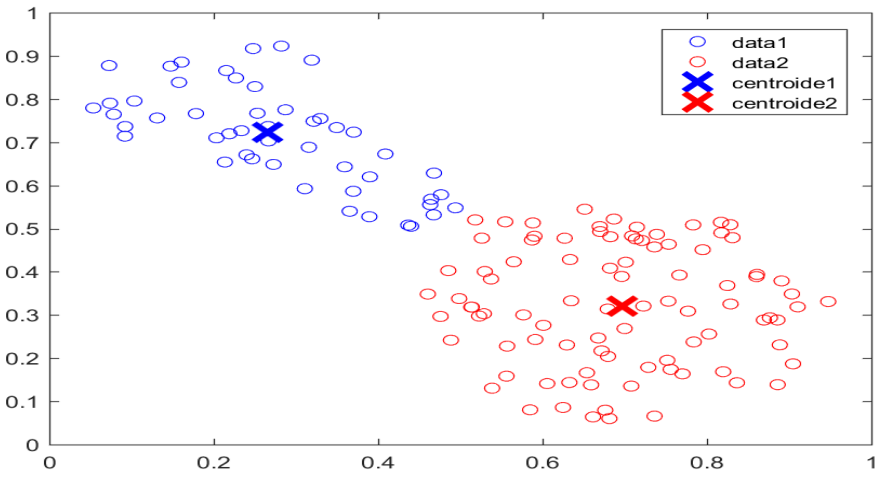

FIGURE 2.4 – Clustering basée sur le centroïde.

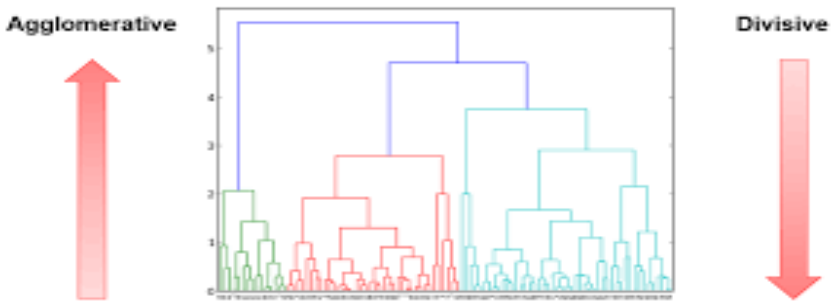

FIGURE 2.5 – Clustering hiérarchique : agglomératif et divisif.

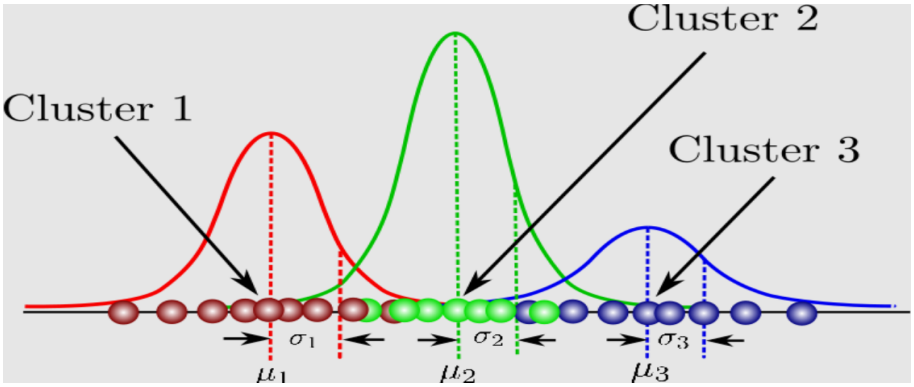

FIGURE 2.6 – Clustering basé sur le Modèle de Mélange Gaussien.

FIGURE 2.7 – Clustering basé sur la densité.

des données [13], et l'exploration de données [14], etc. Les analystes de données et les chercheurs ont adopté la théorie du clustering et algorithmes pour donner un sens aux données dans de nombreux domaines. Les méthodes de clustering sont utilisées dans de nombreuses applications, telles que :

1. **La reconnaissance de forms :**
   Les techniques de clustering tentent de diviser les modèles en leurs groupes appropriés (clusters) en fonction d'une mesure de similarité spécifiée.
2. **Analyse des données spatiales :**
   Le regroupement spatial [15] vise à diviser les données spatiales en clusters significatifs. L'analyse de grappes spatiales joue un rôle important dans la détermination des modèles de variance géographique. Généralement, il est utilisé dans de nombreux domaines tels que l'épidémiologie spatiale, la surveillance des maladies, l'écologie du paysage, la génétique des populations, etc.
3. **Étude de marché :**
   L'étude de marché [16] vise à recueillir des informations sur les clients et les marchés cibles. Dans les études de marché, un cluster est un groupe d'objets de données similaires et différents les uns des autres. Ces objets sont catégorisés en fonction d'un ensemble de variables telles que la démographie, les comportements d'achat, les préférences, etc. Ces variables peuvent être sélectionnées en fonction des cibles de l'étude de marché.
4. **Web mining :**
   C'est un domaine de recherche autonome qui peut être classé en trois catégories différentes en fonction de la partie du Web à exploiter [17]. Ces trois catégories du Web sont le contenu Web ; la structure Web ; et l'utilisation du Web. Dans ce dernier cas, les clusters d'utilisation et les clusters de pages sont des types de clusters qui doivent être découverts. Le clustering des utilisateurs vise à établir des clusters d'utilisateurs ayant le même comportement de navigation.

2.7 Avantages des algorithmes de clustering

Le but des algorithmes de clustering est de comprendre les données et d'extraire la valeur de grands volumes de données structurées et non structurées. Ces algorithmes vous permettent de séparer les données en fonction de leurs propriétés ou caractéristiques, et de les regrouper en différents clusters en fonction de leur similarité. Les algorithmes de clustering ont une variété d'utilisations dans différents domaines. Dans ce qui suit, nous montrerons quelques points nécessaires qui expliquent l'intérêt des algorithmes de clustering :

1. **Comprendre les données :**
   La collecte de données est une étape initiale dans tout projet d'apprentissage automatique. Après cela, nous avons besoin d'une meilleure compréhension de toutes les caractéristiques pour prédire la variable cible. La plupart des algorithmes de clustering sont utilisés pour l'analyse exploratoire des données. Par exemple, identifier les communautés dans les réseaux sociaux ou les clients qui se comportent de la même manière.
2. **Visualisation des données :**
   En plus de l'algorithme de réduction de dimensionnalité qui permet la visualisation des données en deux ou trois dimensions, nous pouvons utiliser un algorithme de clustering pour former des sous-groupes ou des clusters de ces points de données. Il est ainsi possible de représenter visuellement les relations entre les points de données. Ainsi, au lieu de représenter toutes les données, il est également possible d'afficher un seul point de données représentatif par cluster.
3. **L'inférence des propriétés à partir des données :**
   Cette utilisation d'algorithmes de clustering est particulièrement utile dans les cas fréquents où l'étiquetage des données est coûteux. L'utilisation d'algorithmes de clustering est une approche particulièrement utile. Il s'agit d'identifier : dans une image, les points qui appartiennent à un même objet (c'est ce qu'on appelle par segmentation) ; des images similaires, susceptibles de représenter le même objet (c'est-à-dire la même personne, le même animal, la même voiture) ; des textes similaires, susceptibles de parler du même sujet, etc.

2.8 Limitations des algorithmes du clustring

Bien que les algorithmes de clustering présentent de nombreux avantages, ils ont également certaines limitations et inconvénients. Voici quelques-uns :

- **Sensibilité aux paramètres :**
  Les résultats du clustering peuvent être influencés par les paramètres choisis pour l'algorithme. Si les paramètres ne sont pas choisis avec soin, le clustering peut produire des résultats incorrectes ou non pertinents.
- **Interprétation subjective :**
  La sélection des variables à inclure dans l'analyse peut être subjective, ce qui peut conduire à des interprétations différentes des résultats. Les résultats peuvent également varier en fonction de la méthode de clustering choisie.

— **Problèmes de taille d'échantillon :**
Les algorithmes de clustering peuvent ne pas fonctionner correctement si l'échantillon de données est trop petit ou trop grand. Les grands ensembles de données peuvent entraîner des problèmes de performance, tandis que les petits ensembles de données peuvent ne pas contenir suffisamment d'informations pour produire des résultats significatifs.
— **Sensibilité aux valeurs aberrantes :**
Les algorithmes de clustering peuvent être sensibles aux valeurs aberrantes ou aux erreurs dans les données. Les valeurs aberrantes peuvent influencer les résultats du clustering, ce qui peut conduire à des conclusions erronées.
— **Difficulté à choisir le nombre de clusters :**
Il peut être difficile de choisir le nombre de clusters approprié pour une analyse de clustering donnée. Un nombre trop élevé de clusters peut produire des groupes trop petits pour être significatifs, tandis qu'un nombre trop faible de clusters peut ne pas capturer toute la variance des données.

2.9 Algorithme de K-Means

2.9.1 Principe du clustering

L'objectif du regroupement K-means est de partitionner les données en K groupes de manière à minimiser la somme des carrés des distances intra-clusters. Dans cette technique du regroupement, nous devons spécifier le nombre de groupes ou de clusters que nous recherchons. Cette méthode est supervisée dans le sens où le nombre de classes doit être donné, mais pas nécessairement leurs paramètres [21].

Le regroupement K-means est composé de deux étapes. Tout d'abord, nous attribuons les observations à leur groupe le plus proche, en utilisant généralement la distance euclidienne entre l'observation et le centre de gravité du cluster. La deuxième étape consiste à calculer le nouveau centre de gravité du cluster en utilisant les objets attribués. Ces étapes sont alternées jusqu'à ce qu'il n'y ait plus de changements dans l'appartenance aux clusters ou jusqu'à ce que les centres de gravité ne changent pas (convergence). Cette dernière est assurée par un algorithme appelé HMEANS ou la méthode ISODATA. L'algorithme HMEANS est ainsi (Figure 2.8) :

1. Spécifier le nombre de clusters k.
2. Déterminer les centres de gravité du cluster initiaux. Ils peuvent être choisis au hasard (aléatoire) ou l'utilisateur peut les spécifier.
3. Calculer la distance entre chaque observation et chaque centre de gravité du cluster.
4. Attribuer chaque observation au cluster le plus proche.
5. Calculer le centre de gravité (c.-à-d., la moyenne d-dimensionnelle) de chaque cluster en utilisant les observations qui viennent d'y être regroupées.
6. Répéter les étapes 3 à 5 jusqu'à ce qu'il n'y ait plus de changements.

Généralement, il y a deux problèmes avec l'utilisation de l'algorithme HMEANS. Le premier est que cet algorithme pourrait conduire à des clusters vides, donc les utilisateurs devraient en être conscients. Comme le centre de gravité est recalculé et que les observations sont réattribuées à des groupes, certains clusters pourraient devenir vides. Le deuxième problème concerne l'optimalité des partitions. Avec k-means, nous cherchons des partitions où la somme des carrés intra-cluster est minimale. Il peut être démontré que dans certains cas, l'attribution finale de cluster k-means n'est pas optimale, dans le sens où le déplacement d'un seul point d'un cluster à un autre peut réduire la somme des carrés d'erreur.

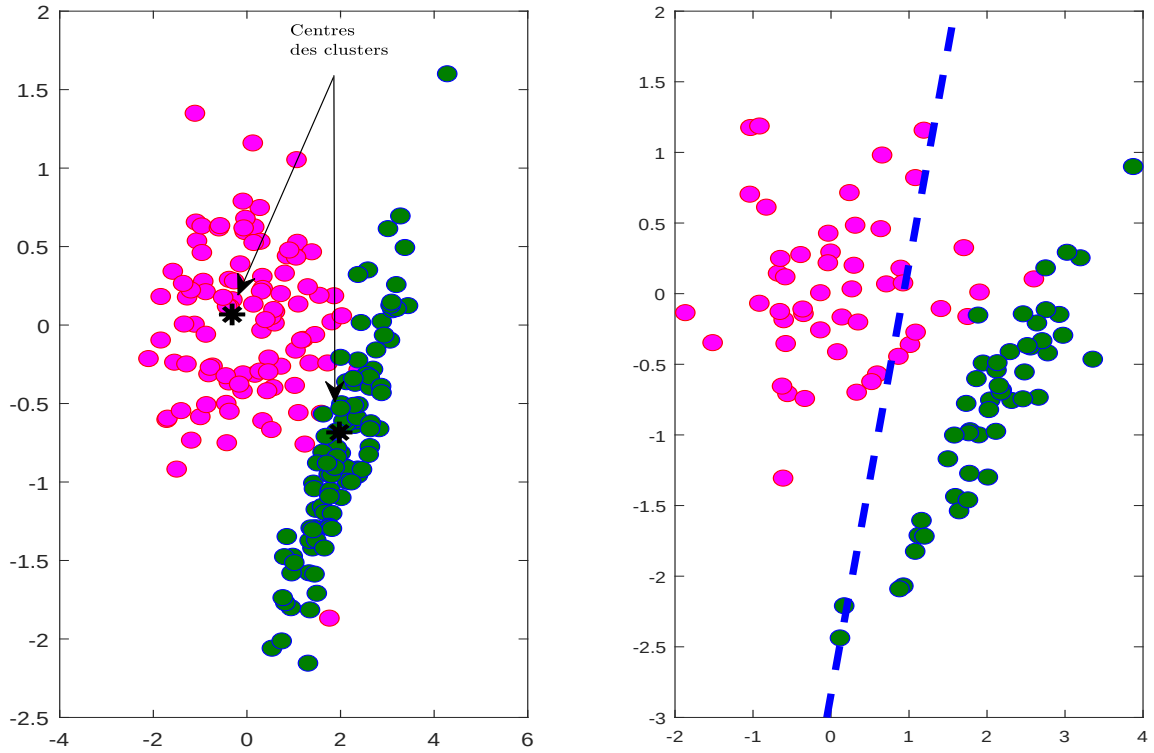

FIGURE 2.8 – Exemple de HMEANS clustring (ISODATA).

2.9.2 Fonctionnement de l'algorithme K-Means

L'algorithm de K-Means est donnée ci dessous (Algorithm 1) :

Le principe algorithmique de la méthode K-Means est décrite alors comme suit (voir figure 2.9) :

1. Obtenir une partition de k groupes, éventuellement à partir de l'algorithme HMEANS.
2. Prendre chaque point de données x_j et calculer la distance euclidienne entre celui-ci et chaque centroïde de cluster et attribuer chaque point de données x_j au groupe le plus proche.
3. Déterminer le nouveau centre de chaque classe en calculant la moyenne des points attribués.

Algorithme 1 K-Means.

1. **Entrées :** Ensemble de n données, noté par $X = \{x_1, x_2, ..., x_n\}$, Nombre de groupes souhaité, noté par k.
2. **Sorties :** Une partition de k groupes C_i, $i = 1, ..., k$.
3. **Initialisation :** initialiser aléatoirement les centres C_i
4. **Répéter :**
   — **Affectation :** Générer une nouvelle partition en assignant chaque objet au groupe dont le centre est le plus proche : $x_j \in C_k$ si $\forall i$

$$d(x_j, \mu_k) = \min d(x_j, \mu_i), \tag{2.1}$$

   avec μ_k est le centre de la classe k.
   — **Représentation :** Calculer les centres associe à la nouvelle partition ;

$$\mu_k = \frac{1}{n} \sum_{x \in C_k} x_j. \tag{2.2}$$

5. **Jusqu' à** convergence de l'algorithme vers une partition stable.
6. **fin**

4. Répétez les étapes 2 à 3 jusqu'à ce qu'il n'y ait plus de changements effectués. Nous notons qu'il existe de nombreux algorithmes de regroupement k-means décrits dans la littérature.

Ce processus tente de maximiser la similarité intra-cluster représentée sous forme d'une fonction objectif :

$$\sum_{i=1}^{k} \sum_{x_j \in C_i} d(x_j, C_i), \tag{2.3}$$

Dans le cas de la distance euclidienne cette fonction est appelée fonction d'erreur quadratique. La convergence est atteinte quand il n'y a plus de changement. La limite principale de cette méthode dépendant des résultats des valeurs de départ (centres initiaux). À chaque initialisation correspond une solution différente (optimum local) qui peut dans certain cas être très loin de la solution optimale (optimum global). Une solution naïve À ce probleme consiste À lancer l'algorithme plusieurs fois avec différentes initialisations et retenir le meilleur regroupement trouvé. L'usage de cette solution reste limité du fait de son coût et que l'on peut trouver une meilleure partition en une seule exécution [4].

2.10 Avantages et Inconvénients du K-means

2.10.1 Avantages

L'algorithme K-means présente plusieurs avantages, notamment :

1. L'algorithme K-Means est très populaire car il est simple, facile à comprendre et à mettre en oeuvre.

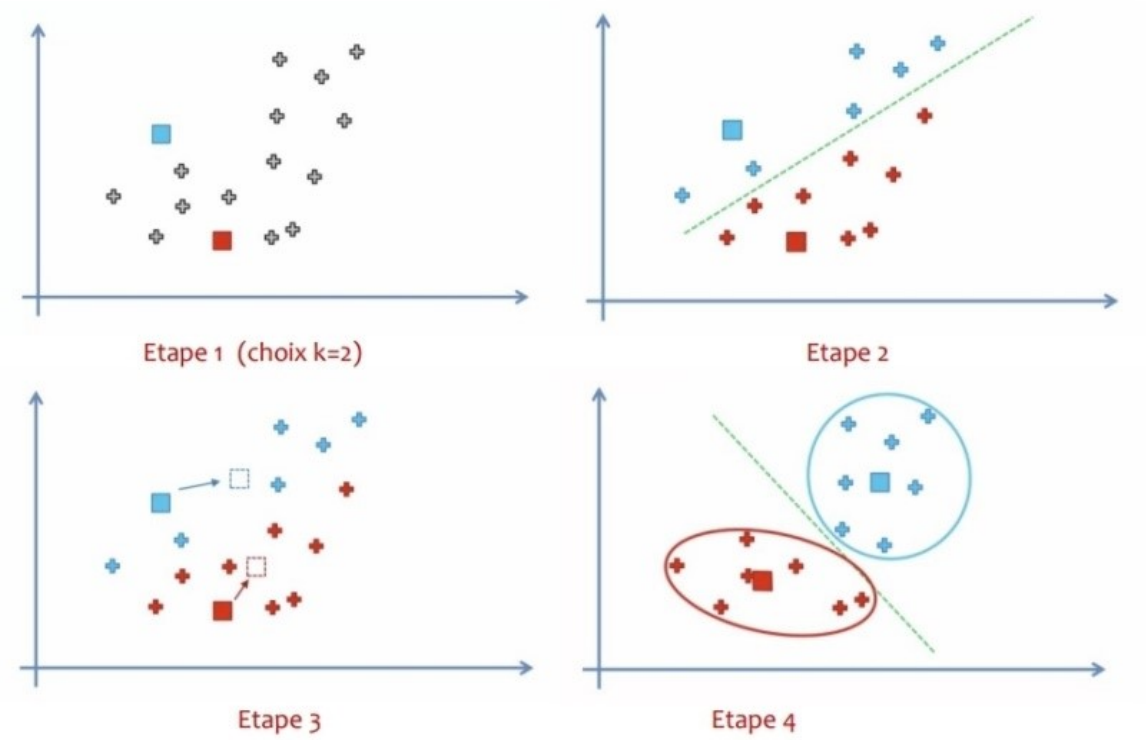

FIGURE 2.9 – Étapes de K-means clustring.

2. Il tend à réduire l'erreur quadratique.
3. Il est applicable à des données de grandes tailles.
4. Cette méthode est adaptée à des tâches non supervisées, donc elle ne nécessite aucune information sur les données.
5. Elle est rapide et nécessite peu de mémoire.
6. Elle peut être appliquée à tout type de données, y compris les données textuelles, en choisissant une bonne notion de distance.

2.10.2 Inconvénients

Il existe plusieurs inconvénients associés à l'algorithme du K-means :

1. Besoin de spécifier le nombre de clusters k à l'avance, ce qui peut être difficile à déterminer en pratique.
2. Sensible aux valeurs aberrantes : Le k-means peut être sensible aux valeurs aberrantes ou aux points qui sont très différents de la plupart des autres points dans le jeu de données, car il cherche à minimiser la somme des carrés des distances entre les points et les centroïdes.
3. Sensible au choix initial des centroïdes, ce qui peut conduire à des solutions suboptimales ou différentes selon les choix initiaux.

4. Non approprié pour des données non-sphériques : L'algorithme du k-means est conçu pour des données sphériques, ce qui signifie que les clusters doivent être de forme sphérique, avec une variance égale dans toutes les dimensions. Il n'est donc pas approprié pour des données non-sphériques ou ayant des structures plus complexes.
5. Problèmes de convergence : Il y a des situations où l'algorithme du k-means peut ne pas converger vers une solution, ou peut converger vers une solution locale plutôt que globale, ce qui signifie que la solution trouvée peut ne pas être optimale.

2.11 Conclusion

Ce chapitre présente un bref aperçu sur le clustering de données et ses applications dans le monde réel. Le principe de fonctionnement des méthodes de clustering est envisagé et notamment celui qui en relation avec la méthode K-means. Les avantages et les limitations de clustering sont aussi abordés.

En fin, on peut dire que le K-means (K-moyennes) est un algorithme non supervisé de clustering, populaire en Machine Learning. Il n'essaie pas d'apprendre une relation de corrélation entre un ensemble de caractéristiques X d'une observation et une valeur à prédire Y, comme c'est le cas pour l'apprentissage supervisé. L'apprentissage non supervisé va plutôt trouver des modéles dans les données. Notamment, en regroupant les choses qui se ressemblent.

Chapitre 3

La méthode Fuzzy C-Means (FCM)

3.1 Introduction

Précédemment, nous avons décrit le problème général de clustering et avons également discuté de la méthode de clustering K-means. Dans ce chapitre, nous visons à définir et décrire l'approche de clustering flou, en particulier la méthode Fuzzy c-means "FCM"(c-moyennes floue). FCM est un algorithme de classification floue non supervisée. Il introduit la notion d'ensemble flou dans la définition des classes : chaque point de l'ensemble de données appartient à chaque cluster avec un certain degré, et tous les clusters sont caractérisés par leur centre de gravité. De plus, les avantages et les inconvénients de cette méthode sont également mises en évidence. Dans le chapitre suivant, nous appliquerons la méthode de FCM clustering à plusieurs ensembles de données.

3.2 La logique floue

La logique floue est une extension de la logique booléenne créée par Lotfi Zadeh en 1965 [17] en se basant sur sa théorie mathématique des ensembles flous, c'est une généralisation de la théorie des ensembles classiques. En logique booléenne, une variable ne peut prendre que deux valeurs : vrai (1) ou faux (0). Les propositions énoncées en prémisse d'une règle et en conclusion ne peuvent être que totalement vraies ou totalement fausses. En revanche, la logique floue, qui se rapproche du raisonnement humain, et il s'agit d'une logique linguistique, floue ou approximative, où les valeurs de vérité sont exprimées en termes du langage courant, tels que "plutôt vrai", "presque faux", etc. Il introduise la notion de degré d'appartenance dans la vérification d'une condition, nous donnons la possibilité à une condition d'être dans un autre état que vrai ou faux. La logique floue confère ainsi une flexibilité très appréciable aux raisonnements qui l'utilisent, ce qui rend possible la prise en compte des imprécisions et des incertitudes. Donc, la logique floue permet de traiter les données imprécises et incertaines de manière efficace, en fournissant des résultats adaptés aux variations et aux imprécisions du système. Cela en fait un outil précieux pour de nombreux domaines, tels que la robotique, la reconnaissance de formes, la prise de décision, la surveillance et le contrôle de processus complexes.

3.2.1 Remarques

— La logique floue diffère de la logique classique parce qu'elle permet des définitions partielles ou “floues” des règles de contrôle.
— La logique floue est une branche mathématique qui permet aux ordinateurs de modéliser le monde réel de manière similaire à celle des êtres humains.

— Elle s'intéresse à la quantification et au raisonnement en utilisant un langage qui permet des définitions ambiguës telles que "beaucoup", "peu", "petit", "haut", "dangereux", etc.
— Elle traite des situations où les questions posées et les réponses obtenues impliquent des concepts vagues.

3.2.2 Théorie des ensembles flous

La logique floue repose sur les mathématiques et se présente comme une méthode de raisonnement qui ressemble à celui d'un être humain. Elle permet de traiter des vérités partielles, ce qui peut être utile dans des situations où l'on ne peut pas décider si une affirmation est vraie ou fausse. Depuis les années 1920, la logique floue a été étudiée, mais le terme "logique floue" a été utilisé pour la première fois en 1965 par Lotfi Zadeh, un professeur iranien à l'Université de Californie à Berkeley.

Le professeur Zadeh a constaté que la logique informatique classique était incapable de traiter des données représentant des pensées humaines subjectives ou imprécises. Les catégories définies de manière floue sont importantes dans la communication et l'apprentissage humains. Pour résoudre ce problème, Zadeh a introduit une classe d'objets appelée ensemble flou, qui est plus générale que les ensembles mathématiques ordinaires et possède des degrés d'appartenance (voir Fig. 3.1). Malgré l'ambiguïté et l'imprécision qui existent dans la vie, les humains cherchent toujours à prendre des décisions raisonnables. Ainsi, la théorie des ensembles flous a été développée pour décrire mathématiquement l'imprécision du flou présent dans notre langage de la vie quotidienne.

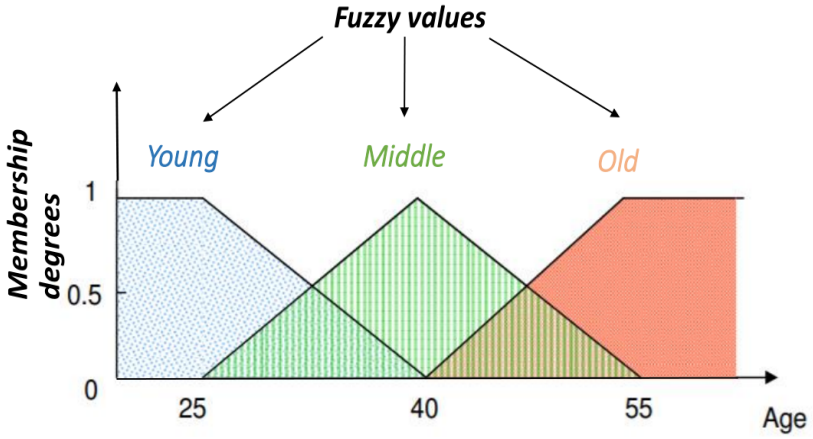

FIGURE 3.1 – Degrés d'appartenance des ensembles flous.

Soit X un espace de données, dans lequel chaque élément est désigné par x. Un ensemble flou A, défini dans X, est représenté par une fonction d'appartenance $U_A(x)$, qui relie chaque point x de X à un nombre compris entre 0 et 1 (c'est-à-dire $x \rightarrow U_A(x) \in [0,1]$). La valeur de $U_A(x)$ représente le degré d'appartenance de x à A, où une valeur plus proche de 1 (unité) indique que x appartient fortement à A. Différentes définitions existent pour les ensembles flous, qui étendent simplement les définitions des ensembles ordinaires [17].

Définitions :

1. *On peut nommer un ensemble flou "ensemble vide" si sa fonction d'appartenance, notée $U_A(x)$, est identiquement nulle sur X, c'est-à-dire si $U_A(x) = 0$, $\forall\ x \in X$.*
2. *On peut dire que deux ensembles flous (A et B) sont égaux si $U_A(x) = U_B(x)$, $\forall\ x \in X$.*
3. *Si A' est ,e complément de A, on aura :*

$$U_{A'} = 1 - U_A, \tag{3.1}$$

Où U_A et $U_{A'}$ sont les notations simplifiées de $U_A(x)$ et $U_{A'}(x)$, respectivement.

4. *On dit que B contient A si et seulement si $U_A \leq U_B$, et il peut être écrit comme suit :*

$$A \subset B \Longleftrightarrow U_A \leq U_B, \tag{3.2}$$

5. *L'union des ensembles floues A et B avec les degrés d'appartenances respectifs $U_A(x)$ et $U_B(x)$ est un ensemble flou $C : C = A \cup B$. Où le degré d'appartenance de C est lié á ceux de A et B et il est donné par :*

$$U_C(x) = Max[U_A(x), U_B(x)], \tag{3.3}$$

3.2.3 Avantages et Inconvénients des systèmes flous

Certains avantages des systèmes logiques flous sont énumérés ci-dessous :

1. La structure de la logique floue est facilement compréhensible.
2. Elle peut fournir une solution efficace pour des problèmes complexes.
3. La logique floue peut aider à contrôler les machines et les produits de consommation dans le domaine de l'IA.
4. Elle peut aider les utilisateurs à gérer l'incertitude en ingénierie lors de l'exploration de données.
5. Bien qu'elle ne fournisse pas un raisonnement précis, elle est souvent la seule forme de raisonnement acceptable.

Cependant, certains inconvénients des systèmes de logique floue sont également à prendre en compte :

1. Parfois, la logique floue peut donner des résultats inexacts, de sorte qu'ils sont considérés comme des hypothèses et peuvent ne pas être largement acceptés.
2. L'établissement de règles précises et floues ainsi que la définition des degrés d'appartenance (fonctions) est une tâche difficile.
3. La validation d'un système flou basé sur les connaissances nécessite des tests matériels approfondis.
4. Les systèmes flous ne peuvent pas reconnaître l'apprentissage automatique et les réseaux neuronaux pour la reconnaissance de modèles.

3.3 Fuzzy C-Means Clustering

L'algorithme C-means (C-means) est une méthode de regroupement couramment utilisée pour effectuer une analyse de clustering. Il est également connu sous le nom d'algorithme des moyennes floues (fuzzy c-means clustering FCM). FCM est une extension de l'algorithme K-means qui permet une affectation floue des points de données à différents clusters, plutôt qu'une affectation binaire.

3.3.1 Modèle d'optimisation FCM

Le clustering FCM minimise la fonction objectif donnée ci-dessous :

$$J_{FCM}(U, V, X) = \sum_{i=1}^{c} \sum_{j=1}^{n} u_{ij}^{m} d_{ij}^{2}, \tag{3.4}$$

Avec les contraintes

$$\sum_{i=1}^{c} u_{ij} = 1, \quad 0 \leq u_{ij} \leq 1, \quad 1 \leq j \leq n, \tag{3.5}$$

Dans la formule (3.4), U et V représentent respectivement les degrés d'appartenance et les centres de cluster. Les paramètres X, m et c sont les entrées de la fonction objectif J_{FCM}, et voici une explication détaillée :

- c : Il s'agit du nombre supposé de clusters dans l'ensemble de données. C'est un paramètre déterminé par l'utilisateur qui spécifie le nombre de clusters souhaité dans la partition.
- m : C'est un exposant de pondération pour la fuzzification qui doit être supérieur ou égal à 1. Il contrôle le degré de flou de la solution. Lorsque la valeur de m devient plus grande, la solution devient plus floue. En revanche, lorsque m devient égal à 1, le modèle se rapproche du clustering c-means dur (HCM), ce qui conduit à des résultats nets.
- u_{ij} : Il représente le degré d'appartenance du j-ème point de données x_j dans le i-ème cluster μ_i. Les clusters sont représentés par leurs centres (V = $\{\mu_i\}$ avec $i = 1$ à c). $U = [u_{ij}]$ est une matrice de partition floue ($c \times n$) qui satisfait la contrainte donnée par l'équation (3.5).
- n : Il représente le nombre de points de données, c'est-à-dire la taille de l'ensemble de données.
- d_{ij}^2 représente la mesure de la distance entre le centre du cluster μ_i et le point de données x_j. avec

$$d_{ij}^{2}(x_j, \mu_i) = \|x_j - \mu_i\|^2, \tag{3.6}$$

3.3.2 Conditions pour l'optimalité

Soit $(U^*; V^*)$ les minimiseurs de la fonction objectif $J_{FCM}(U, V)$ de la méthode FCM. Les conditions requises pour optimiser la fonction objectif sont énumérées comme suit :

$$u_{ij}^* = \frac{1}{\sum_{k=1}^{c} \left(\frac{d_{ij}^2}{d_{kj}^2} \right)^{\frac{1}{m-1}}}, \tag{3.7}$$

$$\mu_i^* = \frac{\sum_{j=1}^{n} u_{ij}^m x_j}{\sum_{j=1}^{n} u_{ij}^m}, \tag{3.8}$$

Le partitionnement flou est réalisé par une optimisation itérative de la fonction objectif avec une mise à jour des degrés d'appartenance u_{ij} et des centres des classes μ_i comme dans le cas de l'algorithme K-means.

Preuve :

Les formules de mise à jour sont obtenues par l'introduction d'un multiplicateur de Lagrange λ associé à la contrainte de normalisation donnée dans l'équation (3.5). En appliquant le Lagrangien par rapport à x_j on aura :

$$L(x_j) = \sum_{i=1}^{c} u_{ij}^m \|x_j - \mu_i\|^2 + \lambda(\sum_{i=1}^{c} u_{ij} - 1), \lambda > 0, \tag{3.9}$$

La minimisation du Lagrangien par rapport aux degrés d'appartenance u_{ij} et au coefficient de Lagrange λ, nous donne :

$$\frac{\partial L(x_j)}{\partial u_{ij}} = 0 \Rightarrow m(u_{ij}^{m-1}) \|x_i - \mu_i\|^2 + \lambda = 0, \tag{3.10}$$

$$\frac{\partial L(x_j)}{\partial \lambda} = 0 \Rightarrow \sum_{i=1}^{c} u_{ij} - 1 = 0 \Rightarrow \sum_{i=1}^{c} u_{ij} = 1, \tag{3.11}$$

A partir de l'équation (3.10), on déduit :

$$u_{ij} = \left(\frac{-\lambda}{m} \right)^{\frac{1}{m-1}} (\frac{1}{\|x_j - \mu_i\|^2})^{\frac{1}{m-1}}, \tag{3.12}$$

Et à partir de l'équations (3.11) et (3.12), on obtient :

$$\left(\frac{-\lambda}{m} \right)^{\frac{1}{m-1}} (\frac{1}{\|x_j - \mu_i\|^2})^{\frac{1}{m-1}} = 1, \tag{3.13}$$

$$\Leftrightarrow \left(\frac{-\lambda}{m} \right)^{\frac{1}{m-1}} = \frac{1}{\sum_{l=1}^{c} (\frac{1}{\|x_j - \mu_l\|^2})^{\frac{1}{m-1}}}, \tag{3.14}$$

$$\Leftrightarrow \frac{-\lambda}{m} = \frac{1}{\sum_{l=1}^{c} (\frac{1}{\|x_j - \mu_l\|^2})}, \tag{3.15}$$

En remplaçant l'équation (3.15) dans l'équation (3.12), on aboutit à :

$$u_{ij} = \left(\frac{1}{\sum_{l=1}^{c}\left(\frac{1}{\|x_j-\mu_l\|^2}\right)}\right)^{\frac{1}{m-1}}\left(\frac{1}{\|x_j-\mu_i\|^2}\right)^{\frac{1}{m-1}} \Leftrightarrow u_{ij} = \frac{1}{\sum_{l=1}^{c}\left(\frac{\|x_j-\mu_i\|^2}{\|x_j-\mu_l\|^2}\right)^{\frac{1}{m-1}}}, \tag{3.16}$$

D'où la formule (3.7). La minimisation du (3.9) par rapport à μ_i, nous donne :

$$\frac{\partial L(x_j)}{\partial \mu_i} = 0 \Leftrightarrow \sum_{j=1}^{n} 2\|x_j-\mu_i\|u_{ij}^m = 0, \tag{3.17}$$

$$\Leftrightarrow \sum_{j=1}^{n} u_{ij}^m x_j - \sum_{j=1}^{n} u_{ij}^m \mu_i = 0. \tag{3.18}$$

D'où la formule de mise à jour des centres de classes (3.8).■

3.3.3 Algorithme FCM

L'algorithme FCM cherche à minimiser une fonction d'objectif qui est basée sur la distance euclidienne entre les points de données et les centroïdes des clusters, pondérée par les degrés d'appartenance. L'algorithme itère jusqu'à atteindre une convergence où les centroïdes ne changent plus significativement et les degrés d'appartenance se stabilisent.
L'algorithme FCM (Algorithme 2) fonctionne de la manière suivante :
1. Spécifier le nombre de clusters c et un degré de flouté m.
2. Sélectionner les centroïdes initiaux pour chaque cluster de manière aléatoire ou en utilisant une méthode spécifique.
3. Répéter les étapes suivantes (a et b) jusqu'à ce que les centroïdes convergent et que les degrés d'appartenance se stabilisent.
a. Affecter chaque point de données à un cluster en calculant sa "flouité" (degré d'appartenance) à chaque cluster en fonction de la distance euclidienne par rapport aux centroïdes.
b. Mettre à jour les centroïdes de chaque cluster en calculant la moyenne pondérée des points de données en fonction de leur degré d'appartenance.

3.3.4 Notes sur le FCM clustering

Plusieurs travaux de recherche ont été réalisés pour étudier l'effet de la variable m sur les résultats obtenus par la méthode FCM (Fuzzy C-Means) [18] [19]. Les études montrent que l'utilisation d'une valeur élevée de m rend les partitions floues et les centres de cluster se concentrent vers le centre des données. Les conclusions expérimentales indiquent que la valeur $m = 2$ est favorable, car elle permet de simplifier les équations mises à jour.

La contrainte de l'équation (3-5) impose le degré d'appartenance d'un point de données à l'intérieur d'un cluster, en prenant en compte les distances des points par rapport aux prototypes des autres clusters. Ainsi, les points qui sont très proches d'un cluster par rapport aux autres auront un degré d'appartenance presque nul pour les autres clusters. Tandis que

Algorithme 2 : FCM

1. **Entrée :** c : 2≤ c < N and m : 1≤ m < ∞, ε.
2. **Sorties :** Centres de clusters, et le degré d'appartenance.
3. **Initialisation :** les centres du clusters floues, V, $k = 1$.
4. **Répéter :**
   - Calculer les degrés d'appartenances floues u^*_{ij}, on utilisant l'equation (3-7).
   - Mettre a jours les centres du clusters floues, μ^*_i on utilisant l'equation (3-8).
   - Incrémenter k.
5. **Jusqu' à** $\|u^{k-1}_{ij} - u^k_{ij}\| < \varepsilon$
6. **fin**

les points situés au milieu entre deux clusters auront des degrés d'appartenance proches de $0,5$.

Les clusters avec un plus grand nombre d'échantillons et des diamètres plus larges vont dominer efficacement la solution de FCM, car la somme est liée à n. De plus, la fonction objectif de FCM dépend d'une distance de séparation, dans laquelle les distances pondérées entre les points de données et les prototypes sont accumulées. Ainsi, un cluster avec un grand diamètre aura plus d'impact sur la fonction objectif qu'un cluster avec un petit diamètre, en raison des distances plus élevées. En conclusion, les diamètres relatifs des clusters jouent un rôle dans la détermination de quel cluster contribue le plus à la fonction objectif.

3.3.5 Les avantages et les inconvénients du FCM

Comme toute méthode de regroupement, le FCM clustering (Fuzzy C-Means) présente des des forces et des faiblesses. La classification FCM a été largement utilisée par les chercheurs, où elle a montré son efficacité dans plusieurs applications. Cela est dû à plusieurs raisons :

1. La méthode FCM utilise une fonction objectif intuitive et facile à comprendre.
2. En raison de sa base floue, elle offre une robustesse : elle converge toujours vers une solution et fournit des degrés d'appartenance cohérents.
3. En ce qui concerne sa mise en œuvre en programmation, elle est relativement claire.
4. Elle fournit des résultats de regroupement précis, en particulier pour les ensembles de données présentant des clusters bien séparés en forme d'hyper-sphères.

Malgré son grand succès, cet algorithme de regroupement présente encore plusieurs inconvénients. Nous en énumérons certains ci-dessous :

1. L'algorithme nécessite de connaître à l'avance le nombre de clusters souhaité.
2. Il nécessite une bonne initialisation des centres de cluster, ce qui est difficile à réaliser dans les expériences.
3. La solution de la fonction objectif de FCM peut varier d'un essai à l'autre en fonction de l'initialisation, ce qui est lié au problème général de l'optimisation locale et globale.

4. L'algorithme est sensible aux points de bruit et aux valeurs aberrantes en raison de la contrainte probabiliste, ainsi que du carré de l'erreur entre les points de données et les centres, ce qui confirme l'influence du bruit et des valeurs aberrantes.
5. L'algorithme ne cherche pas à détecter des ensembles de données présentant différentes formes de clusters.

3.4 Conclusion

Dans ce chapitre, nous avons présentées l'idée de base du clustering flou, précisément le clustering FCM. Les principales démarches de la logique floue, la théorie des ensembles flous, et la notion des fonctions d'appartenances sont présentées.

L'algorithme Fuzzy C-Means est une extension de l'algorithme de k-means, où la notion d'ensemble flou dans la définition des classes est introduite. Cet algorithme utilise un critère de minimisation des distances intra-classes et de maximisation des distances interclasses, mais en tenant compte des degrés d'appartenance et en optimisant une fonction objectif J. La méthode présentée fait partie des techniques de classification non supervisée et qui est très souvent utilisée dans le domaine de l'imagerie pour la segmentation des images.

Chapitre 4

Réalisation et expérimentation

4.1 Introduction

Dans ce chapitre, on s'intéresse à la segmentation d'images en niveaux de gris en utilisant la méthode Fuzzy C-means. Une image en niveaux de gris permet une dégradation de gris entre le noir et le blanc. En général, on code le niveau de gris sur un octet (8 bits) ce qui donne 256 nuances dégradé. L'expression de la valeur du niveau de gris avec $Ng = 256$ devient : $p(i, j) \in [0, 255]$. Où, la valeur 0 représente la brillance ou l'intensité minimale (le noir) et 255 la brillance ou l'intensité maximale (le blanc). En effet, chaque pixel de l'image est caractérisé par son attribut en qualité d'un niveau de gris.

D'abord, nous aurons un apercu du Software utilisé pour appliquer l'algorithme FCM, puis nous réaliserons notre expérience sur deux images différentes.

4.2 Le Software (Matlab)

MATLAB est une plateforme de calcul scientifique et de programmation de haut niveau, utilisant un environnement interactif qui permet de résoudre de manière précise des tâches de calcul complexes plus rapidement qu'avec les langages de programmation traditionnels. Il est actuellement la plateforme de calcul privilégiée dans les domaines des sciences, de l'ingénierie et de nombreux secteurs techniques.

MATLAB est également un environnement interactif de calcul technique de haut niveau pour le développement d'algorithmes, la visualisation de données, l'analyse de données et les calculs numériques. MATLAB est adapté pour résoudre des problèmes de calcul technique à l'aide d'algorithmes optimisés qui sont faciles à utiliser pour l'utilisateur final grâce à des commandes simples. Il est possible d'utiliser MATLAB dans un large éventail d'applications, notamment le calcul mathématique, l'algèbre, les statistiques, l'économétrie, le contrôle qualité, les séries temporelles, le traitement des signaux et des images, les communications, la conception de systèmes de contrôle, les systèmes de test et de mesure, la modélisation et l'analyse financière, la biologie computationnelle, etc. Les ensembles d'outils complémentaires appelés "boîtes à outils" (collections de fonctions MATLAB pour des objectifs spécifiques, disponibles séparément) étendent l'environnement de MATLAB et vous permettent de résoudre des problèmes spécifiques dans différents domaines d'application [20].

Il est possible d'intégrer du code MATLAB avec d'autres langages et applications, en plus des algorithmes distribués et des applications développés avec MATLAB. Dans l'ensemble, les fonctions, les commandes et les capacités de programmation de l'écosystème MATLAB constituent une collection vraiment impressionnante. Voici les caractéristiques importantes liées aux graphiques de MATLAB :

— Un langage de calcul technique de haut niveau.
— Un environnement de développement pour gérer le code, les fichiers et les données.
— Des outils interactifs pour l'exploration, la conception et les solutions itératives.
— Des fonctions mathématiques pour l'algèbre linéaire, les statistiques, les transformations de Fourier, le filtrage, l'optimisation et l'analyse d'intégration numérique.
— Des fonctions graphiques bidimensionnelles et tridimensionnelles pour visualiser les données.
— Des outils pour créer des interfaces graphiques utilisateur personnalisées.
— Des fonctions pour intégrer des algorithmes basés sur des applications MATLAB avec des langages externes tels que C/C++, Fortran, Java, Microsoft .Net, Excel, et d'autres.

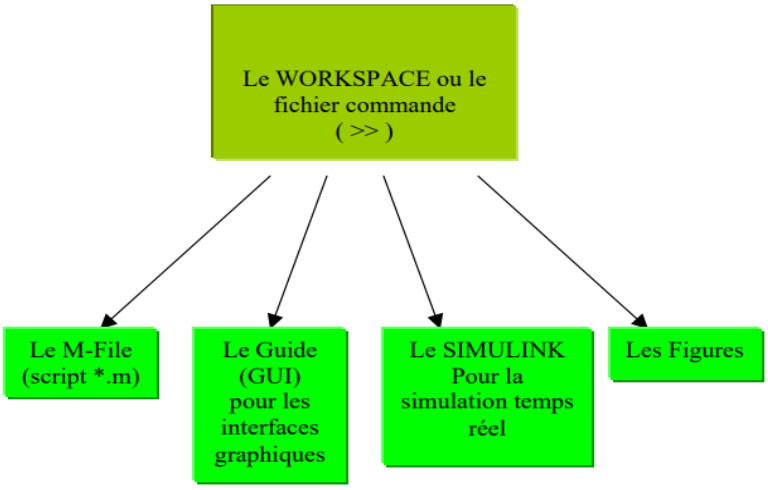

FIGURE 4.1 – La structure MATLAB

4.3 Différente Fonctionnalités de Matlab

L'environnement de développement MATLAB vous permet de développer des algorithmes, d'analyser des données, d'afficher des fichiers de données et de gérer des projets en mode interactif avec la fenêtre de commande, qui est le centre d'activité principal, comme illustré dans la Figure 4.2.

4.3.1 Développement d'algorithmes et d'applications

MATLAB vous permet d'exécuter des commandes ou des groupes de commandes, un par un, sans compilation ni liaison, afin d'obtenir la solution optimale. Pour effectuer rapidement des calculs complexes sur des vecteurs et des matrices, MATLAB utilise des bibliothèques optimisées pour le processeur. Pour de nombreux calculs, MATLAB génère des instructions en code machine en utilisant la technologie JIT (Just-In-Time). Grâce à cette technologie,

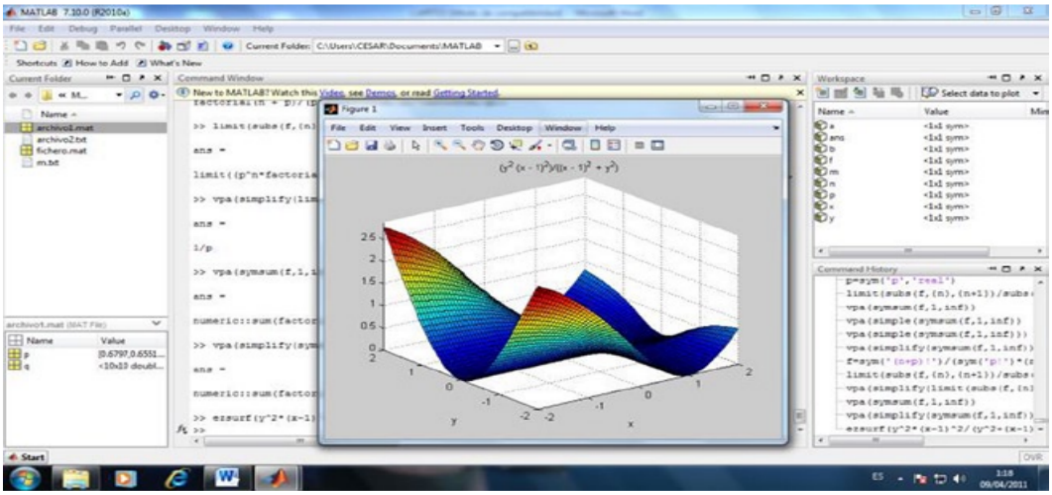

FIGURE 4.2 – Environement Matlab

disponible sur la plupart des plates-formes, les temps d'exécution sont beaucoup plus rapides que ceux des langages de programmation traditionnels.

MATLAB comprend des outils de développement qui aident à mettre en œuvre efficacement les algorithmes. Voici quelques-uns d'entre eux :

— Éditeur MATLAB : doté de fonctions d'édition et d'offres de débogage standard telles que la définition de points d'arrêt et les simulations pas à pas.
— Vérificateur de code M-Lint : analyse le code et recommande des modifications pour améliorer les performances et la maintenance.
— Profiler MATLAB : Enregistre le temps nécessaire à l'exécution de chaque ligne de code.
— Rapports de répertoire : Analyse tous les fichiers d'un répertoire et crée des rapports sur l'efficacité du code, les différences entre les fichiers, les dépendances entre les fichiers et la couverture du code.

Il permet également d'utiliser l'outil interactif *GUIDE* (Graphical User Interface Development Environment) pour concevoir et éditer des interfaces utilisateur. Cet outil vous permet d'inclure des listes de sélection, des menus déroulants, des boutons poussoirs, des boutons radio et des curseurs, ainsi que des diagrammes MATLAB et des contrôles ActiveX. Vous pouvez également créer des interfaces utilisateur graphiques en utilisant des fonctions MATLAB. Vous pouvez également rendre des figures ou des algorithmes MATLAB accessibles sur le Web.

4.3.2 Accès aux données et analyse

MATLAB prend en charge l'ensemble du processus d'analyse des données, de l'acquisition des données à partir de périphériques externes et de bases de données, en passant par le prétraitement, la visualisation et l'analyse numérique, jusqu'à la production de résultats de qualité de présentation.

MATLAB propose des outils interactifs et des opérations en ligne de commande pour

l'analyse des données, qui comprennent notamment :

— Sélection de sections de données
— Mise à l'échelle et moyenne
— Interpolation
— Seuillage et lissage
— Corrélation
— Analyse de matrices, etc.

4.3.3 Visualisation de données

Toutes les fonctions graphiques nécessaires pour visualiser des données scientifiques et techniques sont disponibles dans MATLAB. MATLAB comprend des fonctionnalités pour la représentation de diagrammes bidimensionnels et tridimensionnels, la visualisation de volumes tridimensionnels, des outils pour créer des diagrammes de manière interactive et la possibilité d'exporter vers les formats graphiques les plus populaires. Il est possible de personnaliser les diagrammes en ajoutant des axes multiples, en modifiant les couleurs des lignes et des marqueurs, en ajoutant des annotations, des équations LaTeX, des légendes et d'autres options de tracé.

Les fonctions vectorielles représentées par des diagrammes bidimensionnels peuvent être utilisées pour créer :

— Diagrammes de lignes, d'aires, de barres et de secteurs.
— Diagrammes de direction et de vitesse.
— Histogrammes.
— Polygones et surfaces.
— Diagrammes de dispersion de bulles.
— Animations.

4.3.4 Calcul numérique

MATLAB contient des fonctions mathématiques, statistiques et d'ingénierie qui prennent en charge la plupart des opérations à effectuer dans ces domaines. Ces fonctions, développées par des experts en mathématiques, constituent la base du langage MATLAB. Voici quelques exemples des fonctions mathématiques et d'analyse implémentées par MATLAB dans les domaines suivants :

— Manipulation de matrices et algèbre linéaire
— Polynômes et interpolation
— Calcul différentiel et intégral
— Analyse de Fourier et filtres
— Statistiques et analyse de données
— Optimisation et intégration numérique
— Équations différentielles ordinaires (EDO)
— Équations différentielles aux dérivées partielles (EDP)
— Opérations sur les matrices creuses

4.4 Segmentation d'image

Le processus d'analyse d'image peut être défini comme l'ensemble des méthodes et outils utilisés pour décrire quantitativement le contenu d'une image.

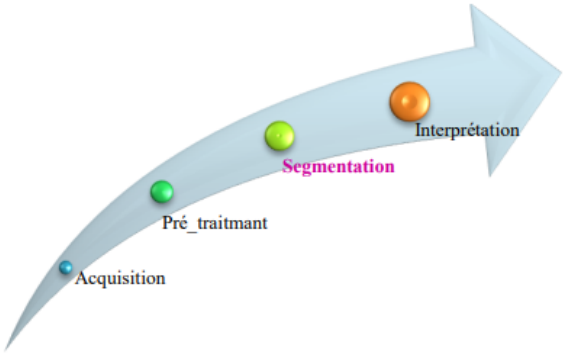

FIGURE 4.3 – Analyse de l'image

La segmentation d'une image est une étape essentielle et critique de l'analyse d'images (figure 4.3). Une segmentation d'image de qualité permet une analyse précise, car c'est à partir de l'image segmentée que les mesures sont effectuées pour extraire les paramètres discriminants nécessaires à la classification ou à l'interprétation. L'objectif de la segmentation est de fournir une description concise et représentative du contenu informationnel de l'image, plus exploitable que l'ensemble de ses pixels individuels.

4.4.1 Définition de la segmentation

La segmentation est un processus de traitement d'image qui vise à diviser l'image A en plusieurs sous-ensembles appelés régions, de telle manière qu'aucune région ne soit vide, qu'il n'y ait pas d'intersection entre deux régions et que l'ensemble des régions recouvre l'intégralité de l'image. Une région est définie comme un ensemble de pixels connectés qui partagent des propriétés communes les distinguant des pixels des régions voisines.

FIGURE 4.4 – Exemple d'image en couleur segmentée

Quelques règles à suivre pour obtenir une segmentation sont ainsi :
— Les régions doivent être uniformes et homogènes par rapport à certaines caractéristiques (niveau de gris, écart type, gradient).
— Leurs intérieurs doivent être simple et sans beaucoup de petits trous (des parties de région non segmentés).
— Les régions adjacentes doivent avoir des valeurs très différentes par rapport à la caractéristique prise en compte dans la segmentation.
— Les limites de chaque région doivent être simples et spatialement précises.

4.4.2 Types d'approches pour la segmentation

Il existe deux types d'approches pour la segmentation : l'approche contour et l'approche région.

Dans **l'approche contour**, l'objectif est de détecter et d'isoler les contours des objets d'intérêt dans l'image. Le résultat de cette méthode est généralement une représentation sous la forme d'un ensemble de chaînes de pixels formant les contours, et des traitements supplémentaires sont souvent nécessaires pour associer ces contours aux objets correspondants.

L'autre approche consiste **à identifier des régions de pixels** homogènes à l'intérieur de l'image. Cette méthode se base sur des critères d'homogénéité tels que l'intensité, la couleur ou même la texture locale. Le résultat peut être présenté sous la forme d'une image binaire, où les pixels appartenant à une région sont mis en évidence, ou sous la forme d'une image étiquetée, où chaque région est identifiée par une étiquette spécifique.

4.4.3 Le choix d'une technique de segmentation

Le choix est lié à :
— La nature de l'image (éclairage, contours, texture, etc.).
— Aux opérations en aval de la segmentation (compression, reconnaissance des formes, Mesures, etc.).
— Aux primitives à extraire (droites, régions, textures, etc.).
— Aux contraintes d'exploitation (temps réel, espace mémoire, etc.).

4.4.4 Objectifs de la segmentation

— Fournir des régions homogènes (selon un critère donné).
— Localiser de manière précise les contours des régions.
— L'étude et l'interprétation des structures anatomiques.
— Réduction de bruit.

4.4.5 Implémentation et résultats

La segmentation d'image par classification ("en utilisant la méthode FCM") fournie une partition de l'image en regroupant des pixels ayant des niveaux de gris similaires dans une même classe de pixels. Par conséquent, nous donnons les résultats des segmentations effectués sur un ensemble d'image en niveaux de gris. Tout d'abord on a choisi deux images en niveaux

de gris fréquemment utilisées pour l'évaluation des différents algorithmes du domaine de traitement d'image (voir les Figures 4.5 et 4.10).

Image 01 (`Baboon`)

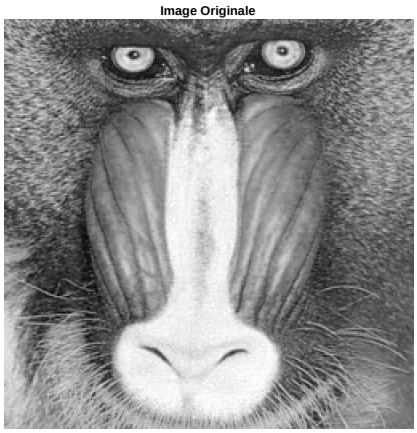

FIGURE 4.5 – Image en niveaux de gris.

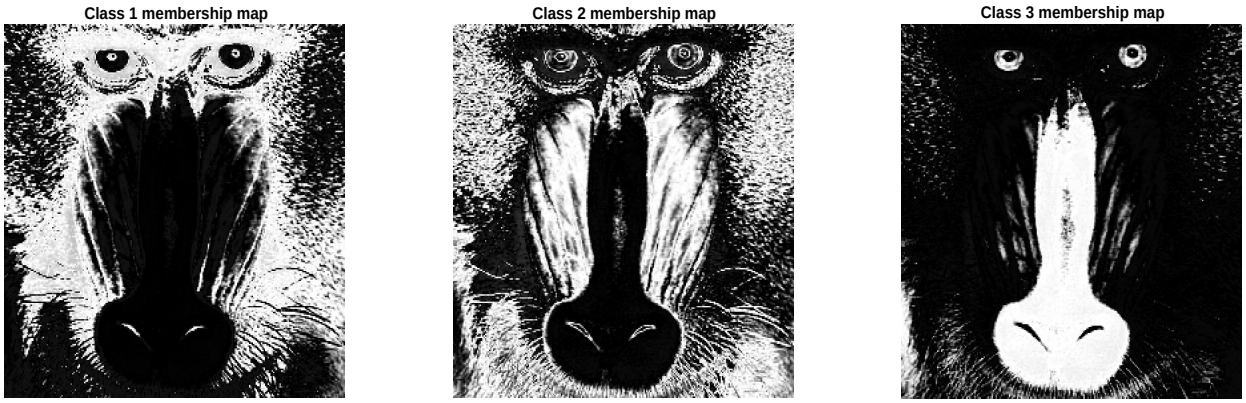

FIGURE 4.6 – Image segmentée par FCM (c = 3).

Les résultats de la segmentation de l'image 1 (image originale du Baboon, Figure 4.5) avec un nombre de cluster égale à 3 (c=3) sont illustrés dans la Figure 4.6. L'application de l'algorithme FCM a donc pour but de partitionner l'image en trois classes en regroupant des pixels ayant des niveaux de gris similaire dans une même classe. Les trois différentes classes de pixels sont illustrées respectivement dans les sous-figures (a), (b), et (c) de la Figure 4.6.

Les pixels de faible intensité correspondent aux parties sombres ou bien noir du visage du Baboon sont capturées dans le premier cluster (sous-figures (a)). Tandis que les pixels de forte intensité correspondent aux parties claires ou bien blanches du visage du Baboon sont présentes dans le troisième cluster (sous-figures (c)). Les nuances de

moyenne intensité correspondent aux parties grisées de son visage sont prises par le deuxième cluster (sous-figures (b)).

Afin de bien montrer les résultats de la segmentation dans chaque classe comme illustré dans la figure, les différentes parties du visage du Baboon (image originale) capturé par chaque cluster sont colorées en blanc (codé par 1), le reste des pixels de l'image non capturé par le cluster sont pondérés par 0, c'est-à-dire coloré en noir. Autrement dit, tout ce qui est retenu dans chaque cluster est coloré en blanc, tandis que le reste des pixels non capturés par le cluster en question sont coloré en noir (voir les sous-figures (a), (b), et (c) de la Figure 4.6).

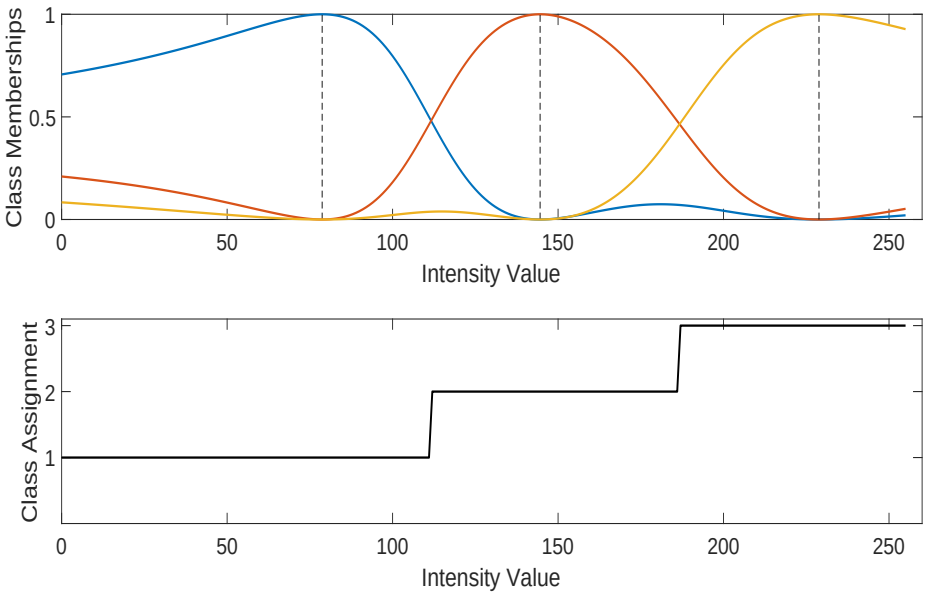

FIGURE 4.7 – Fonctions d'appartenances pour c = 3.

La Figure 4.7 montre la forme des fonctions d'appartenances associes aux trois classes de pixels. La plage de variation des intensités de pixels pour chaque classe est aussi présenté. A titre d'exemple, les pixels de l'image qui ont une intensité qui varie entre 0 et environ 110 appartient au premier cluster avec un degré d'appartenance plus élevé (proche de 1) par rapport aux autres pixels de l'image. Le deuxième cluster comporte les pixels ayant une intensité moyenne variante entre 110 et 186. Pour les nuances du dernier cluster, elles appartiennent à l'intervalle $[187 - 255]$.

Les résultats de la segmentation de la même image (Baboon de la Figure 4.5) avec un nombre de cluster égale à 4 (c=4) sont illustrés dans la figure 4.8. Les quatre différentes classes de pixel sont illustrées respectivement dans les sous-figures (a), (b), (c), et (d) de la Figure 4.8. Cependant que les pixels de faible intensité correspondent aux parties sombres ou bien noir du visage du Baboon sont capturées dans le premier cluster (sous-figures (a)). Tandis que les pixels de forte intensité correspondent aux

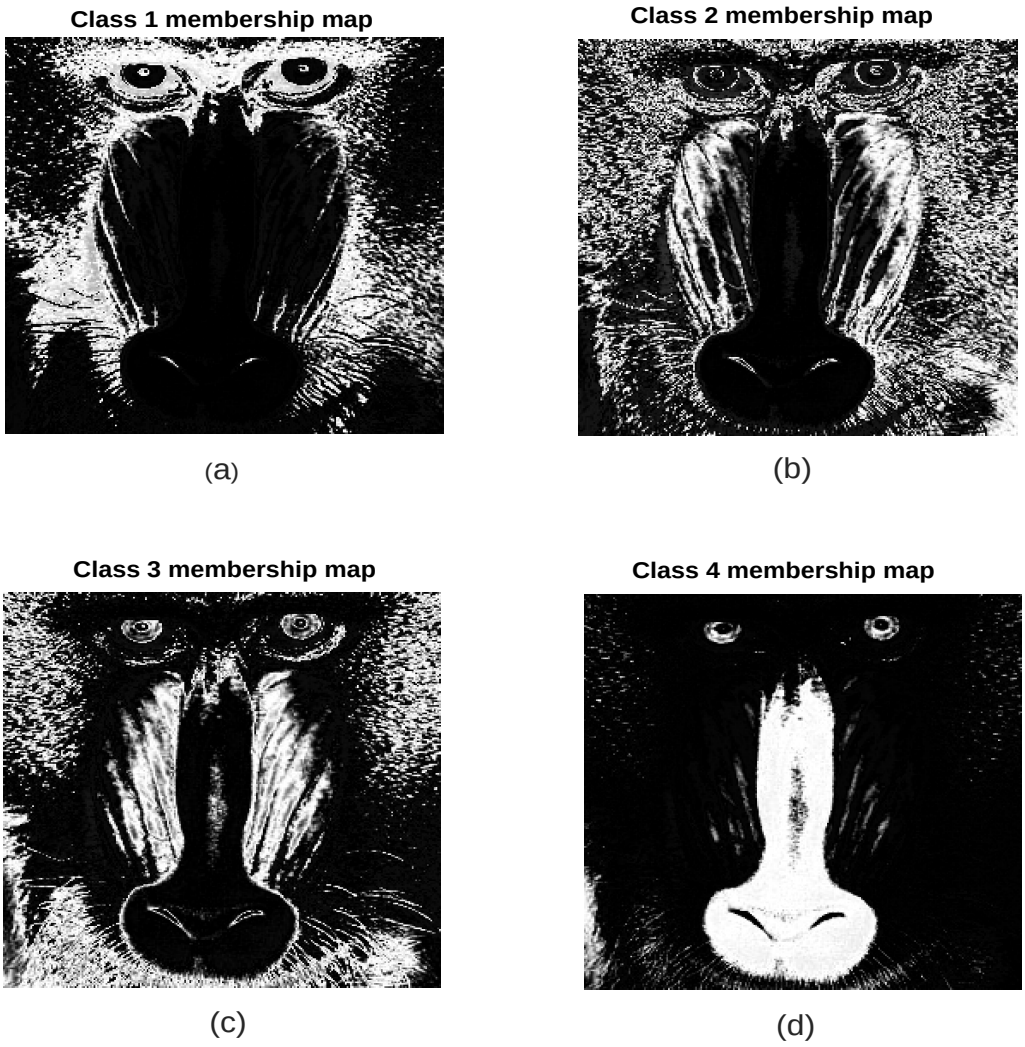

FIGURE 4.8 – Image segmentée par FCM (c = 4).

parties claires ou bien blanches du visage du Baboon sont présentes dans le quatrième cluster (sous-figures (d)). Les nuances les plus foncés correspondent aux parties proches du sombre de son visage sont prises par le deuxième cluster (sous-figures (b)). Alors que les nuances plus ou moins claires correspondent aux parties proches du blanc sont prises par le troisième cluster (sous-figures (c)).

Pour raison d'affichage des résultats de la segmentation dans chaque classe comme illustré dans la figure, les différentes parties du visage du Baboon (image originale) capturé par chaque cluster sont colorées en blanc, le reste des pixels de l'image non capturé par le cluster sont pondérés par 0, c'est-à-dire coloré en noir.

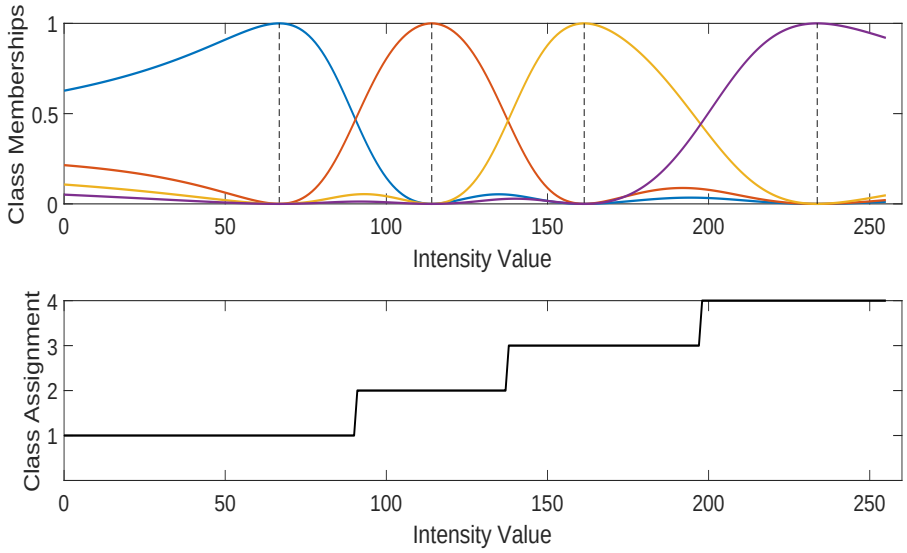

FIGURE 4.9 – Fonctions d'appartenances pour c = 4.

La Figure 4.9 présente les fonctions d'appartenance associées aux quatre classes de pixels de l'image segmanté, ainsi que la plage de variation des intensités de pixels pour chaque classe. Les pixels de l'image dont l'intensité varie entre 0 et environ 90 appartiennent au premier cluster avec un degré d'appartenance plus élevé (proche de 1) par rapport aux autres pixels de l'image. Le deuxième cluster regroupe les pixels ayant une intensité foncée variant de 90 à 137. Le troisième cluster comprend les pixels ayant une intensité plus ou moins claire variant de 137 à 197. Quant aux nuances du dernier cluster, elles appartiennent à l'intervalle [197, 255].

Image 02 (`Cameraman`)

Les résultats de la segmentation de l'image 2 (image du cameraman, Figure 4.10) avec un nombre de clusters égal à 2 (c=2) sont illustrés dans la Figure 4.11. Les deux classes de segmentation sont représentées respectivement dans les sous-figures (a) et (b) de la Figure 4.11. La Figure 4.12 présente les fonctions d'appartenance associées aux deux

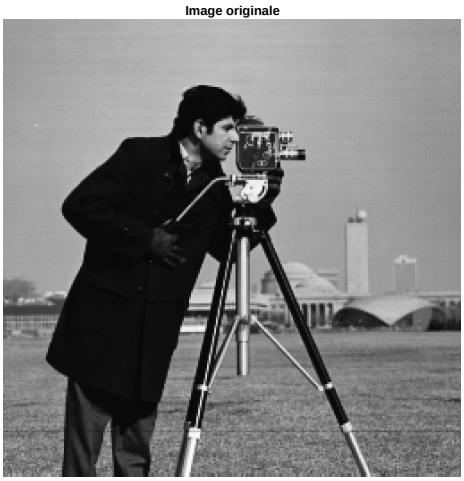

FIGURE 4.10 – Image en niveaux de gris.

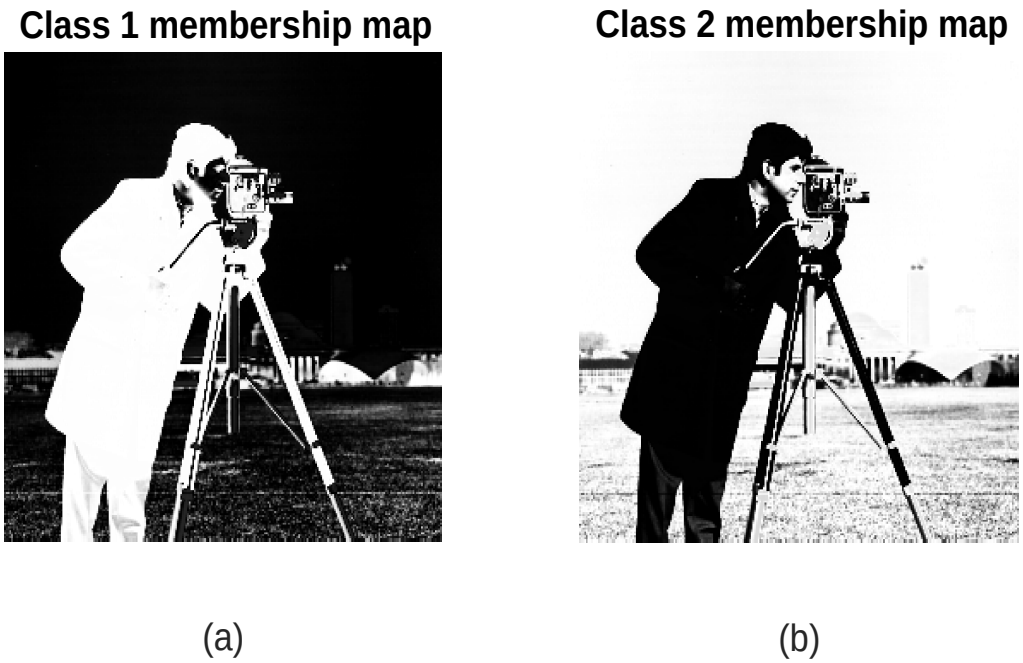

FIGURE 4.11 – Image segmentée par FCM (c = 2).

classes de pixels, ainsi que la plage de variation des intensités de pixels pour chaque classe. Les pixels de l'image dont l'intensité varie entre 0 et environ 89 appartiennent au premier cluster avec un degré d'appartenance plus élevé (proche de 1) par rapport aux autres pixels de l'image. Quant aux nuances du deuxième cluster, elles appartiennent à l'intervalle $[90, 255]$.

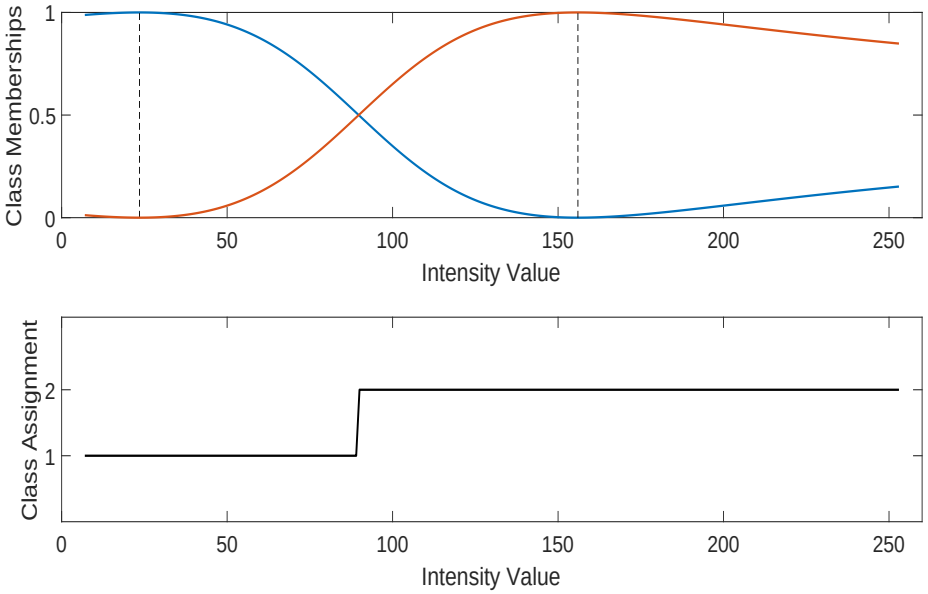

FIGURE 4.12 – Fonctions d'appartenances pour c = 2.

Les résultats de la segmentation de la même image (cameraman) avec un nombre de cluster égale à 3 (c=3) sont illustrés dans la Figure 4.13. Les trois différentes classes de pixel sont illustrées respectivement dans les sous-figures (a), (b), et (c), de la Figure 4.13.

La Figure 4.14 présente les fonctions d'appartenance associées aux trois classes de pixels, ainsi que la plage de variation des intensités de pixels pour chaque classe. Les pixels de l'image dont l'intensité varie entre 0 et environ 69 appartiennent au premier cluster avec un degré d'appartenance proche de 1 par rapport aux autres pixels de l'image. Le deuxième cluster regroupe les pixels ayant une intensité foncée variant de 90 à 143. Quant aux nuances du dernier cluster (de forte intensité), elles appartiennent à l'intervalle $[144, 255]$.

En effet, les objets sombres ou noires de l'image du cameraman, correspondant aux pixels de faible intensité, sont capturées dans le premier cluster (sous-figure (a)). En revanche, les objets claires ou blanches de l'image, représentées par les pixels de forte intensité, sont présentes dans le troisième cluster (sous-figure (c)). Les nuances de moyenne intensité, correspondant aux objets grisés de l'image, sont regroupées dans le deuxième cluster (sous-figure (b)).

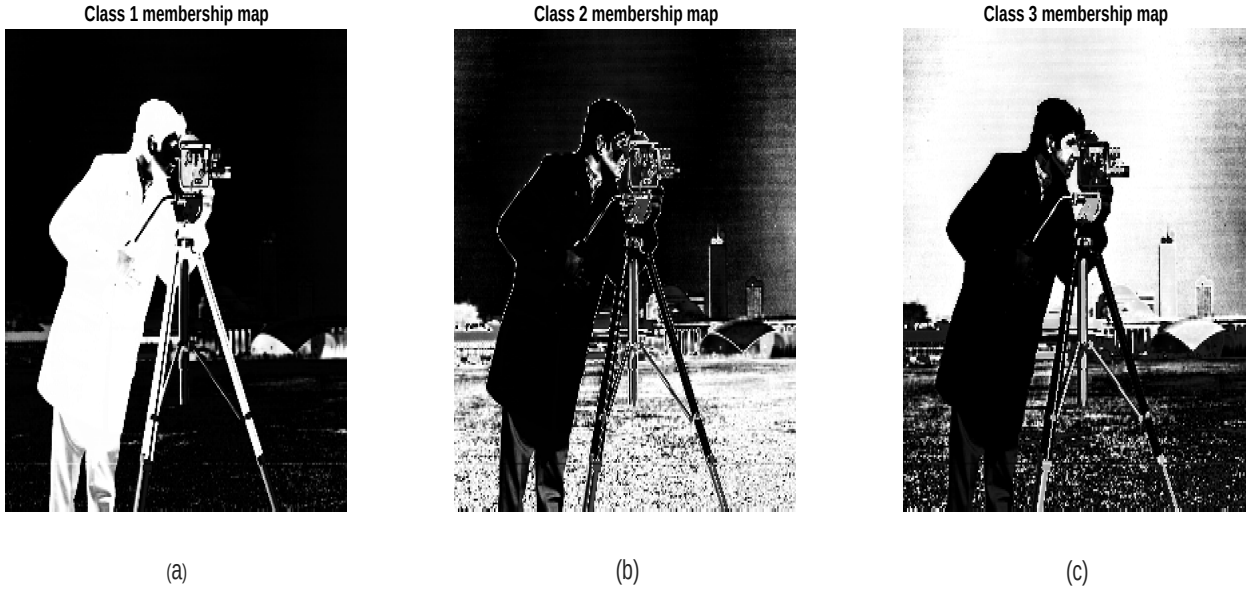

FIGURE 4.13 – Image segmentée par FCM (c = 3).

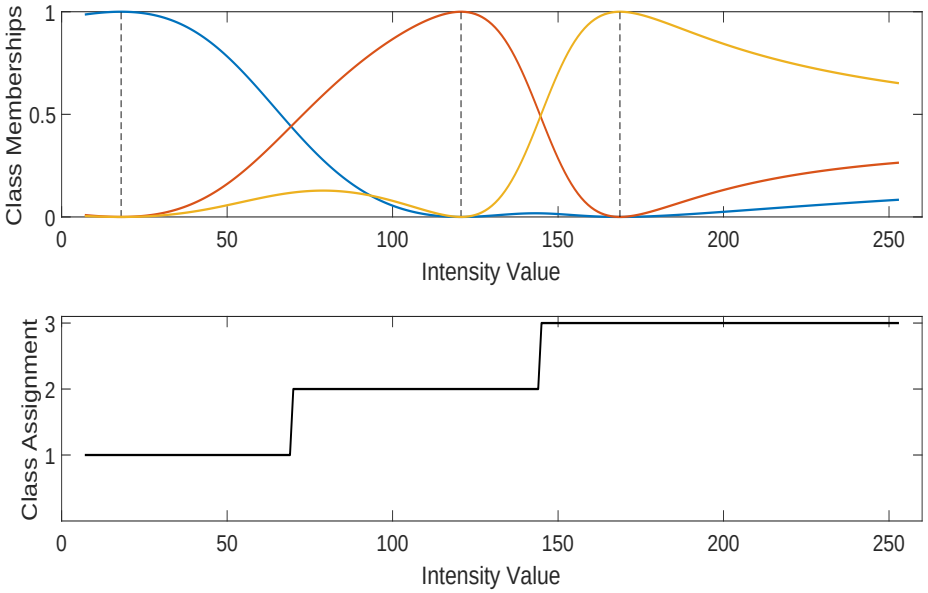

FIGURE 4.14 – Fonctions d'appartenances pour c = 3.

4.5 Conclusion

Ce chapitre porte sur la segmentation des images en niveaux de grés en utilisant l'environnement Software Matlab. Dans une première partie du chapitre, on a présenté les éléments de base de l'environnement Matlab et ses différentes fonctionnalisées. Ainsi qu'un aperçu général sur le processus de segmentation d'image est introduit. La seconde partie a consisté à déterminer les résultats de segmentation d'un ensemble d'images par la méthode de classification non supervisé Fuzzy C-Means. Plusieurs expérimentations de segmentation des images en niveaux de gris (de baboon et celle du cameraman) avec différents nombres de clusters (c=2, 3 et 4) ont été réalisées. Les résultats se présente en général sous la forme d'un ensemble de classes de pixels.

D'après ses résultats, on déduit que FCM permet une segmentation floue où chaque pixel peut avoir un degré d'appartenance à plusieurs clusters. Cela permet une représentation plus précise des frontières entre les objets dans une image, car les pixels près des frontières peuvent avoir des degrés d'appartenance partagés entre plusieurs clusters.

Conclusion générale

L'apprentissage artificiel est une branche essentielle des mathématiques appliquées qui se situe à l'intersection des statistiques et de l'intelligence artificielle. Son objectif principal est de développer des modèles capables d'apprendre à partir d'exemples. Contrairement aux modèles basés sur des connaissances préexistantes, qui reposent sur des équations dérivées des principes fondamentaux de la physique, de la chimie, de la biologie, de l'économie, etc., l'apprentissage artificiel se fonde sur des données numériques issues de mesures ou de simulations.

L'apprentissage statistique se révèle extrêmement utile lorsqu'il s'agit de modéliser des processus complexes, pour lesquels les connaissances théoriques sont trop approximatives pour permettre des prédictions précises. Ses domaines d'application sont nombreux et variés, tels que la fouille de données, la bio-informatique, le génie des procédés, l'aide au diagnostic médical, l'imagerie, les télécommunications, et bien d'autres encore.

Le clustering est une méthode d'apprentissage statistique qui consiste à regrouper des points de données par similarité ou par distance. C'est une méthode d'apprentissage non supervisée. Pour un ensemble donné de points, le clustering est utilisé pour classer ces points de données individuels dans des groupes spécifiques. En conséquence, les points de données d'un groupe particulier présentent des propriétés similaires. Dans le même temps, les points de données de différents groupes ont des caractéristiques différentes.

Ce projet de fin d'étude consiste en l'approfondissement de l'analyse en clusters des données réelles en utilisant la méthode des k-moyennes et la méthode C-moyenne floue. Nous avons appliqué la méthode FCM pour la segmentation des images grises sous le software Matlab. Les résultats obtenus montrent l'efficacité de la méthode FCM pour la segmentation floue des images, dont chaque pixel de l'image peut avoir un degré d'appartenance à plusieurs clusters. Cela permet une représentation plus précise des frontières entre les objets dans une image, car les pixels proches des frontières peuvent avoir des degrés d'appartenance partagés entre plusieurs clusters.

À la suite des résultats obtenus, plusieurs perspectives s'ouvrent à nous. Une perspective majeure est l'application de la méthode FCM (Fuzzy C-Means) pour la segmentation d'images en couleur.

Bibliographie

[1] Jean-Luc Raffaëlli, Médéric Morel, Pirmin P. Lemberger, Marc Batty. Big data et machine learning : manuel du data scientist. InfoPro. Management des systèmes d'information. Dunod (2015). (Cité en page 1.)

[2] Dunn J. A fuzzy relative of the ISODATA process and its use in detecting compact well separated clusters [J]. Journal of Cybernetics, 1973, 3 : 32-57. (Cité en page 1.)

[3] Bezdek J C. Pattern recognition with fuzzy objective function algorithms [J]. 1981. (Cité en page 1.)

[4] Macqueen J. Some methods for classification and analysis of multivariate observations[C]. Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, F, 1967. Oakland, CA, USA. (Cité en pages 1 et 18.)

[5] Al-sultan K S, Selim S Z. A global algorithm for the fuzzy clustering problem [J]. Pattern Recognition, 1993, 26(9) : 1357-1361. (Cité en page 1.)

[6] Al-sultan K S, Fedjki C A. A tabu search-based algorithm for the fuzzy clustering problem [J]. Pattern Recognition, 1997, 30(12) : 2023-2030. (Cité en page 1.)

[7] Hathaway R J, Bezdek J C. Optimization of clustering criteria by reformulation [J]. IEEE Transactions on Fuzzy Systems, 1995, 3(2) : 241-245. (Cité en page 1.)

[8] Hall L O, Ozyurt I B, Bezdek J C. Clustering with a genetically optimized approach [J]. IEEE Transactions on Evolutionary computation, 1999, 3(2) : 103-112. (Cité en page 1.)

[9] Runkler T A, Bezdek J C. Alternating cluster estimation : A new tool for clustering and function approximation [J]. IEEE Transactions on Fuzzy Systems, 1999, 7(4) : 377-393. (Cité en page 1.)

[10] Rokach L, Maimon O. Clustering methods [J]. Data Mining and Knowledge Discovery Handbook, 2005, 3(3) : 321-352. (Cité en page 9.)

[11] Gan G, Ma C, Wu J. Data clustering : theory, algorithms and applications [M]. Philadelphia : SIAM, 2007. (Cité en page 12.)

[12] Yu Z, Au O C, Zou R, et al. An adaptive unsupervised approach toward pixel clustering and color image segmentation [J]. Pattern Recognition, 2010, 43 : 1889-1906. (Cité en page 12.)

[13] Zarinbal M, Zarandi M F, Turksen I. Relative entropy fuzzy c-means clustering [J]. Information Sciences, 2014, 260 : 74-97. (Cité en page 14.)

[14] Gong M, Liang Y, Shi J, et al. Fuzzy c-means clustering with local information and kernel metric for image segmentation [J]. IEEE Transactions on Image Processing, 2012, 22(2) : 573-584. (Cité en page 14.)

[15] Ji Z, Xia Y, Sun Q, et al. Adaptive scale fuzzy local Gaussian mixture model for brain MR image segmentation [J]. Neurocomputing, 2014, 134 : 60-69. (Cité en page 14.)

[16] Jacobs D, Weinshall D, Gdalyahu Y. Classification with nonmetric distances : image retrieval and class representation [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2000, 22 : 583-600. (Cité en page 14.)

[17] Zadeh L. Fuzzy sets [J]. Information and Control, 1965, 8(3) : 338-353. (Cité en pages 14, 21 et 22.)

[18] Pal N R, Bezdek J C. On cluster validity for the fuzzy c-means model [J]. IEEE Transactions on Fuzzy systems, 1995, 3(3) : 370-379. (Cité en page 26.)

[19] Torra V. On the selection of m for Fuzzy c-Means[C]. Proceedings of the 2015 Conference of the International Fuzzy Systems Association and the European Society for Fuzzy Logic and Technology (IFSA-EUSFLAT-15), Atlantis Press, F, 2015. (Cité en page 26.)

[20] Misza Kalechman. PRACTICAL MATLAB BASICS FOR ENGINEERS. International Standard Book Number-13 : 978-1-4200-4774-5 (Softcover), 2009. (Cité en page 29.)

[21] Brunton, Steven L. (Steven Lee), Kutz, Jose Nathan. Data-driven science and engineering : machine learning, dynamical systems, and control. Cambridge University Press, 2019.

(Cité en page 16.)