ÈRE DE L'ENSEIGNEMENT SUPÉRIEUR ET DE LA RECHERCHE SCIENTIFIQUE

JNIVERSITÉ ABDELHAMID IBN BADIS - MOSTAGANEM

**Faculté des Sciences Exactes et d'Informatique**

**Département de Mathématiques et informatique**

**Filière : Informatique**

RAPPORT DE PROJET DE FIN D'ÉTIDE

Option : **Ingénierie des Systèmes d'Information**

THÈME :

# *Le clustering multicritère ordonné: une approche basée sur la détection des anomalies*


Étudiant(e) : BLIDI nour El houda

BENSLIM Izdihar

Encadrant(e) : ROUBA Baroudi


***Année Universitaire 2022-2023***

# Remerciement

La réalisation de ce mémoire a été possible grâce à l'aide de plusieurs personnes à qui nous voudrions témoigner toute nos gratitudes.

Nous voudrions tout d'abord adresser toute notre reconnaissance au directeur de ce mémoire, Monsieur ROUBA BAROUDI pour sa patience, sa disponibilité et surtout ses judicieux conseils, qui ont contribués à alimenter nos réflexions.

Nous désirons aussi remercier les professeurs de l'université LINES, qui nous fournis les outils nécessaires à la réussite de nos études universitaires.

Nos parents, pour leurs soutiens constants et leurs encouragements.

# Résumé

Le clustering ordonné est une technique permettant de classer automatiquement des objets dans des clusters ordonnés. Les techniques du clustering ordonné permettent, en plus de la création des clusters, de définir une relation d'ordre entre ces clusters (les clusters sont classés du meilleur au moins bon). L'objectif du projet est de développer une approche de clustering ordonné dans un contexte décisionnel multicritère. Cette approche est basée sur la détection des objets dits « outliers » (anomalies ou valeurs aberrantes). La détection de ces objets permet de définir les frontières entre les différents clusters. Afin de prendre en considération le caractère multicritère du problème, la méthode multicritère PROEMTHEE est utilisée. Cette méthode est caractérisée par la notion du flux-net qui permet de définir une relation d'ordre entre les objets. Cette relation est réutilisée pour définir un ordre sur les clusters.

**Mot clés :** clustering, aide multicritère à la décision, clustering multicritère ordonné, la détection d'anomalies, test de normalité

# Abstract

Ordered clustering is a technique for automatically classifying objects into ordered clusters. The techniques of ordered clustering allow, in addition to the creation of clusters, to define an order relationship between these clusters (clusters are ranked from best to worst). The objective of the project is to develop an orderly clustering approach in a multi-criteria decision-making context. This approach is based on the detection of so-called "outliers" objects (anomalies or outliers). Detecting these objects helps define the boundaries between the different clusters. In order to take into account the multi-criteria nature of the problem, the PROEMTHEE multicriteria method is used. This method is characterized by the notion of net flow which makes it possible to define an order relationship between objects. This relationship is reused to set an order on clusters.

**Key word:** clustering, multi-criteria decision aid, ordered multi-criteria clustering, anomaly detection, normality test

# Index des figures

# Index des tableaux

# Table des matières

# Introduction Générale

Beaucoup de choses qui nous entourent peuvent être catégorisées. Pour être moins vague et plus précis, nous avons des regroupements qui peuvent être binaires ou plus de deux.

À l'époque, les scientifiques catégorisent les populations en utilisant leurs étiquettes prédéterminées. Ainsi, tout individu répondant à certaines caractéristiques est placé manuellement dans la catégorie des individus ayant les mêmes caractéristiques. Cette façon de procéder a rapidement atteint ses limites avec la volonté d'étudier des jeux de données toujours plus volumineux provenant de sources diverses. Cela a conduit à l'émergence de nouvelles techniques de classification telles que le clustering.

Afin de permettre une meilleure interprétation des résultats d'un clustering, on a recourt, dans certains cas, à étudier les éventuelles relations qui peuvent exister entre les clusters. Ces relations peuvent être des relations de préférence, d'incomparabilité ou des relations d'ordre.

Le présent travail s'intéresse à la construction des clusters ordonnés dans un contexte décisionnelle multicritère. L'objectif n'est pas seulement de construire des clusters mais aussi de définir une relation d'ordre entre eux. Pour répondre à ces objectifs, trois problèmes doivent être résolus :

1- Comment prendre en considération du caractère multicritère du problème lors de la construction des clusters ?
2- Comment construire les clusters ?
3- Comment définir une relation d'ordre entre les clusters ?

Pour prendre en considération le caractère multicritère du problème, nous avons utilisé la méthode multicritère PROMETHEE connue pour sa force et sa simplicité.

Pour construire les clusters, nous avons fait appel à une technique statistique de détection des anomalies. En fait, nous nous sommes basés sur le fait que les objets appartenant à un même cluster suivent une loi normale. Nous avons, ainsi, construit des clusters par un processus itératif, permettant d'ajouter, à chaque itération, un objet à un cluster. Le processus s'arrête dès qu'un objet outlier est détecté.

Pour la définition de la relation d'ordre, nous avons réutilisé le concept du flux-net de la méthode PROMTHEE. Ce flux-net peut être utilisé pour classer les objets du meilleur au moins bon.

Ce rapport se compose de cinq chapitres, dans le premier chapitre, nous donnons d'abord une brève introduction à la classification, en nous concentrant sur la classification non supervisée (clustering). Nous introduisons les concepts et techniques de base de la classification non supervisée. Dans le deuxième chapitre, nous présenterons quelques concepts importants pour liés à l'aide multicritère à la décision. Le troisième chapitre est consacré à la présentation des concepts de base du clustering en aide multicritère à la décision (AMCD), le clustering multicritère ordonné, et les différentes approches développées dans ce domaine. Dans le quatrième chapitre, nous présenterons la détection des anomalies et les méthodes appliquées dans ce domaine. Dans le cinquième chapitre, nous présenterons une approche de clustering multicritère ordonné basée sur la détection des anomalies à l'aide de la méthode PROMETHEE2 et du test de Shapiro-Wilk.

# CHAPITRE 1: CLUSTERING

# 1. Introduction

La classification et comme toutes les méthodes d'analyse de données, a pour objectif de répartir un ensemble de données en groupes homogènes de telle façon que chaque groupe est bien différencié des autres. Ces données représentent des observations, des objets ou des individus décrits par des descripteurs (attributs, variables...). La classification peut être divisée en deux grandes catégories : la classification supervisée où le modèle est construit à partir de données étiquetées. La classification supervisée sert de de construire un modèle qui peut être utilisé pour prédire la classe des objets non étiquetés. Dans la classification non supervisée, les données ne sont pas étiquetées, l'objectif est de trouver des structures cachées dans les données, telles que les groupes ou les clusters.

Dans ce chapitre, nous allons présenter une description détaillée du clustering. Nous abordons par la suite les méthodes de clustering.

# 2. Clustering

Le clustering est le processus qui consiste à diviser les points de données en un certain nombre de groupes de sorte que les points de données des mêmes groupes soient plus similaires aux autres points de données du même groupe et différents des points de données des autres groupes [32].

L'exemple de la figure 1 représente un ensemble de données qui peuvent être regroupées en quatre groupes (clusters).

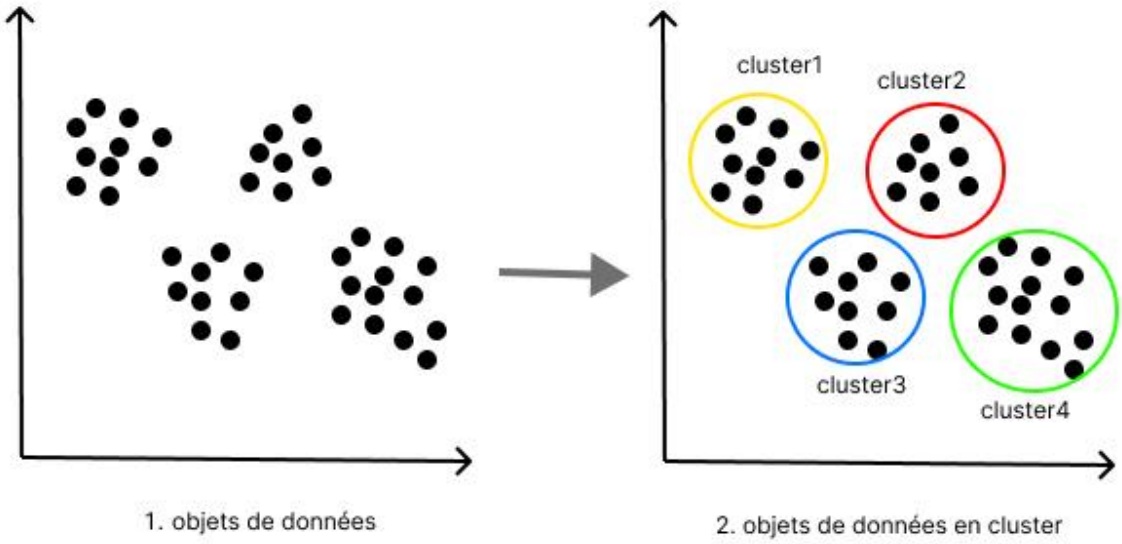


*Figure 1: Exemple de clustering*

## 3. Domaines d'application du clustering

Voici quelques domaines d'application du clustering :

**Marketing :** il peut être utilisé pour caractériser et découvrir des segments de clientèle à des fins de marketing.

**Biologie :** Il peut être utilisé pour la classification entre différentes espèces de plantes et d'animaux.

**Bibliothèques :** Il est utilisé pour regrouper différents livres sur la base de sujets et d'informations.

**Assurance :** il est utilisée pour reconnaître les clients, leurs polices et identifier les fraudes.

## 4. Les étapes d'un processus de clustering

Un processus de clustering généralement comporte les étapes suivantes [1] :

- **Préparation des données :** Charger et nettoyer les données, en éliminant les données manquantes et en normalisant les variables si nécessaire.
- **Définition des critères de similarité ou de distance :** Déterminer la mesure à utiliser pour évaluer la proximité entre les points de données. Cela peut inclure la définition d'une distance Euclidienne, de la similarité de Jaccard, etc.[16]
- **Sélection du nombre de clusters :** fixer le nombre de clusters souhaité.
- **Application de l'algorithme de clustering :** Appliquer l'algorithme de clustering choisi (par exemple, K-means, DBSCAN, agrégation hiérarchique, etc.) aux données préparées en utilisant les critères de similarité ou de distance définis.
- **Evaluation des clusters :** Evaluer la qualité des clusters formés en utilisant des métriques telles que le coefficient de silhouette, l'indice de Davies-Bouldin, etc. [3]
- **Interprétation et visualisation des résultats :** Interpréter les résultats et les visualiser à l'aide de graphiques et de tableaux pour une analyse plus approfondie.

Ces étapes peuvent varier en fonction du type d'algorithme de clustering utilisé et des spécificités des données, mais représentent un processus généralement suivi pour un clustering réussi.

## 5. Similarité, dissimilarité et distance

La similarité, dissimilarité et distance sont des concepts importants en analyse de données et en clustering. Elles sont utilisées pour mesurer la proximité entre les points de données et pour déterminer comment les groupes de points de données sont formés [32].

- **Similarité :** La similarité mesure à quel point deux objets ou deux instances sont similaires entre eux. Cela peut être mesuré en utilisant des mesures telles que la similarité de Jaccard [22] ou la similarité cosinus [12].
- **Dissimilarité :** La dissimilarité mesure à quel point deux instances sont différentes. Elle est souvent utilisée en tant que complément inverse de la similarité.
- **Distance :** La distance est une mesure de la dissimilarité entre deux objets. Elle est généralement définie comme la longueur d'un vecteur reliant deux points dans cet espace. En d'autres termes, la distance est une mesure quantitative de la dissimilarité. Il existe plusieurs types de distances, notamment la distance Euclidienne, la distance de Manhattan et la distance de Mahalanobis.

Il est important de choisir une mesure de similarité ou de distance appropriée en fonction de la nature des données et des objectifs de clustering pour obtenir des résultats significatifs.

## 6. Méthodes de clustering

Les algorithmes de clustering permettent de diviser les données en sous-groupes ou clusters de manière non supervisée. Ces sous-groupes regroupent des observations similaires. L'élément de base de tout algorithme de clustering est une mesure de proximité, de dissimilarité ou de distance. Les algorithmes de clustering peuvent être divisés en plusieurs classes. Nous en présentons ci-dessous les principales d'entre elles.

### 6.1 Méthodes basées sur la hiérarchie

Les clusters formés dans cette approche forment une structure arborescente basée sur la hiérarchie. Le processus de clustering est exécuté de façon itérative. Chaque itération permet de former de nouveaux clusters en utilisant des clusters précédemment formés. On distingue deux approches de clustering hiérarchique :

- Approche agglomérative (approche ascendante).
- Approche Divisive (approche descendante).

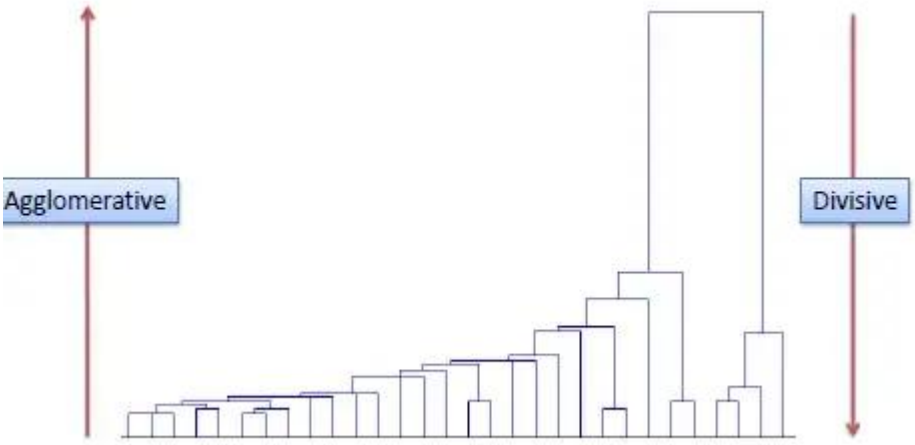


*Figure 2: Types de clustering hiérarchique*

### 6.1.1 Approche descendante

Elle consiste à commencer avec chaque point de données comme un cluster individuel, puis à combiner les clusters les plus proches jusqu'à ce qu'il n'en reste plus qu'un seul.

### 6.1.2 Approche ascendant

Elle consiste à commencer avec tous les points de données comme un seul grand cluster, puis à scinder répétitivement ce cluster en sous-groupes plus petits jusqu'à ce que chaque sous-groupe ne contienne plus qu'un seul point.

## 6.2 Méthodes de partitionnement

Les méthodes de partitionnement font partie des trois familles d'outils d'analyse non supervisée les plus répandues avec la classification ascendante hiérarchique (CAH) et les méthodes d' estimation de densité.

### 6.2.1 Algorithme K-means

L'algorithme le plus classique est l'algorithme des k-moyennes. Il ne nécessite qu'un seul choix de départ qui est le nombre de classes voulues **(*k*)**.

K-means permet de regrouper des individus ayant des caractéristiques similaires à travers l'analyse de jeux de données caractérisée par un ensemble de descripteurs [26]. Cet algorithme fonctionne en selon les étapes suivantes :

1. Spécifier le nombre souhaité de clusters K
2. Attribuer au hasard chaque point de données à un cluster
3. Calculer les centroïdes de cluster et ré-affecter chaque point au centroïde du cluster le plus proche.
4. Re-calculer les centroides des clusters.

5. Répétez les étapes 3 et 4 pour un nombre défini d'itérations ou jusqu'à ce que les groupes ne changent plus.

Nous pouvons également choisir d'initialiser au hasard les centres de groupe plusieurs fois, puis de sélectionner l'exécution qui semble avoir fourni les meilleurs résultats.

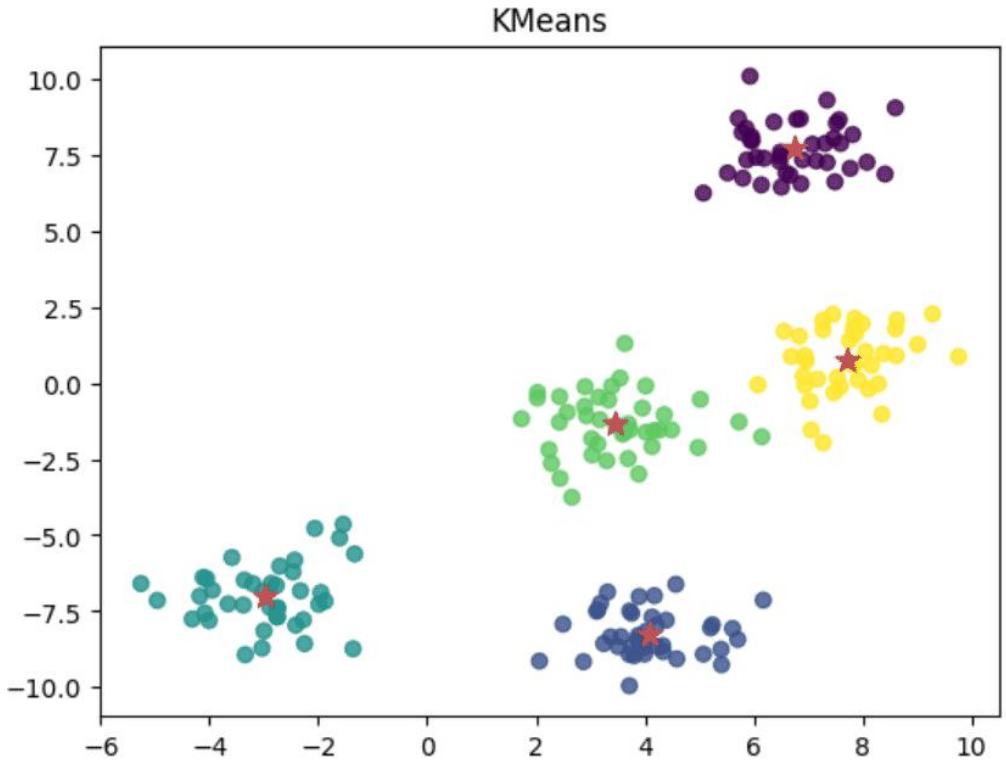


*Figure 3: Clustering avec k-means*

L'avantage de K-Means est qu'il est assez rapide, car tout ce que nous faisons vraiment est de calculer les distances entre les points et les centres de groupe, très peu de calculs, Il a une complexité linéaire O(n).

### 6.2.2 Algorithme K-medoids

K-medoids ou partitionnement autour de medoids est un algorithme de clustering légèrement modifié par rapport à l'algorithme K-means. En fait, c'est une variante de K-means. Il s'avère que le calcul des K-médoïdes est plus robuste au bruit que celui des K-means [24]. Cet algorithme fonctionne en selon les étapes suivantes :

1. Choisir la valeur du nombre de clusters k
2. L'algorithme part d'un ensemble de n points de données en sélectionnant arbitrairement k objets comme points centraux.
3. Après avoir sélectionné k-medoids, Chaque objet dans le jeu de données est associé au médoid le plus proche.

4. Sélectionner au hasard un objet non médoid O.
5. Calculer le coût total S de l'échange de l'objet initial par l'objet O.
6. Si S<0, échangez le médoïde initial avec le nouvel objet O (si S<0 alors il y aura un nouvel ensemble de médoïdes).
7. Répétez les étapes de 3 à 5 jusqu'à ce que nous n'ayons plus de changement de médoïdes.

### 6.3 Méthodes de statistique

La logique de ce type de clustering est de considérer que les objets qui sont générés à partir d'un mélange de k distributions de probabilité (k : nombre de clusters). Le but est d'estimer les paramètres permettant de définir chaque distribution afin de mieux s'adapter aux observations (objets). L'algorithme le plus couramment utilisé dans ce type de clustering est l'algorithme EM (Expectation Maximisation) [20]. L'algorithme est basé sur une méthode d'estimation gaussienne avec maximum de vraisemblance. Par rapport aux k-moyennes, l'algorithme EM présente les avantages suivants :

- Gestion de clusters serrés de tailles très variables.
- Gérer le l'éventuel bruit dans les données.

Cependant, l'algorithme EM caractérisé par une complexité quadratique, et il nécessite de nombreuses itérations avant de converger.

### 6.4 Méthodes basées sur la densité

Ces méthodes traitent les clusters comme des régions denses qui partagent des similitudes et des différences avec des régions moins denses de l'espace [23]. La méthode phare de cette catégorie est appelée DBSCAN « density-based spatial clustering of applications with noise ». L'algorithme DBSCAN prend deux paramètres en entrée [23] :
**Ɛ :** la distance maximale qui peut définir deux individus comme voisins.
**minPts:** détermine le nombre minimum de points requis pour former un groupe.
Les étapes d'exécution de cet algorithme sont les suivantes :

- Choisir un point au hasard dans le jeu de données et récupérer tous les points à l'intérieur d'un certain rayon (ε) autour de celui-ci.
- S'il y a au moins minPts points à l'intérieur du rayon, un nouveau groupe est formé et tous les points à l'intérieur du rayon sont ajoutés au groupe.
- Pour chacun des points récemment ajoutés, répéter le processus de récupération de tous les points à l'intérieur du rayon et les ajouter au groupe.

- Continuer le processus jusqu'à ce que tous les points à l'intérieur du rayon aient été ajoutés au groupe ou qu'il ne soit plus possible d'ajouter de points.
- Répéter le processus pour chaque point non visité dans le jeu de données jusqu'à ce que tous les points soient affectés à un groupe ou considérés comme du bruit (une anomalie) [27].

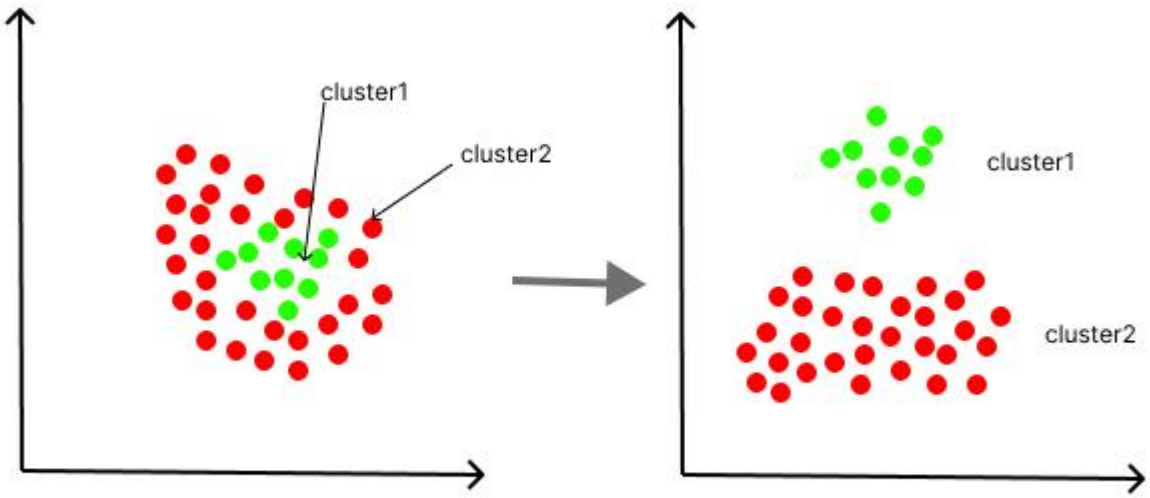


*Figure 4: DBSCAN clustering*

### 6.5 Méthodes basé sur les grilles

Les méthodes basées sur les grilles sont une famille d'algorithmes de clustering qui divisent l'espace de données en une grille de cellules et attribuent chaque point de données à une cellule spécifique. Ces méthodes sont efficaces pour traiter les grands jeux de données et sont souvent utilisées pour traiter les données à haute dimension. L'idée de base derrière ces méthodes est de diviser l'espace de données en une grille régulière de cellules et de regrouper les points qui tombent dans la même cellule. Les cellules avec un grand nombre de points sont considérées comme des régions denses et sont utilisées pour former des groupes [32]. Il existe plusieurs algorithmes de clustering basés sur les grilles, tels que STING (STatistical INformation Grid), CLIQUE (CLustering In QUEst) et BIRCH (Balanced Iterative Reducing and Clustering using Hierarchies). Ces algorithmes ont leurs propres forces et faiblesses et conviennent à différents types de données et de problèmes de clustering.

En général, les méthodes basées sur les grilles sont rapides et peuvent gérer les grands jeux de données, mais elles peuvent être sensibles au choix de la taille de la grille et ne peuvent pas être appropriées pour trouver des groupes de formes

arbitraires. Ils sont souvent utilisés comme étape de pré-traitement pour d'autres algorithmes de clustering ou pour la réduction de la dimensionnalité.

### 6.6 Méthode basé sur les graphes

Les méthodes basées sur les graphes représentent les données sous forme de nœuds dans un graphe et utilisent la théorie des graphes pour identifier des groupes de nœuds similaires. Ce type d'algorithme est particulièrement utile pour les données complexes et non linéaires. Les méthodes basées sur les graphes sont souvent combinées avec d'autres techniques pour une analyse plus approfondie. L'algorithme de Louvain, utilise une approche de partitionnement de graphe pour détecter les communautés dans un réseau [8].

### 6.7 Le clustering stochastique

Le clustering stochastique utilise des méthodes aléatoires pour effectuer le regroupement. Au lieu de suivre une démarche déterminée pour diviser les données en groupes, le clustering stochastique utilise des techniques aléatoires pour explorer les données et trouver des structures cachées. Ce type d'algorithme peut être utile pour des données complexes ou des situations où il y a plusieurs structures cachées dans les données. Le clustering stochastique est souvent combiné avec d'autres algorithmes pour une analyse plus approfondie. Par exemple l'algorithme de recuit simulé (simulated annealing en anglais). Cet algorithme est souvent utilisé pour résoudre des problèmes d'optimisation combinatoire complexes en explorant de manière stochastique l'espace des solutions [19].

## 7. Caractéristiques d'un algorithme de clustering

Comme tout algorithme, un algorithme de clustering doit répondre à certaines exigences, parmi lesquelles nous citons :

- Évolutivité.
- Gestion de différents types d'attributs.
- Découverte de clusters de forme arbitraire.
- Exigence minimale de connaissance du domaine pour déterminer les paramètres d'entrée.
- capacité à gérer le bruit et les anomalies (outliers), insensibilité à l'ordre des enregistrements d'entrée.

## 8. Conclusion

Dans ce chapitre, nous avons introduit les concepts de base du clustering et ces différentes méthodes existantes. Dans la suite de ce rapport, nous nous plaçons dans le contexte des processus de clustering dans le cadre de l'aide à la décision multicritère. Pour cela, nous donnons un aperçu du domaine de l'aide à la décision multicritères.

# CHAPITRE 2: AIDE MULTICRITERE À LA DÉCISION

# 1. Introduction

Dans ce chapitre, nous parlerons des concepts de base du domaine d'aide multicritère à la décision et nous présentons les principales problématiques et les méthodes de ce domaine.

# 2. Aide Multicritère à la décision

L'aide multicritère à la décision est un processus qui utilise un ensemble d'informations pour pouvoir formuler un problème et aboutir à une décision. « l'aide à la décision est l'activité de celui qui, prenant appui sur des modèles clairement explicités mais non nécessairement complètement formalisés, aide à obtenir des éléments de réponses aux questions que se pose un intervenant dans le processus de décision, éléments concourant à éclairer la décision et normalement à prescrire, ou simplement à favoriser un comportement de nature à accroître la cohérence entre l'évolution du processus d'une part, les objectifs et le système de valeurs au service desquels cet intervenant se trouve placé d'autre part» [6].

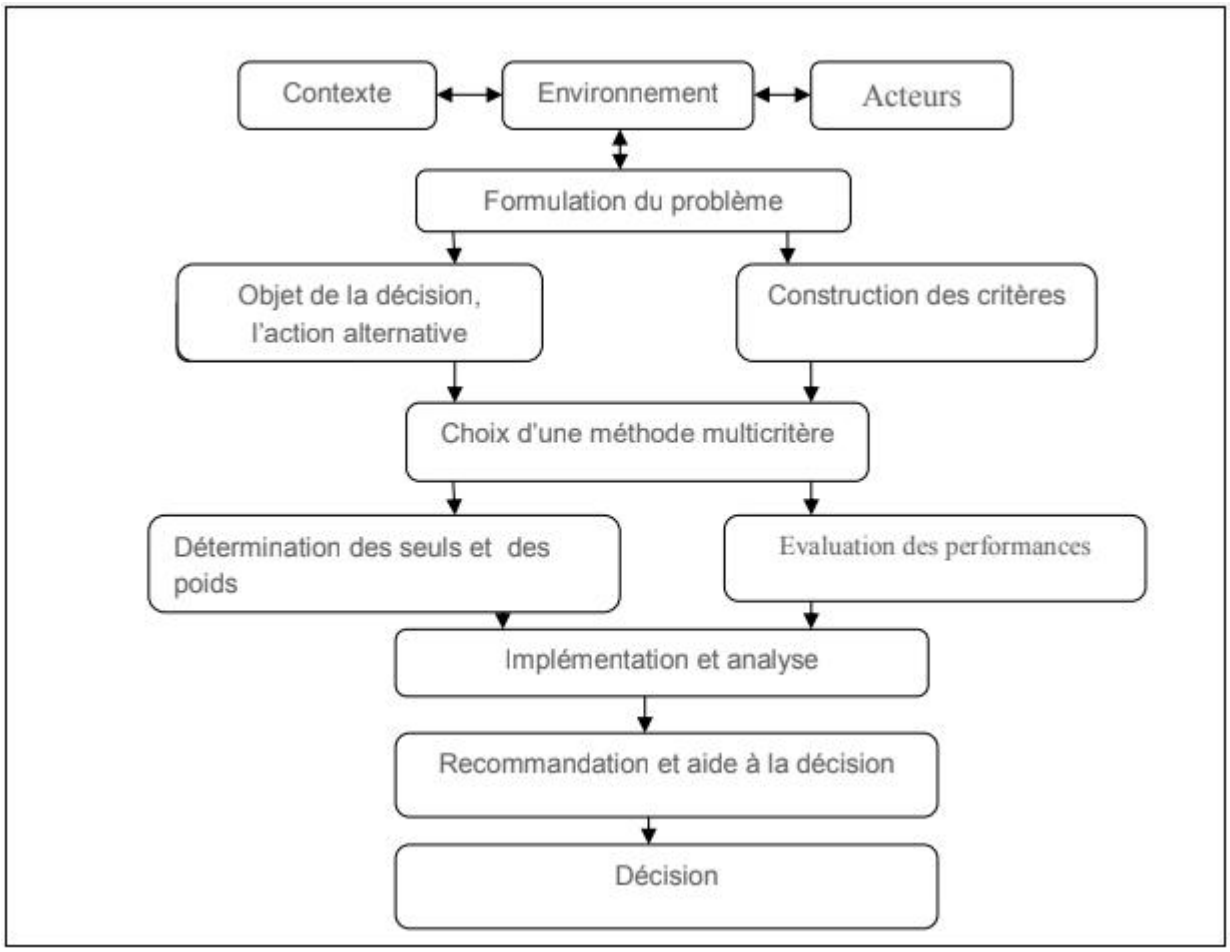


*Figure 5: Le processus d'aide multicritère à la décision[25].*

# 3. Concepts de base

## 3.1 Le concept de critère

Un critère est une fonction g à valeurs réelles définie sur l'ensemble des actions potentielles. Cette fonction est définie de telle sorte que deux alternatives $a_1$ et $a_2$ puissent être comparées en s'appuyant sur les valeurs $g(a_1)$ et $g(a_2)$.

## 3.2 Les actions potentielles

L'ensemble des actions potentielle noté ***X*** est l'ensemble des objets, solution, décisions, candidats, etc. que l'on va explorer dans le processus de décision.

## 3.3 Le tableau de performances

Un tableau des performances contient des actions de ***X*** dans les lignes et l'ensemble des critères ***G*** dans les colonnes.

*Tableau 1: tableau de performances.*

| | $g_1$ | $g_2$ | … | $g_i$ | … | $g_m$ |
|---|---|---|---|---|---|---|
| $x_1$ | | | | | | |
| $x_2$ | | | | | | |
| … | | | | | | |
| $x_i$ | | | | $g_i(x_i)$ | | |
| … | | | | | | |
| $x_m$ | | | | | | |

## 3.4 Le système relationnel de préférences

Afin de déterminer la décision finale en fonction de problème traité, un ensemble de relation de préférence est utilisé pour comparer deux actions $x_1$ et $x_2$ , le décideur ne doit choisir qu'une seule des 4 attitudes suivantes [32] :

- $x_1$ est préféré à $x_2$ . Noté $x_1$ **P** $x_2$ . **P** est irréflexive et asymétrique.
- $x_2$ est préféré à $x_1$ , noté $x_2$ **P** $x_1$ .
- $x_1$ est indifférent à $x_2$ . Notée $x_1$ **I** $x_2$ . **I** est une relation réflexive et symétrique.
- $x_1$ est incomparable à $x_2$ . Notée $x_1$ **R** $x_2$ . **R** est la relation irréflexive et symétrique.

# 4. Problématiques en aide multicritère à la décision

les problèmes réels peuvent être formulés à l'aide des méthodes d'analyse multicritère, selon trois formulations de bases : problématique de choix notée (P,$\alpha$) , la problématique de tri ou d'affectation notée (P,$\beta$) et la problématique de rangement noté (P,$\gamma$).

## 4.1 Problématique de choix (P,$\alpha$)

La décision par le choix d'un sous ensemble aussi restreint que possible en vue d'un choix final d'une seule action.

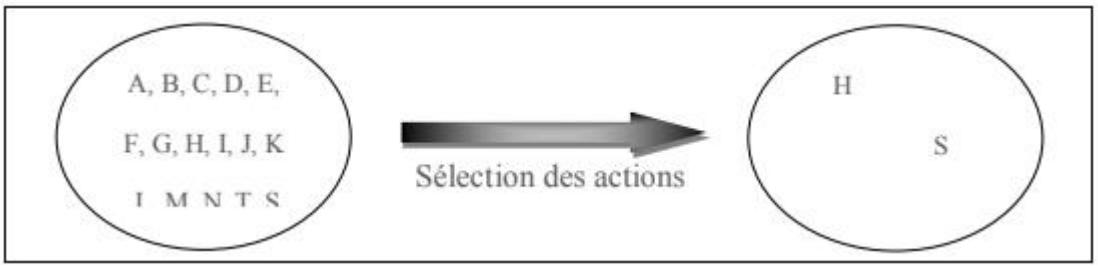


*Figure 6: problématique de choix [32]*

## 4.2 Problématique de tri (P,$\beta$)

La décision se traduit par l'affectation de chaque action à une catégorie, les catégories étant définies à priori.

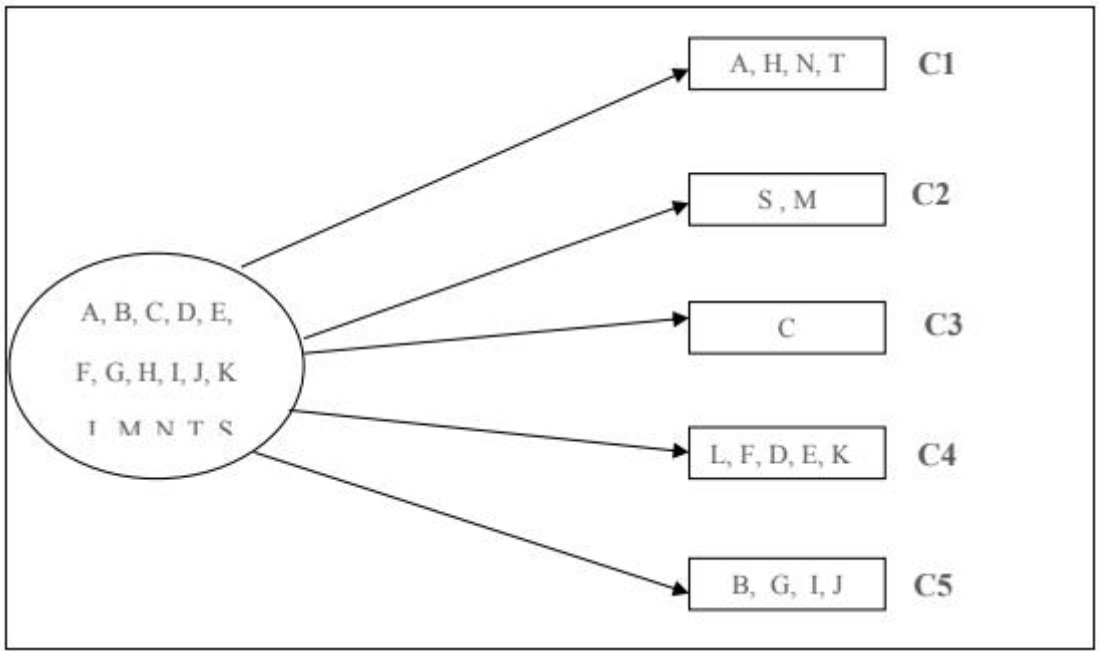


*Figure 7: problématique de tri [4]*

### 4.3 Problématique de rangement (P,$\gamma$)

La décision par un rangement obtenu en regroupement tout ou partie des actions en classe d'équivalence, ces classe étant ordonnées de façons complète ou partielle.

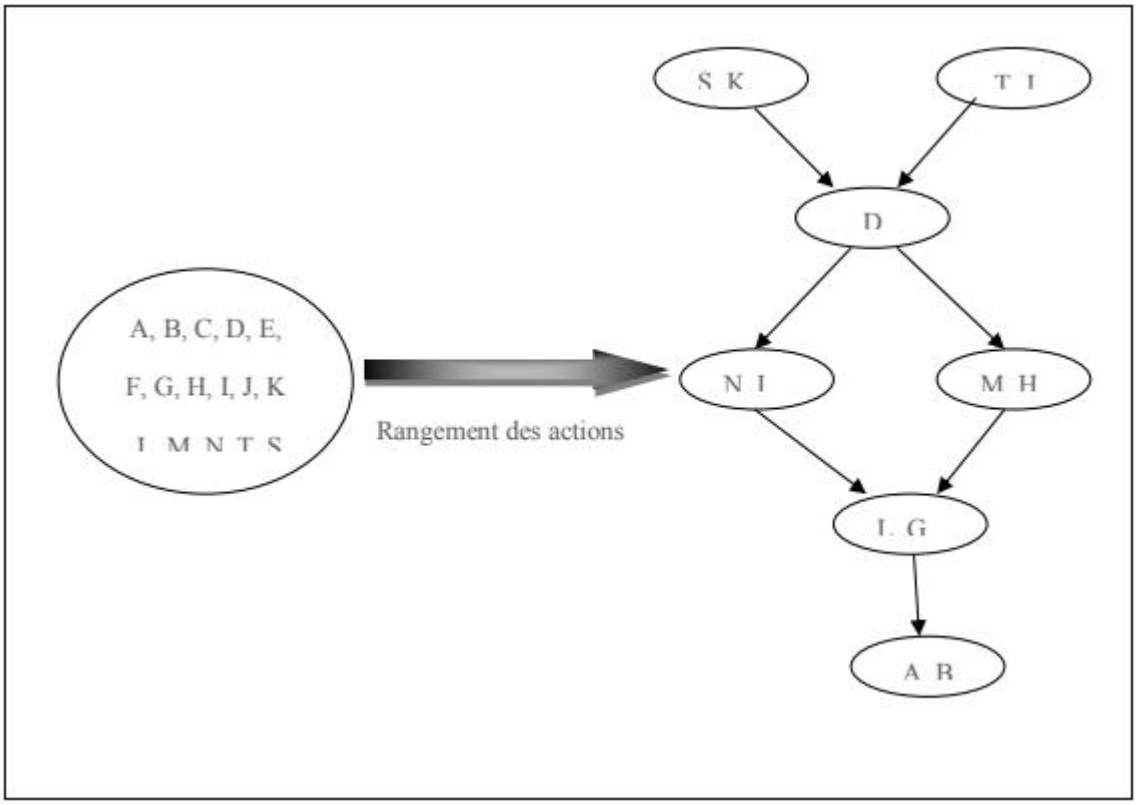


*Figure 8: problématique de rangement [4]*

## 5. Méthodes d'aide multicritère à la décision

Les méthodes d'analyse multicritères sont une étape importante dans le processus de prise de décision, développé depuis 1960, basé sur l'identification et la définition du problème et conduit au choix de la solution par le décideur.

L'adoption d'une stratégie décisionnelle multicritère impose le passage par les étapes suivantes [37] :

- Dresser la liste des actions potentielles
- Dresser la liste des critères à prendre en considération
- Agréger les performances

C'est l'étape de l'agrégation des performances qui fait la différence entre les méthodes d'aide multicritère. On distingue les classes de méthodes suivantes

### 5.1 Les méthodes d'agrégation complète (top-down approach)

Cette méthode consiste à assembler un ensemble de critères et à les réduire à un seul critère.

1. **La méthode Weight SumMethod (WSM):**

C'est la méthode de prise de décision multicritère connue la plus simple connue pour évaluer un certain nombre de solutions de rechange en fonction d'un certain nombre de critères de décision.

$$max_i oumin_i \sum_{j=1}^{m} e_{ij} \times p_j, i = 1,2,\ldots,n$$

2. **La méthode Weight ProductMethod (WPM):**

Elle est similaire au modèle de WSM.

$$max_i oumin_i \prod_{j=1}^{m} \left(\frac{a_{ij}}{a_{Lj}}\right)^{p_{ij}}, pour\ i = 1,2,\ldots,n$$

## 5.2 Les méthodes d'agrégation partielle

Consiste à comparer les actions potentielles et établir des relations de surclassement entre ces éléments.

### 5.2.1 ELECTRE

Il s'agit des méthodes multicritère basées sur la comparaison d'actions. La famille ELECTRE (ELimination Et Choix Traduisant la REalité) contient plusieurs méthodes : Electre I, Electre II, Electre III, Electre IV, Electre Is, Electre Tri.

La méthode Electre I [33] est une méthode proposée par Roy en 1968 permet de résoudre les problématique des choix. Elle construit une relation de surclassement S qui servira à comparer deux actions a et b. On considère un ensemble A évaluées sur un ensemble de critère Gj, l'objectif de la méthode est de sélectionner un sous ensemble d'actions offrant un meilleur compromis parmi l'ensemble de départ. Les étapes de méthode Electre I sont :

**a. Transformation des données :**

Dans cette étape, il faut transformer les performances en notes et dans des valeurs varie sur des échelles dont la longueur évoluera de la même façon que les poids accordés aux critères, en utilisant cette formule de transformation des données :

$$X \in [g\ min, g\ max]\ ,\quad y = \alpha\ x + \beta$$

$$\alpha = \frac{poids-0}{g- \ g\,min} \quad , \quad \beta = 0 - \alpha * g\,min \ = \text{poids} - \alpha * \text{g max}$$

**b. L'évaluation des indices de concordance locale :**

L'indice de concordance locale varie de 0 à 1 .Il est calculé pour chaque critère $g_j$ et il permet de mesure l'hypothèse a surclassement b pour le critère $g_j$. Il calculé par la formule suivantes [34]**:**

$$C(a, b) = \begin{cases} 1 & \text{si } g(b) \geq g(a) \\ 0 & \text{si } g(b) > g(a) \end{cases}$$

**c. L'évaluation des indices de concordance globale :**

L'indice de concordance globale compris entre 0 à 1 , il permet de mesurer les arguments en faveur de « a surclassement b » pour le critère $g_j$. . Calculé par[34]:

$$C(a, b) = \frac{\sum_{gj\,(a) \geq gj\,(b)}^{kj}}{\sum_{j=1}^{n} \quad kj}$$

**d. L'évaluation des indices de discordance :**

L'indice de discordance permet de mesurer la plus grand différance discordant de performance de chaque couple d'actions. Il varie de 0 à 1 et calculé par la formule suivante [31]:

$$D(a, b) = \frac{1}{\delta} \max j\,[gj(b) - gj(a)]$$

δ : la différence maximale entre le même critère pour deux actions donnée.

**e. L'algorithme de classement :**

La relation de surclassement de a S b doit satisfaire un test de concordance et un test de discordance :

- Si (C (a, b)>c)l'indice de concordance C (a, b) supérieure que seuil de concordance c, on passe à la deuxième vérification .Sinon l'hypothèse « a est au moins aussi bonne que b » rejetée.

-si la deuxième condition (Dj(a, b)< d) qui représenté dans l'indice de discordance Dj(a, b) inferieur que seuil de discordance d ,donc l'hypothèse « a est au moins aussi bonne que b » acceptée. Sinon l'hypothèse rejetée.

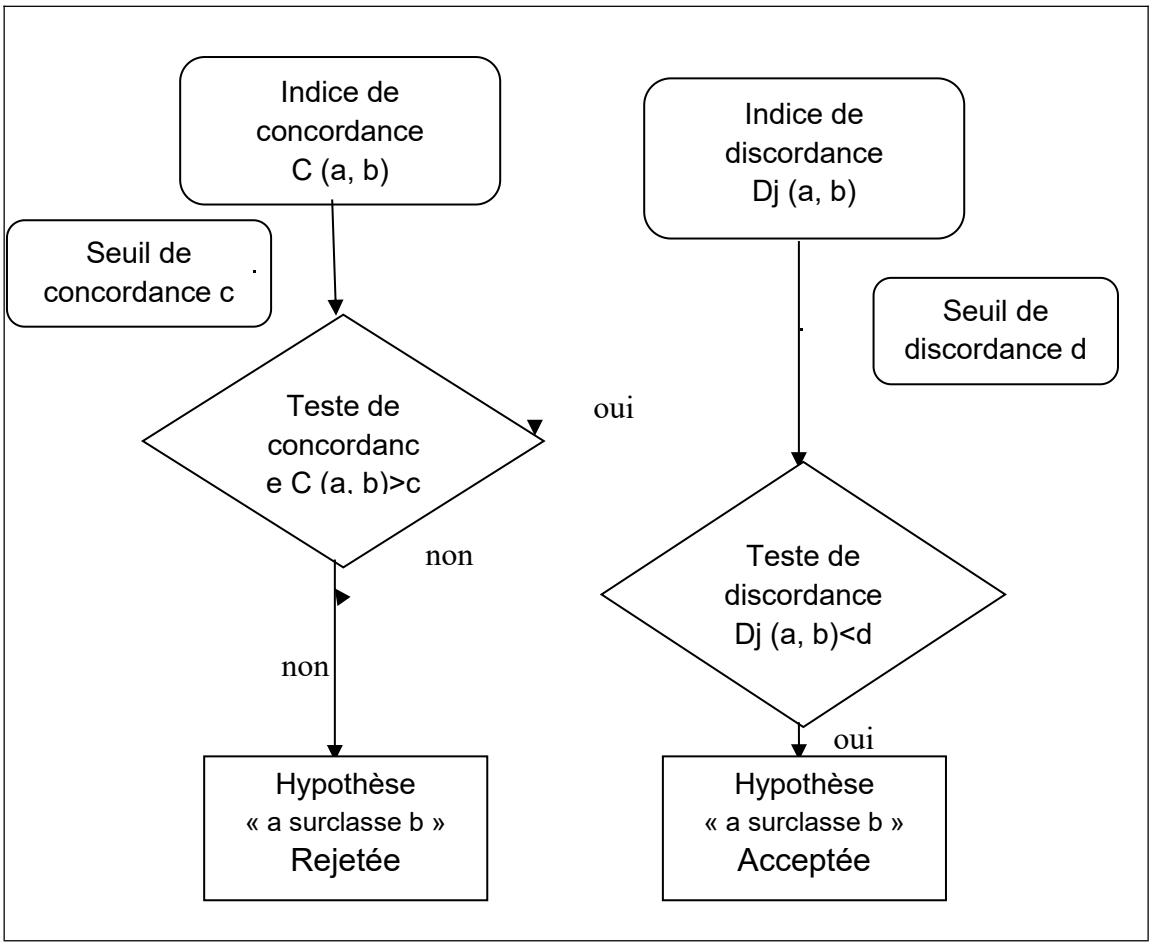


*Figure 9: l'algorithme de surclassement [34]*

### f. Exploiter les relations de surclassement:

Cette étape consiste à déterminer le sous ensemble d'actions N (Noyau).La recherche de N est équivalente à la recherche du noyau de graph G représentant S.

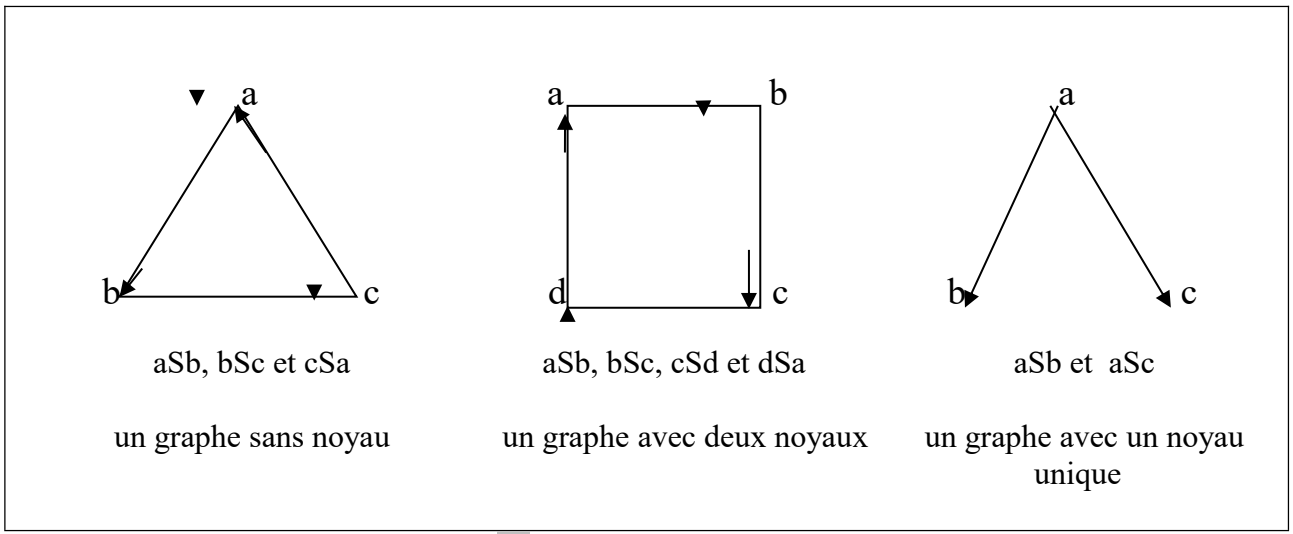


*Figure 10: Exemples de noyaux dans un graphe [31]*

### 5.2.2 PROMETHEE

La méthode PROMETHEE a été proposée pour la première fois en 1982 par Jean Pierre Brans [9], est une partie de la famille de la méthode de surclassement. Le principe général de cette méthode se base sur la comparaison des valeurs de flux (positifs, négatifs ou nets) de profils de préférence avec ceux d'une action à classer. Les étapes de la méthode PROMETHEE sont :

- **Choix des fonctions de préférence :**

Consiste à associer à chaque critère $g_i$ une fonction de préférence $P_j$ : X* X [0,1]. Cette fonction parmi de modéliser les préférences du décideur sur le critère $g_i$ .Lorsque le décideur compare deux alternatives $x_i$ et $x_k$ , $P_j$ ( $x_i$ , $x_k$ )représente le degré de préférence pour $x_i$ , en ne considérant que le critère $x_i$ .Le couple ($g_i$ , $P_j$) est appelé critère généralisé.

- **Degré de préférence multicritère :**

Le degré de préférence multicritère est la moyenne pondérée des fonctions de préférence $P_j$. Il représente le degré de préférence de $x_i$ sur $x_k$ en considèrent simultanément tous les critères.

- **Les flux de surclassement :**

Afin d'apprécier comment chaque action de ***X*** se comporte face aux (n-1) autres action, nous introduisons ici 3 flux de surclassement :

- Le flux de sortant : c'est le flux positif, noté $Ø^+(x)$ ce flux exprime la puissance. Il est donné par : $Ø^+(x) = \frac{1}{n-1}\sum_{x_k \in X} (\delta P(x_i, x_k))$ [32].

- Le flux de sortant : c'est le flux négative, noté $Ø^+(x)$ ce flux exprime la faiblesse. Il est donné par : $Ø^-(x) = \frac{1}{n-1}\sum_{x_k \in X} (\delta P(x_k, x_i))$ [32].

- Le flux net : exprime le bilan des flux entrant et sortant de l'action x. noté Ø (x). . Il est donné par :
  $Ø(x) = Ø^+(x) - Ø^-(x)$[35].

On utilisera PROMETHEE I si on souhaite disposer d'un rangement partiel, certaines actions restent incompatibles, on s'ait qu'une action est

d'autant meilleure que son flux sortant est élevé, et que sont flux entrant est faible.

**PROMETHEE I :**

$$x_i \text{ est préféré à } x_k \text{ si } \begin{cases} \emptyset^+(x) > \emptyset^+(x) \text{ et } \emptyset^-(x) < \emptyset^-(x) \\ \emptyset^+(x) > \emptyset^+(x) \text{ et } \emptyset^-(x) = \emptyset^-(x) \\ \emptyset^+(x) = \emptyset^+(x) \text{ et } \emptyset^-(x) < \emptyset^-(x) \end{cases}$$

$x_i$ est indifférent à $x_k$ si $\emptyset^+(x) = \emptyset^+(x)$ et $\emptyset^-(x) = \emptyset^-(x)$ [22]

PROMETHEE II est un rangement complet de toutes les actions, ce rangement est obtenu en rangeant les actions de l'ordre décroissant des Ø, cette méthode est plus facile à utiliser par rapport à PROMETHEE I .

**PROMETHEE II :**

$x_i$ est préféré à $x_k$ si Ø (x) > Ø (x)

$x_i$est indifférent à $x_k$ si Ø (x) = Ø (x) [32]

### 5.2.3 La méthode d'agrégation locale

Cette méthode est basée sur l'exploitation, interactive et itérative de toutes les actions, ou dans une itération donnée, l'action le plus intéressante est d'abord identifiée puis les actions les plus proches de celle ici sont déterminées. Parmi ce groupe d'actions, nous recherchons s'il existe une meilleure action que l'action initiale qui devient l'action initiale pour l'itération suivante [11].

## 6. Conclusion

Dans ce chapitre nous avons présenté les concepts de base et les problématiques ainsi que les méthodes du domaine d'aide multicritère à la décision. Dans le chapitre suivant, nous présenterons le concept clustering multicritère.

# CHAPITRE 3 : LE CLUSTERING MULTICRITERE ORDONNE

# 1. Introduction

La clustering multicritère est une technique qui permet de classer des objets en fonction de plusieurs critères simultanément. Cette méthode est souvent utilisée dans les domaines tels que la décision, la gestion de projet, la stratégie d'entreprise, entre autres. Elle consiste à utiliser plusieurs critères pour évaluer et classer les objets en question, ce qui permet de prendre en compte différents aspects de chaque objet et d'obtenir une analyse plus complète et plus précise. Dans ce chapitre nous allons présenter les différentes approches de clustering en AMCD.

# 2. Le clustering multicritère

Plusieurs définitions du clustering multicritère ont été proposée, nous en citons, dans ce qui suit, les principales d'entre-elles.

Une première définition a été proposée par Ferligoj et Batagelj [18] : Le clustering multicritère est une méthode d'analyse de données qui vise à regrouper des objets en fonction de plusieurs critères simultanément. Le problème du clustering multicritère consiste à trouver le schéma de clustering optimal c* parmi un ensemble de schémas de clustering possibles C, de manière à minimiser les fonctions d'évaluation $g_j$ (c*) pour chaque critère j de l'ensemble G = $\{g_1 , g_2,...,g_m \}$.

Plus précisément, pour évaluer un schéma de clustering c* par rapport à un critère g j , Ferligoj et Batagelj proposent d'utiliser une fonction $g_j$ (c*) qui associe une valeur numérique à c*. Cette valeur numérique représente l'évaluation de c* par rapport à $g_j$ , et doit être minimale pour que c* soit considéré comme optimal pour le critère $g_j$ . Le schéma de clustering optimal c* est celui qui minimise simultanément toutes les fonctions d'évaluation $g_j$ (c*) pour j = 1, ..., m.

Eppe, Roland et De Smet [15] ont proposé une définition plus générale du clustering multicritère qui intègre les concepts du domaine d'aide multicritère à la décision. Selon cette définition, le clustering multicritère désigne toutes les méthodes de clustering qui prennent en compte les relations de préférence entre les objets, telles que définies par une procédure d'aide multicritère à la décision.

Les relations de préférence incluent les relations d'indifférence (I), de préférence (P) et d'incomparabilité (R).

Contrairement à la définition proposée par Ferligoj et Batagelj, cette définition plus générale permet d'inclure des méthodes de clustering qui ne se limitent pas à la minimisation de plusieurs critères, mais qui prennent en compte les préférences et les priorités de l'utilisateur. Cette approche peut être utilisée dans différents domaines où les décisions doivent être prises en tenant compte de plusieurs critères et de préférences subjectives.

Meyer & Olteanu [28] ont aussi proposé une autre définition du clustering multicritère en se basant sur la définition traditionnelle du clustering comme étant le processus visant à regrouper les objets similaires et à séparer les objets qui ne le sont pas, mais en utilisant la notion des relations de préférence. Selon ces auteurs, le clustering multicritère consiste à regrouper les objets qui sont indifférents les uns aux autres, et à séparer ceux qui ne le sont pas. Cette définition inclut la relation d'indifférence (I) pour regrouper les objets, tandis que les relations de préférence (P) et d'incomparabilité (R) sont utilisées pour séparer les objets.

Contrairement à la définition proposée par Ferligoj et Batagelj, cette définition met davantage l'accent sur la notion de similarité entre les objets et la séparation entre les objets qui ne sont pas similaires, tout en prenant en compte les préférences du décideur. Cela permet d'élargir la portée du clustering multicritère à des problèmes où la notion de similarité est plus pertinente que la simple minimisation de critères.

## 3. clustering multicritère ordonné

Un schéma de clustering ordonné dans le contexte d’aide multicritère à la décision est le processus permettant de regrouper les objets qui sont indifférents tout en séparant les clusters qui sont préférables ou incomparables à d’autres, de façon à ce qu’un ordre soit défini entre ces clusters. Dans ce cas, on ne se contente pas de définir des clusters mais on cherche à définir un ordre entre ces clusters. Cet ordre peut être total s’il est basé sur la relation de préférence seulement. Il est considéré comme partiel s’il existe des clusters incomparables.

Il existe plusieurs approches pour le clustering en Analyse Multicritère des Données (AMCD). Les principales approches sont représentée dans le tableau suivant :

*Tableau 2:Les approches de clustering en AMCD*

| Méthode du clustering en AMCD | Année | points traités | les nouveautés des approches |
|---|---|---|---|
| **L'approche de De Smet & Guzman [13]** | 2004 | contexte d'aide multicritère à la décision. | proposer une approche de clustering basée sur une nouvelle mesure de distance intégrant les préférences du décideur. et sur une extension de l'algorithme k-means au contexte multicritère. |
| **L'approche de De Smet et Eppe [14]** | 2009 | extension de l'approche développée par De Smet et Guzman vers la notion du clustering relationnel. | proposer une définition des relations entre les clusters. |
| **L'approche de Eppe, Roland et De Sme** | 2011 | •extension de celle développée par De Smet et Eppe. •prise en considération du caractère flou de la relation de surclassement. | proposer une nouvelle définition de la notion du profile à travers l'intégration de deux ensembles(des actions surclassant et surclassées) |
| **L'approche de De Smet, Nemery et Selvaraj [30]** | 2012 | déterminer la partition ordonnée caracatérisée par la meilleure matrice d'inconsistance | générer un schéma de clustering dans lequel une relation d'ordre est définie entre les clusters. |
| **Les travaux de Rocha et Dias** | 2013 | produire un ordre partiel entre les clusters. | adapter un algorithme de clustering hiérarchique ascendant au contexte multicritère |
| **L'approche de Meyer et Olteanu [29]** | 2013 | une approche permettant de regrouper les objets en utilisant la relation d'indifférence. | •modifier la structure des clusters pou obtenir un schéma de clustering optimal. •réduire la complexité induite par les déplacements. |

# 4. Conclusion

La classification multicritère est une méthode d'analyse des données qui permet de classer des objets en utilisant plusieurs critères simultanément. Ce chapitre avait pour but de présenter les concepts fondamentaux du clustering en AMCD ainsi que les différentes approches associées à ce domaine. Nous avons abordé également le sujet du clustering multicritère ordonné, qui est une variante de la clustering multicritère dans lequel une relation d'ordre est définie entre les clusters. Nous avons aussi présenté les différentes approches de clustering en AMCD.

# CHAPITRE 4 : LA DETECTION DES ANOMALIES

# 1. Introduction

Dans ce chapitre, nous commençons par introduire la définition de la détection d'anomalies, puis nous présentons les méthodes appliquées pour détecter les anomalies (outlier).

# 2. Détection d'anomalies

La détection d'anomalies consiste à trouver des objets qui sont différents ou incohérents avec la plupart des objets dans les ensembles de données et sont appelés anomalies ou outliers (Aggrawal, 2017) [1]. Il existe plusieurs définitions de terme anomalie mais la définition qui apparait fréquemment dans la littérature est la définition de Hawkins en 1980 [21] : « Une anomalie est une observation qui diffère tellement d'autres observations au point d'éveiller des soupçons qu'elle soit générée par un mécanisme différent ». Nous prenons l'exemple de la figure 11 pour plus de clarté, où nous trouvons deux groupes de points normaux $N_1$ et $N_2$, et nous trouvons des observations $O_1$, $O_2$ et $O_3$ qui sont éloignées de la plupart des points de $N_1$ et $N_2$, elles sont donc considérées comme outliers.

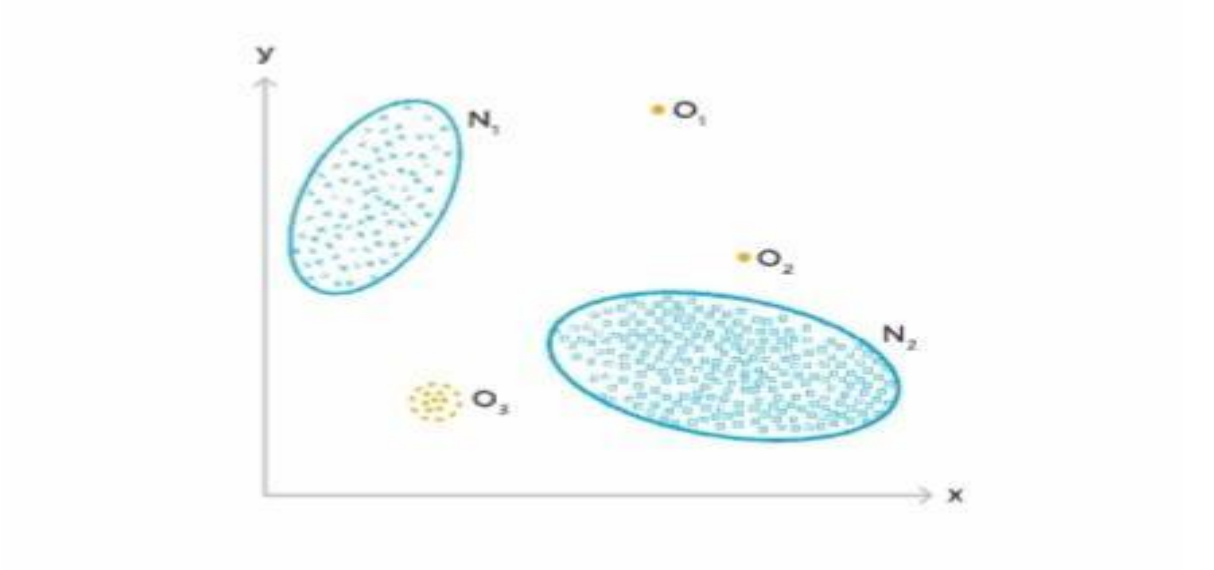


**Figure 11 :** Exemple de détection d'anomalies

Le résultat des méthodes de détection d'anomalies pour chaque observation évaluée peut être de deux types : l'obtention d'une décision (normale ou anormale), tandis que le second type est l'évaluation du degré d'anomalie de l'observation à partir du résultat.

## 3. Méthodes de la détection des anomalies

La détection des anomalies repose sur plusieurs méthodes qui ont été proposées par des chercheurs**.** On distingue les méthodes suivantes :

### 3.1. Méthodes statistiques

Le but de cette méthode est d'estimer empiriquement la distribution statistique en question. Les points normaux apparaissent dans les régions de l'espace où la densité de probabilité est élevée, tandis que des anomalies apparaissent dans les régions de faible densité de probabilité. Les méthodes statistiques sont divisées en deux parties : méthodes paramétriques et méthodes non paramétriques.

#### 3.1.1. Méthodes paramétriques

Cette méthodes consiste à supposer que les données suivent une distribution prédéterminée puis à utiliser les données disponibles pour définir empiriquement les paramètre de ce modèle en minimisant ou en maximisant la mesure choisie. Parmi les modèles de méthodes paramétriques le modèle Gaussiens (Yamanishi et al, 2004) [36] qui vise à déterminer la moyenne et l'écart-type. Ce modèle est populaire dans la détection d'anomalies, car il a été utilisé dans des séries temporelles associées au modèle de régression linéaire. Il existe plusieurs méthodes de paramétriques comme Z-Score, Grubb's, ARIMA.

#### 3.1.2. Méthodes non paramétrique

Ce sont des techniques statistiques qui dépendent de l'estimation de paramètres. Le nombre de paramètre estimés augmente avec la quantité de données disponibles et ne dépend d'aucune hypothèse a priori sur la distribution contrairement aux méthodes paramétriques. Nous trouvons plusieurs méthodes statistiques non paramétriques comme histogramme (Fawcett et Provost, 1999)[17] et les fonctions à noyaux (Bishop, 1994)[7]. Les méthodes à base d'histogramme repose sur l'utilisation de graphes en général afin d'obtenir une représentation visuelle de la distribution unidimensionnelle expérimentale. L'espace est divisé en cellules pour lesquelles les colonnes sont construites, où la hauteur des colonnes correspond au nombre d'échantillons dans lesquels la valeur se trouve dans la cellule, de sorte que la cellule associée à une probabilité élevée aura une colonne plus longue que la cellule associée à une faible probabilité. La forme du graphique lorsque le nombre d'échantillons augmente tend vers la fonction de densité de la distribution, et nous pouvons donc utiliser

la hauteur de la cellule pour déterminer l'anomalie. Nous prenons l'exemple de la figure 12, le point le plus à droite est une anomalie.

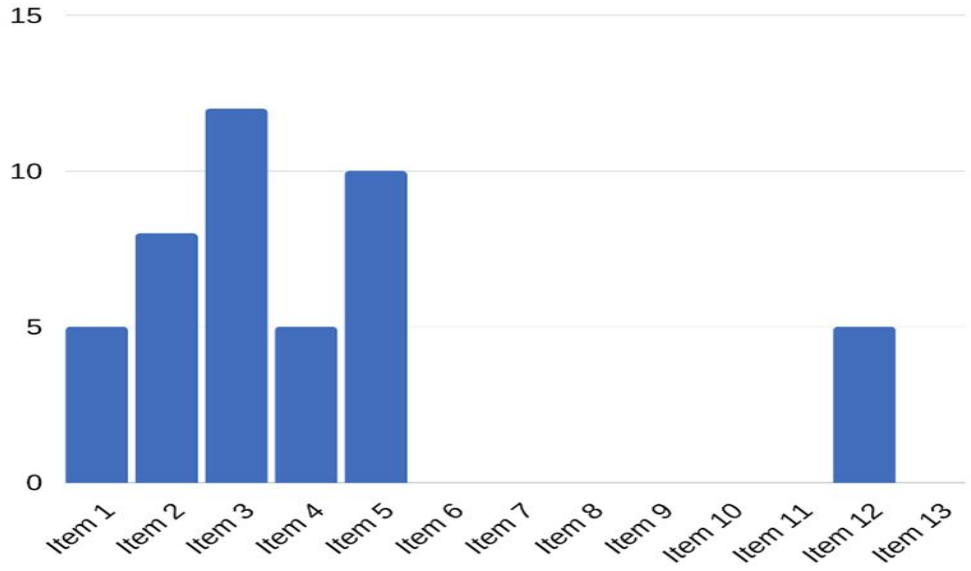


**Figure 12 :** Exemple d'histogramme

## 3.2. Méthodes de Clustering

Les techniques de clustering peuvent aussi être utilisées pour la détection des anomalies. En fait, si l'algorithme de clustering génère un cluster composé d'un seul objet alors ce dernier est considéré comme une anomalie. Les approches clustering par densité basée sur la densité telles DBSCAN, BRICH, CluStream sont considérées comme des techniques de détection d'anomalies.

## 3.3. Méthodes basées sur le principe du plus proches voisins

Ces méthodes dépendent du calcul la distance entre toutes les observations de l'ensemble de données afin de déterminer les voisins les plus proches de l'ensemble. Il existe de classes de méthodes basées sur le concept du plus proche voisins : l'approche basée sur la distance ( Angiulli et Pizzuti (2002)[2] ; Yamanishi et al (2004) [36]) comme KNN, STROM , Abstract-C. L'approche basée sur la densité (Breunig et al (2000)[5] ) comme LOF, iLOF, GLOF.

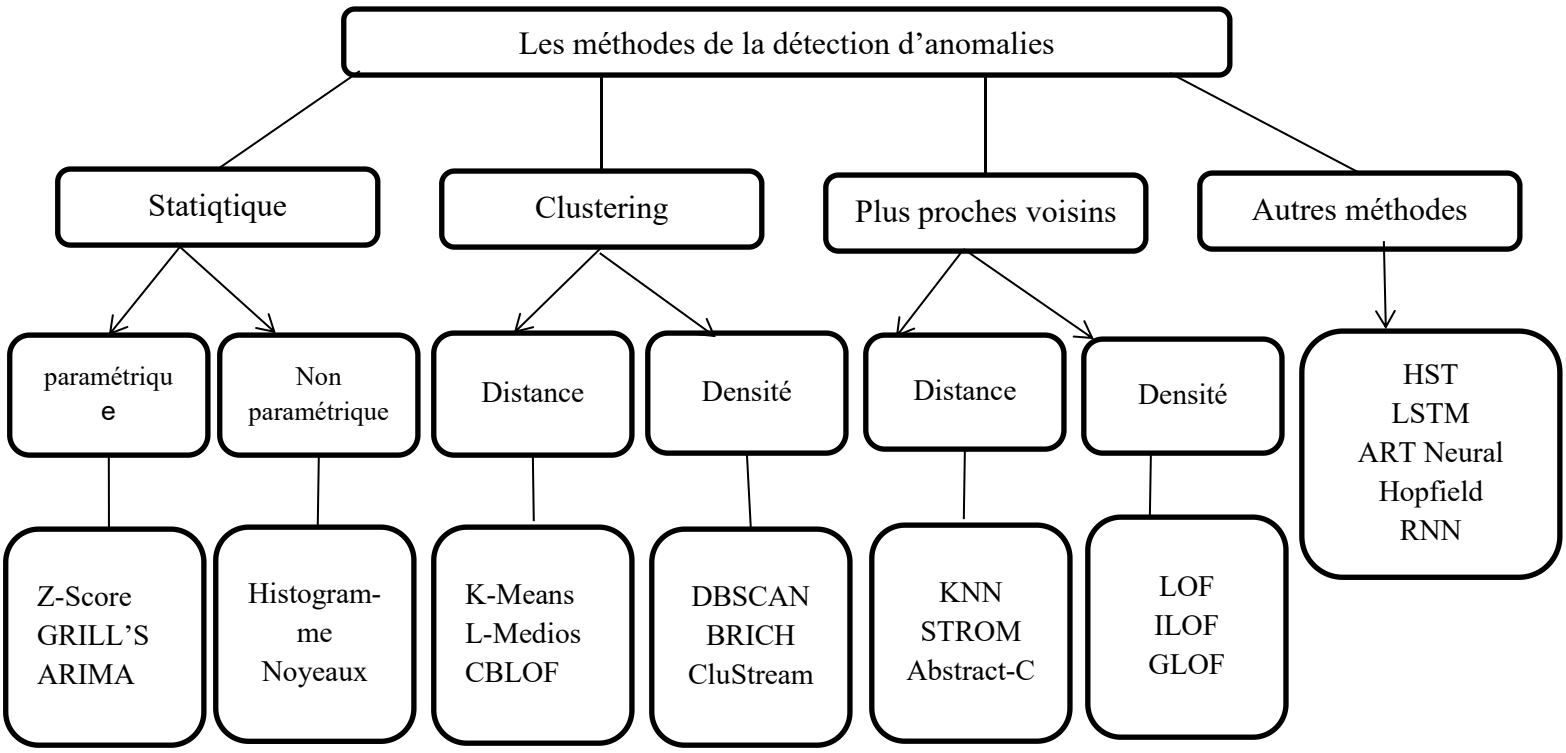


**Figure 13 :** Les méthodes de la détection d'anomalies

## 4. Conclusion

Dans ce chapitre nous avons donné une définition de la détection des anomalies et évoqué les méthodes existantes dans ce domaine. Dans ce qui suit, nous verrons comment nous utilisons la détection d'anomalies dans le clustering multicritère ordonné.

# CHAPITRE 5 : UNE APPROCHE DE CLUSTERING MULTICRITERE ORDONNEES BASEE SUR LA DETECTION DES ANOMALIES

# 1. Introduction

Ce chapitre présente une approche du clustering multicritère ordonné basée sur la détection d'anomalies. Nous présentons également des résultats expérimentaux de cette approche sur de jeux de données pour montrer l'efficacité et la pertinence de cette approche.

# 2. Clustering multicritère ordonné basée sur la détection des anomalies

Nous nous sommes appuyés sur le travail de ROUBA [10] qui a expérimenté une approche de détection des anomalies basée sur les méthodes statistiques et le flux net de la méthode PROMETHEE. En fait, nous nous sommes basés sur le fait que les objets appartenant à un même cluster suivent une loi normale. Nous avons, ainsi, construit des clusters par un processus itératif, permettant d'ajouter, à chaque itération, un objet à un cluster. Le processus s'arrête dès qu'un objet outlier est détecté. Ce dernier permet représente la frontière entre les différents clusters. L'ordre entre les clusters est assuré par les valeurs du flux net. Le diagramme suivant illustre le processus de notre approche.

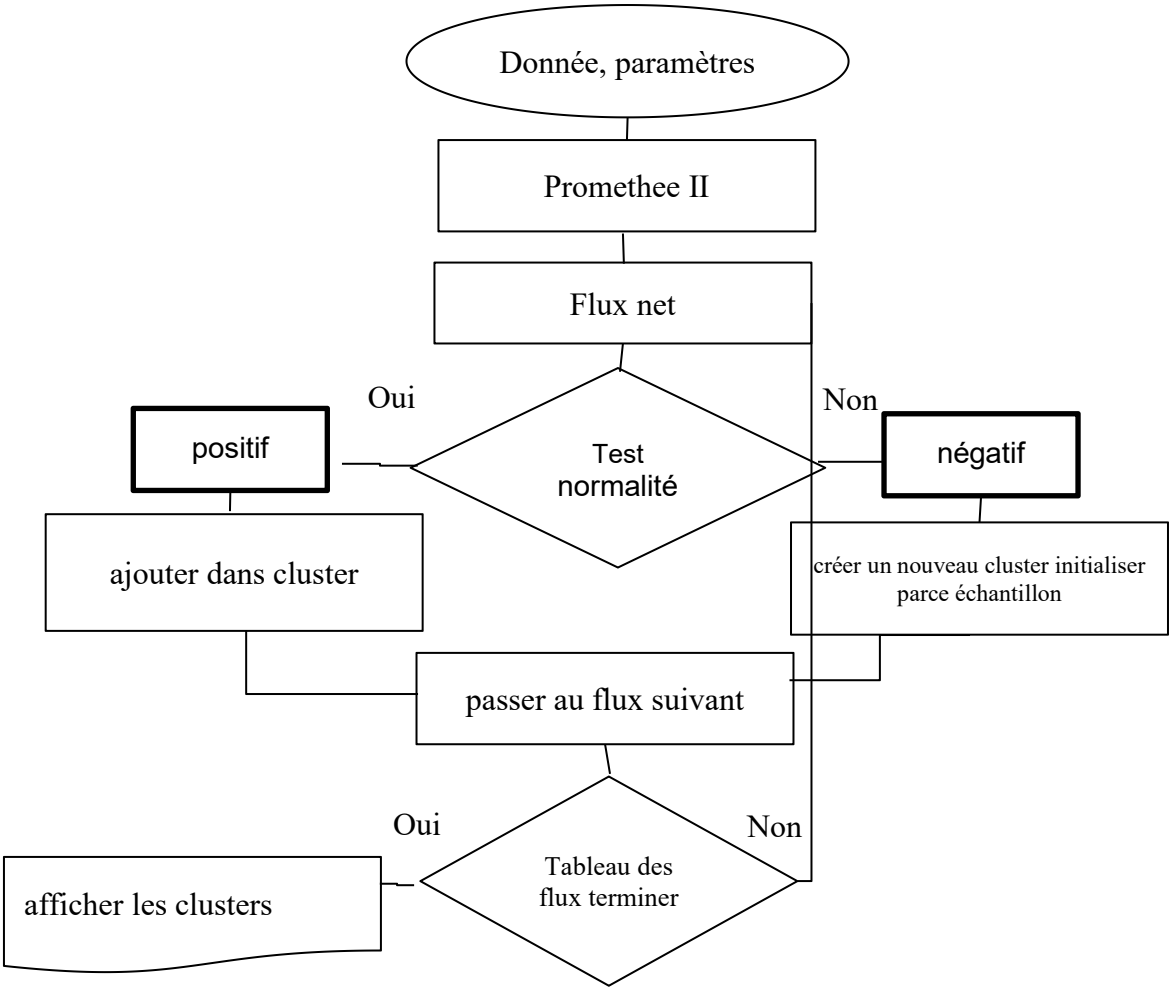


**Figure 14 :** Processus de notre approche

## 3. Environnement de travail

L'environnement de travail que nous avons utilisé pour développer notre projet est basé sur Python et la bibliothèque Hugging Face, et utilise également l'interface Streamlit pour faciliter l'utilisation. Python est un langage de programmation populaire et polyvalent largement utilisé dans les domaines de l'apprentissage automatique et de l'analyse de données. La bibliothèque Hugging Face est connue pour ses outils et modèles avancés d'apprentissage automatique, en particulier dans le traitement du langage naturel. Streamlit est un framework qui facilite la création d'applications Web interactives à partir de code Python.

Grâce à ces outils, nous avons pu développer une application pour implémenter notre méthode de clustering ordinal basée sur la détection d'anomalies.

## 4. L’interface graphique de l'application

Pour mettre en pratique notre approche, nous proposons une méthodologie en plusieurs étapes. L'utilisateur a le choix entre deux options :

Option 1: Importer un fichier CSV contenant les données à analyser. L’utilisateur est invité à importer le fichier. Une fois importé, en cliquant sur le bouton "Calculer", les algorithmes sont exécutés et les résultats sont affichés à l'utilisateur.


**Figure 15 :** L’interface dans le cas de l’importation un fichier CSV

Option 2: Entrer les informations manuellement. L'utilisateur peut choisir de saisir les données directement dans l'application.


**Figure 16 :** L’interface dans le cas de saisie manuelle

Pour cela, il doit fournir le nombre d'alternatives (n) et le nombre de critères (m). Ensuite, il lui sera demandé de sélectionner les options pertinentes pour chaque critère, telles que la fonction (Usuel, Forme_en_U, Forme_en_V, Level, linéaire ou Gaussien) et les critères de minimisation et de maximisation Min/ Max.

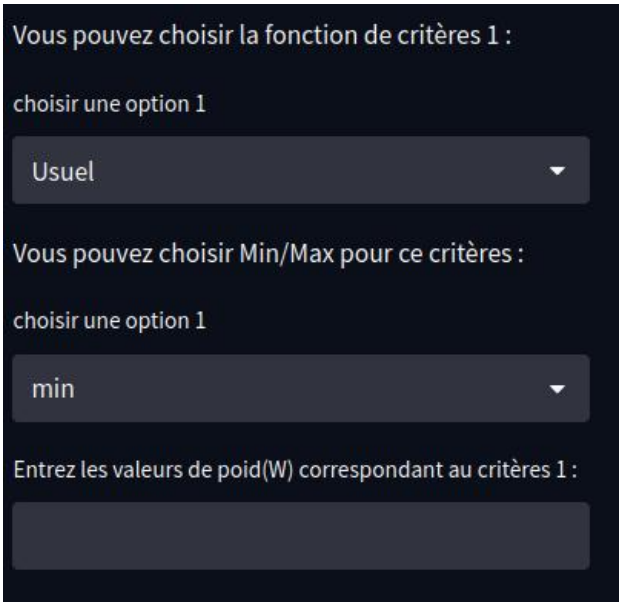


**Figure 17 :** Le choix de fonction et le paramètres correspondants aux critères

L'utilisateur sera ensuite invité à saisir les valeurs des paramètres correspondant à chaque alternative en fonction de la fonction choisie.

Pour chaque alternative, il devra entrer le nom et les valeurs correspondantes de la ligne.


**Figure 18 :** Remplissage de la base de données

Une fois les données stockées, en cliquant sur le bouton "Calculer", l'application effectue les calculs pour obtenir les résultats, détecte les anomalies, regroupe les points de données en clusters pertinents et affiche les résultats correspondants à l'utilisateur.

## 5. Etude de cas

Pour mettre en pratique notre approche, nous proposons dans cette section de l'expérimenter sur un problème réel qui est celui de l'index du développement humanitaire HDI (Human Development Index). Le HDI est un indicateur largement utilisé pour mesurer le niveau de développement humain des pays. L'objectif principal du problème HDI est de classer les pays selon leurs niveaux de vie, en identifiant ceux qui affichent un développement humain plus élevé par rapport aux autres. Cela permet de mettre en évidence les disparités entre les pays et d'orienter les politiques publiques pour améliorer les conditions de vie des populations. Il s'agit d'un problème multicritère qui prend en compte plusieurs critères tels que l'espérance de vie, l'alphabétisation, et le PIB. L'ensemble des pays à classer est composé de 179 pays. Notre objectif est de regrouper ces pays en clusters ordonnés.

Nous avons sélectionné la matrice de performance du problème à partir d'un fichier CSV.

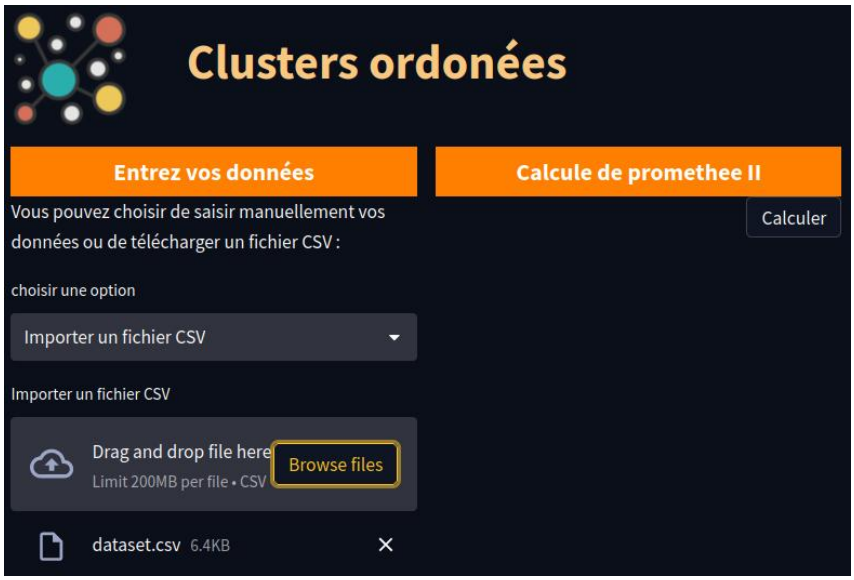


**Figure 19 :** Importation des données du problème

L'application affiche d'abord les données du problème ainsi que les paramètres correspondant.

| | 0 | 1 | 2 | 3 |
|---|---|---|---|---|
| 0 | dimensions | 179 | 3 | None |
| 1 | None | life | litracy | GDP |
| 2 | unit | year | year | millions |
| 3 | Min/Max | max | max | max |
| 4 | Poids | 0,333 | 0,333 | 0,333 |
| 5 | Fonction de pr | Level | Level | Level |
| 6 | Seuils | abs | abs | abs |
| 7 | q | 0 | 0 | 0 |
| 8 | p | 0,704 | 0,719 | 0,828 |
| 9 | Iceland | 0,944 | 0,98 | 0,982 |

**Figure 20 :** Affichage de contenu de fichier importé

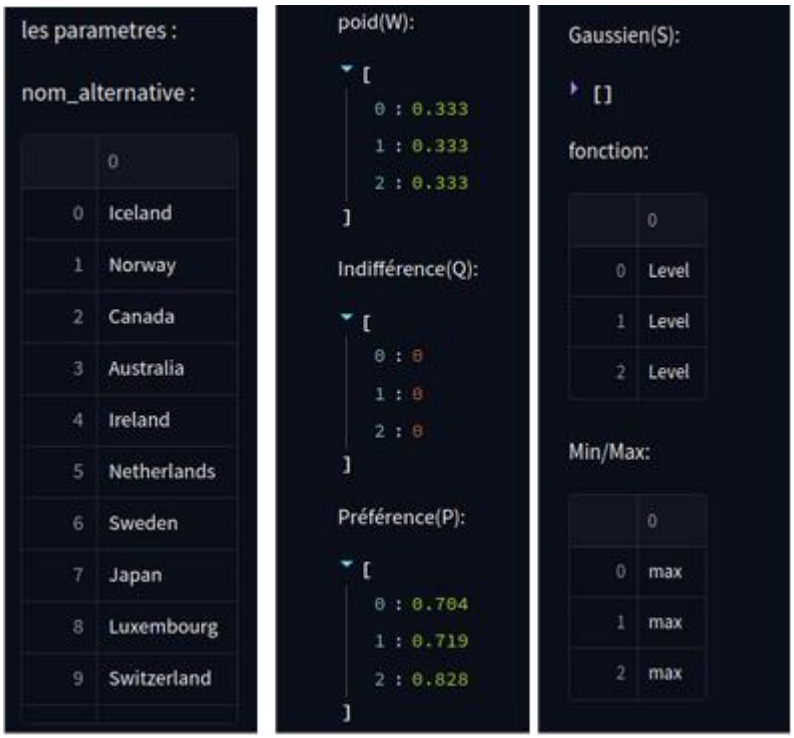


**Figure 21 :** Affichage des paramètres

Ces données ont été utilisées comme base pour les calculs du flux net en se basant sur la méthode PROMETHEE 2. L'algorithme de construction des clusters est ensuite lancé.

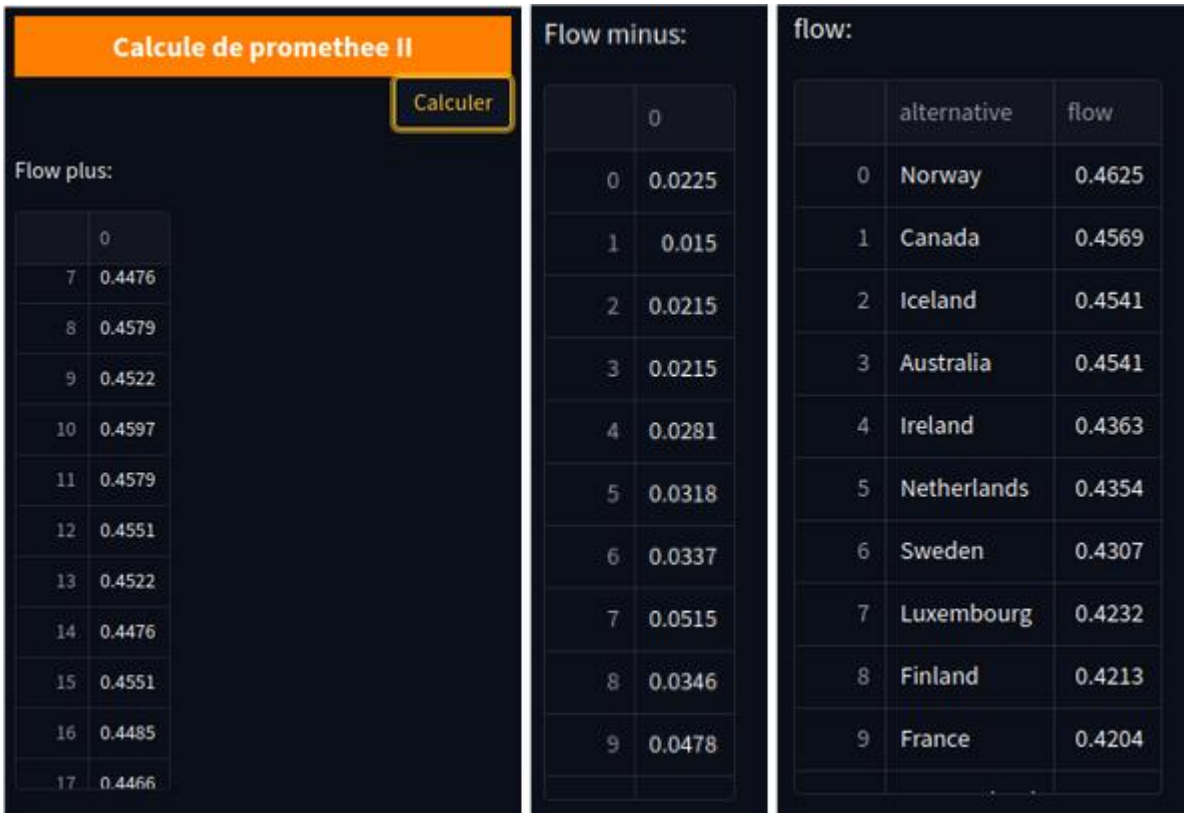


**Figure 22 :** Résultat de l'exécution de la méthode PROMETHEE


**Figure 23 :** Affichage des indices correspondants aux anomalies

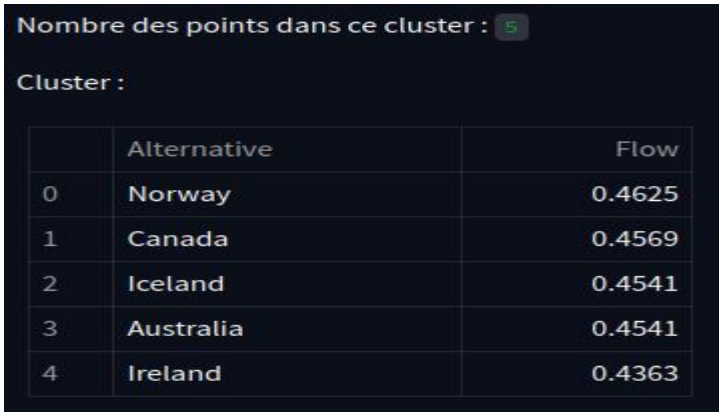


**Figure 24 :** Affichage de premier cluster

Nombre des points dans ce cluster : 1

Cluster :

| | Alternative | Flow |
|---|---|---|
| 0 | Netherlands | 0.4354 |

**Figure 25 :** Affichage de deuxième cluster

Nombre des points dans ce cluster : 101

Cluster :

| | Alternative | Flow |
|---|---|---|
| 0 | Sweden | 0.4307 |
| 1 | Luxembourg | 0.4232 |
| 2 | Finland | 0.4213 |
| 3 | France | 0.4204 |
| 4 | New Zealand | 0.4195 |
| 5 | Denmark | 0.4139 |
| 6 | Spain | 0.4110 |
| 7 | Switzerland | 0.4045 |
| 8 | Austria | 0.4045 |
| 9 | United States | 0.4036 |
| 10 | Belgium | 0.4007 |
| 11 | Greece | 0.3970 |
| 12 | Japan | 0.3961 |
| 13 | Italy | 0.3923 |
| 14 | United Kingdom | 0.3764 |
| 15 | Korea(Republic of) | 0.3745 |
| 16 | Germany | 0.3689 |
| 17 | Israel | 0.3652 |
| 18 | Slovenia | 0.3483 |
| 19 | Hong Kong, China(SAR) | 0.3296 |
| 20 | Brunei Darussalam | 0.3165 |
| 21 | Cyprus | 0.3146 |
| 22 | Czech Republic | 0.2940 |
| 23 | Portugal | 0.2893 |
| 24 | Singapore | 0.2875 |
| 25 | Barbados | 0.2837 |
| 26 | Bahrain | 0.2725 |
| 27 | Poland | 0.2678 |
| 28 | Kuwait | 0.2640 |
| 29 | Chile | 0.2631 |
| 30 | Hungary | 0.2556 |
| 31 | Cuba | 0.2556 |
| 32 | Malta | 0.2537 |
| 33 | Uruguay | 0.2500 |
| 34 | Lithuania | 0.2500 |
| 35 | Slovakia | 0.2472 |
| 36 | United Arab Emirates | 0.2434 |
| 37 | Croatia | 0.2416 |
| 38 | Qatar | 0.2369 |
| 39 | Argentina | 0.2360 |
| 40 | Latvia | 0.2247 |
| 41 | Estonia | 0.2107 |
| 42 | Costa Rica | 0.1845 |
| 43 | Bulgaria | 0.1816 |
| 44 | Libyan_Arab_Jamahiriya | 0.1798 |
| 45 | Panama | 0.1676 |
| 46 | Mexico | 0.1629 |
| 47 | Montenegro | 0.1507 |
| 48 | Romania | 0.1489 |
| 49 | Saint_Lucia | 0.1442 |
| 50 | Bahamas | 0.1442 |
| 51 | Venezuela_Bolivarian_Republic_of | 0.1433 |
| 52 | Serbia | 0.1414 |
| 53 | Seychelles | 0.1320 |
| 54 | Albania | 0.1245 |
| 55 | Saint_Kitts_and_Nevis | 0.1236 |
| 56 | Oman | 0.1227 |
| 57 | Belarus | 0.1180 |
| 58 | Kazakhstan | 0.1161 |
| 59 | Malaysia | 0.1124 |
| 60 | Macedonia_TFYR | 0.1021 |
| 61 | Russian_Federation | 0.0993 |
| 62 | Ecuador | 0.0955 |
| 63 | Saudi Arabia | 0.0955 |
| 64 | Antigua_and_Barbuda | 0.0936 |
| 65 | Brazil | 0.0880 |
| 66 | Bosnia_and_Herzegovina | 0.0805 |
| 67 | Tonga | 0.0805 |
| 68 | Dominica | 0.0787 |
| 69 | Ukraine | 0.0674 |
| 70 | Trinidad_and_Tobago | 0.0646 |
| 71 | Mauritius | 0.0618 |
| 72 | Armenia | 0.0590 |
| 73 | Colombia | 0.0365 |
| 74 | Thailand | 0.0337 |
| 75 | Lebanon | 0.0300 |
| 76 | Peru | 0.0300 |
| 77 | Turkey | 0.0272 |
| 78 | Belize | 0.0253 |
| 79 | Georgia | 0.0187 |
| 80 | Jordan | 0.0131 |
| 81 | Samoa | 0.0112 |
| 82 | China | 0.0056 |
| 83 | Grenada | 0.0019 |
| 84 | Tunisia | -0.0037 |
| 85 | Occupied_Palestinian_Territories | -0.0047 |
| 86 | Jamaica | -0.0056 |
| 87 | Philippines | -0.0169 |
| 88 | Dominican_Republic | -0.0197 |
| 89 | Suriname | -0.0206 |
| 90 | Guyana | -0.0206 |
| 91 | Iran | -0.0225 |
| 92 | Saint_Vincent_and_the_Grenadines | -0.0290 |
| 93 | Azerbaijan | -0.0300 |
| 94 | Maldives | -0.0300 |
| 95 | Turkmenistan | -0.0309 |
| 96 | Paraguay | -0.0346 |
| 97 | Algeria | -0.0375 |
| 98 | Mongolia | -0.0375 |
| 99 | Syrian_Arab_Republic | -0.0384 |
| 100 | Moldova | -0.0412 |

**Figure 26 :** Affichage de troisième cluster

Nombre des points dans ce cluster : 57

Cluster :

| | Alternative | Flow |
|---|---|---|
| 0 | Viet_Nam | -0.0487 |
| 1 | Gabon | -0.0543 |
| 2 | El Salvador | -0.0571 |
| 3 | Sri_Lanka | -0.0590 |
| 4 | Fiji | -0.0637 |
| 5 | Kyrgyzstan | -0.0674 |
| 6 | Equatorial_Guinea | -0.0684 |
| 7 | Bolivia | -0.0740 |
| 8 | Uzbekistan | -0.0758 |
| 9 | Tajikistan | -0.0880 |
| 10 | Nicaragua | -0.1021 |
| 11 | Egypt | -0.1030 |
| 12 | Indonesia | -0.1039 |
| 13 | Cape_Verde | -0.1217 |
| 14 | South_Africa | -0.1255 |
| 15 | Honduras | -0.1283 |
| 16 | Botswana | -0.1358 |
| 17 | Guatemala | -0.1358 |
| 18 | Vanuatu | -0.1629 |
| 19 | Morocco | -0.1751 |
| 20 | Namibia | -0.1863 |
| 21 | Sao_Tome_and_Principe | -0.2097 |
| 22 | Bhutan | -0.2219 |
| 23 | Congo | -0.2303 |
| 24 | India | -0.2378 |
| 25 | Lao_People?_Democratic_Republic | -0.2416 |
| 26 | Swaziland | -0.2650 |
| 27 | Solomon_Islands | -0.2715 |
| 28 | Yemen | -0.2743 |
| 29 | Cambodia | -0.2772 |
| 30 | Myanmar | -0.2790 |
| 31 | Pakistan | -0.2818 |
| 32 | Comoros | -0.2865 |
| 33 | Mauritania | -0.2959 |
| 34 | Kenya | -0.3155 |
| 35 | Ghana | -0.3165 |
| 36 | Cameroon | -0.3221 |
| 37 | Djibouti | -0.3221 |
| 38 | Nepal | -0.3240 |
| 39 | Sudan | -0.3240 |
| 40 | Madagascar | -0.3268 |
| 41 | Haiti | -0.3277 |
| 42 | Angola | -0.3287 |
| 43 | Papua_New_Guinea | -0.3315 |
| 44 | Bangladesh | -0.3333 |
| 45 | Nigeria | -0.3390 |
| 46 | Senegal | -0.3390 |
| 47 | Lesotho | -0.3399 |
| 48 | Tanzania | -0.3483 |
| 49 | Uganda | -0.3511 |
| 50 | Gambia | -0.3652 |
| 51 | Timor-Leste | -0.3670 |
| 52 | Togo | -0.3708 |
| 53 | Zambia | -0.3727 |
| 54 | Benin | -0.3745 |
| 55 | Malawi | -0.3811 |
| 56 | Cote_d?voire | -0.3858 |

**Figure 27 :** Affichage de quatrième cluster

Nombre des points dans ce cluster : 13

Cluster :

| | Alternative | Flow |
|---|---|---|
| 0 | Rwanda | -0.3989 |
| 1 | Guinea | -0.3989 |
| 2 | Eritrea | -0.4026 |
| 3 | Chad | -0.4036 |
| 4 | Mali | -0.4120 |
| 5 | Burkina Faso | -0.4232 |
| 6 | Burundi | -0.4279 |
| 7 | Ethiopia | -0.4279 |
| 8 | Niger | -0.4307 |
| 9 | Liberia | -0.4345 |
| 10 | Guinea-Bissau | -0.4391 |
| 11 | Congo_Democratic | -0.4401 |
| 12 | Mozambique | -0.4476 |

**Figure 28 :** Affichage de cinquième cluster

Nombre des points dans ce cluster : 2

Cluster :

| | Alternative | Flow |
|---|---|---|
| 0 | Central_African_Republic | -0.4588 |
| 1 | Sierra_Leone | -0.4728 |

Nombre des clusters : 6

**Figure 29:** Affichage de sixième cluster

## 6. Conclusion

En conclusion, ce chapitre a présenté une approche de clustering multicritère ordonné basée sur la détection d'anomalies. Nous avons commencé par illustrer notre idée et la représenter à l'aide d'un algorithme. Ensuite, nous avons développé un exemple afin montrer l'applicabilité de cette approche.

# Conclusion générale

Ce travail a exploré plusieurs aspects liés à la classification des données. L'approche traditionnelle de la classification basée sur des étiquettes prédéfinies atteint rapidement ses limites, surtout lorsque l'on travaille avec de grands volumes de données provenant de sources diverses. Cela a conduit à l'émergence de nouvelles techniques de classification, telles que le clustering, qui permettent une meilleure compréhension des relations entre les données.

Dans ce travail, nous nous sommes intéressés plus spécifiquement à la construction de clusters ordonnés dans un contexte décisionnel multicritère. Notre objectif n'était pas seulement de regrouper les données, mais aussi de définir une relation d'ordre entre les clusters. Pour ce faire, nous avons abordé trois problèmes clés : comment prendre en compte le caractère multicritère lors de la construction des clusters ? Comment construire les clusters ? et comment définir une relation d'ordre entre eux.

Pour tenir compte du caractère multicritère, nous avons utilisé la méthode PROMETHEE, reconnue pour sa puissance et sa simplicité. En ce qui concerne la construction des clusters, nous avons employé une technique statistique de détection des anomalies, en exploitant l'hypothèse que les objets appartenant à un même cluster suivent une loi normale. Nous avons ainsi développé un processus itératif permettant d'ajouter progressivement des objets à un cluster, jusqu'à ce qu'un objet "anomalie" soit détecté.

Quant à la définition de la relation d'ordre, nous avons réutilisé le concept du flux-net de la méthode PROMETHEE, qui peut être utilisé pour classer les objets du meilleur au moins bon.

Cependant, nous ne pouvons pas encore nous prononcer sur la qualité des résultats obtenus. En effet, par manque de temps, nous n’avons pas pu comparer cette approche par rapport à d’autres approches du même domaine. Ceci constitue une première perspective de notre travail. Il serait aussi nécessaire d’expérimenter l’approche sur d’autres exemples afin d’évaluer son efficacité.

# Références bibliographiques


- [1] Aggarwal, C. C. (2017). Outlier Analysis (Second Edition ed.). Springer International Publishing AG 2017.
- [2] Angiulli, F. et C. Pizzuti (2002). Fast outlier detection in high dimensional spaces. In European Conference on Principles of Data Mining and Knowledge Discovery, pp. 15–27. Springer.
- [3] Anil K. Jain, M. Narasimha Murty et Patrick J. Flynn : Data clustering: A review, un article de synthèse,1999, 31(3), p 264-323.
- [4] Belacel Nabil :PROCFTN : Une nouvelle procédure du choix flou pour les problèmes d'affectation multicritère ,un article, 2000.
- [5] Breunig, M. M., H.-P. Kriegel, R. T. Ng, et J. Sander (2000). Lof : identifying densitybased local outliers. In ACM sigmod record, Volume 29, pp. 93–104. ACM..
- [6] Bernard Roy :Aide multicritère à la décision : méthodes et cas, un livre par Éditions Economica., 1985.
- [7] Bishop, C. M. (1994). Novelty Detection and Neural Network Validation. IEE Proceedings-Vision, Image and Signal processing, 141(4) , 217- 222.
- [8] Blondel, V. D., Guillaume, J. L., Lambiotte, R., & Lefebvre, E. : Fast unfolding of communities in large networks. un article publié dans la revue "Journal of Statistical Mechanics: Theory and Experiment", 2008, P 10008.
- [9] Brans, J. P. : A preference ranking organization method: The PROMETHEE method for multiple criteria decision-making, un article scientifique publié dans la revue "Recherche Opérationnelle/Operations Research", 1982.
- [10] B ROUBA « A Net-flow based approach to detect outliers in multicriteria decision problems ». Intelligent Decision Technologies (IDT). Vol 15 N°2, 2021, pp.239–250.
- [11] Cengiz Kahraman, İhsan Kaya, et Tugçe Kılınç : Decision Making in Manufacturing Environment Using Graph Theory and Fuzzy Multiple Attribute Decision Making Methods: Volume 1, un livre, 2013.
- [12] Christopher D. Manning, Prabhakar Raghavan et Hinrich Schütze : Introduction to Information Retrieval, un livre publié par Cambridge University Press, 2008, p 135-137.

- [13] De Smet, Y., & Montaño Guzmán, L. : Towards multicriteria clustering: An extension of the k-means algorithm. Revue européenne de recherche opérationnelle, article ,2004 , 317-326.
- [14] De Smet & Eppe : Multicriteria Relational Clustering: The Case of Binary Outranking Matrices, un article scientifique, 2009, 194(2), 406-422.
- [15] Eppe, S., De Smet, Y., & Teghem, J. : Multicriteria clustering and classification: A review. Statistical Science, un article, 2011, 26(4), 531-551.
- [16] Ethem Alpaydin: Introduction to Machine Learning, un livre publié par le MIT Press ,2010, p 196-198.
- [17] Fawcett, T. et Provost, F. (1999). Activity Monitoring : Noticing Interesting Changes in Behavior. Dans Froc. 5th International Conference on K nowledge Discovery and Data Mining, 53- 62 . ACM.
- [18] Ferligoj, A., & Batagelj, V. : Clustering approach to multicriterion problem solving: A case study. Revue européenne de recherche opérationnelle “un articl”, 1992, 56(1), 28-40.
- [19] Fraley, C., & Raftery, A. E. : Model-based clustering, discriminant analysis, and density estimation. Journal de l'Association statistique américaine,2002 ,p 611-631,
- [20] G. J. McLachlan et D. Peel. : Finite Mixture Models, une livre publié par John Wiley & Sons, 2000, le chapitre 1.
- [21] Hawkins, D. M. (1980) . Identifi cation of Outliers., volume 11. Springer.
- [22] Jaccard, P. : The distribution of the flora in the alpine zone. 1. New Phytologist, un article , 1912, 11(2), p 37-50.
- [23] Jain, A. K., Murty, M. N., & Flynn, P. J. : Data clustering: a review. In ACM Computing Surveys (CSUR) ,un livre ,1999, p 264-323..
- [24] Kaushik, S., & Sural, S. : A comparative study of k-means and k-medoids algorithm for clustering. un article publié dans la revue International Journal of Computer Science and Information Security (IJCSIS), 2018, p 72-77.
- [25] Laurent HENRIET : Systèmes d’évaluation et de classification multicritère pour l’aide à la décision, Thèse de Doctorat, université paris dauphine, 2000.
- [26] Liuhao Chen, Zeshui Xu, Hai Wang, Shousheng Liu : An ordered clustering algorithm based on K-means and the PROMETHEE method, un article scientifique, 2018, p 139-150.
- [27] Martine Cadot, Alain Lelu, Michel Zit : BENCHMARKING SEVENTEEN CLUSTERING METHODS, 2018, p 226-231.
- [28] Meyer, P., & Olteanu, A. L. : Multicriteria clustering: A review. Journal international des techniques et stratégies d'analyse de données, 2013, 5(2), 115-142.

- [29] Meyer & Olteanu : Formalizing and solving the problem of clustering in MCDA,un article, 2013, la 3ème section.
- [30] Nemery and Y. de Smet : Multicriteria ordered clustering. Technical Report. TR/SMG/2005-003, Université Libre de Bruxelles/SMG, 2005, p 36, 38, 51 et 79.
- [31] Philippe Lenca : Aide multicritère à la décision Méthodes de surclassement. GET / ENST Bretagne Département lussi, l'université de rennes, 2004.
- [32] ROUBA Baroudi: Minimisation des désagréments dans les clusters agrégés, Thèse de Doctorat, Université d'Oran 1 Ahmed Benbella, Faculté des Sciences Exactes et Appliquées, 2015, p 9-59.
- [33] Roy, B., & Vincke, P. : Multicriteria analysis: survey and new directions. Revue européenne de recherche opérationnelle, 1985, 377-389.
- [34] S. Ben Mena : Méthodes multicritères d'aide à la décision : méthodes de surclassement. une note , 2008.
- [35] Taibi Boumedyen : La méthode PROMETHEE comme outil d'aide à la décision multicritère, un article, 2017.
- [36]Yamanishi, K., J.-I. Takeuchi, G. Williams, et P. Milne (2004). On-line unsupervised outlier detection using finite mixtures with discounting learning algorithms. Data Mining and Knowledge Discovery 8(3), 275–300.
- [37] Zeleny, M. : Multiple criteria decision making. McGraw-Hill. un livre ,1982.