Chapitre 1

Etat de l'art

Le clustering est le regroupement de plusieurs données similaires en un seul groupe appelé cluster. L'analyse de cluster permet ainsi d'identifier des groupes de données relativement homogènes sur la base de leur similarité pour des caractéristiques données ce qui dans notre cas peut par exemple être le type d'emploi occuper en fonction des filières suivies par d'anciens étudiants. Cependant analyser différents profils d'individus peut représenter des difficultés techniques importantes, c'est pourquoi cette première section du document présentera les différentes solutions possibles pouvant apporter une réponse à cette difficulté de catégorisation des parcours.

1.1 La préparation des données

Lorsqu'une segmentation basée sur des clusters est utilisée, il existe plusieurs formes de préparations de données pouvant aider à la formation des différents segments.

1.1.1 La Transformation des variables

Il existe deux types de transformation :

— La modification de la portée des variables connue en tant que la standardisation des variables.

— La modification de la forme de distribution

L'utilisation de la standardisation est motivée par le fait que l'analyse de cluster implique une étude implicite du poids des objets afin de pouvoir se concentrer sur ceux possédant une variance plus élevée. Les méthodes de standardisation les plus communes sont les suivantes [?] :

— Multiplication de chaque variable par une différente constante.

— Utilisation des techniques de réduction de dimensions, qui un processus visant à réduire le nombre de variables aléatoires afin d'obtenir un jeux de variables principal.

— Multiplier chaque variables par une différente constante afin que chacune d'entre elles aient une portée commune.

La modification de la distribution quant à elle est motivée par les mêmes problématiques que dans d'autres secteurs ayant recours aux statistiques qui sont d'extrêmes variations par rapport à ce qui est considéré "normal" dans le cas étudié. Celles-ci entrainent des analyses pouvant induire en erreur. Par conséquent lorsque le poids des variables est modifié le but est d'identifier et supprimer leur longue traine qui correspond à un nombres d'entre elles possédant des valeurs très supérieures ou inférieures à la moyenne.

1.1.2 La mesure de la distance

Le choix de la méthode de mesure de distance est une étape critique pour les méthodes de clustering, son choix ayant une très forte influence sur le résultat final. En effet, la méthodologie choisie définira comment les similarités de deux éléments sont calculés et influera par conséquent sur la forme des clusters également. Les deux méthodes les plus communes de mesure sont la distance Euclidienne illustrée par la formule suivante (figure1.1) :

$$d_{euc}(x,y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$$

FIGURE 1.1 – Exemple de dendrogramme

Dans cette méthode, la distance est calculée en effectuant le carré de la somme des carrés des distances entre les variables répondant à un critère donné.
La seconde méthode communément utilisée est la distance de Manhattan, appelée également "taxi-distance". Celle ci, pour un point A et B de coordonnées respectives (X_a,Y_a) et (X_b,Y_b) est définie de la façon suivante :

$$d(A,B) = [X_b - X_a] + [Y_b - Y_a]$$

A l'inverse de la méthode euclidienne qui pourrait être influencée par des valeurs inhabituelles, le calcul de la distance de Manhattan va s'effectuer selon la différence moyennes entre les dimension. La présence de valeurs aberrantes impactera le résultat de façon réduite étant donné qu'elle ne sera pas élevée au carré contrairement à la méthode euclidienne, ce qui fait que cette méthode aura tendance à donner le même type de résultat.

1.2 Les méthodes de clustering

1.2.1 Le clustering hiérarchique

Parmis les différents types de clustering existant [?], le premier étudié sera le clustering hiérarchique. Très utilisé comme outil d'analyse de données, l'idée principale de cette méthode est de construire un arbre binaire fusionnant de façon successive les groupes de points similaires. L'un des avantages de cette méthode est tout d'abord l'apport de l'arbre qui permet d'avoir une vision globale des données traitées. De plus, cette méthodologie possède ses propres outils de visualisation qui sont le dendrogramme et la classification double. Le dendrogramme permet d'illustrer l'arrangement des clusters (figure 1.2)[?] :

— la racine de l'arbre est formée par un cluster contenant l'ensemble des objets.
— chaque nœud de l'arbre consitue un cluster.
— l'union des objets des noeuds fils correspond aux objets présent dans ce noeud.
— les paliers sont indexés en fonction de l'ordre de construction.

Tandis que la classification double est une technique d'exploration de données non-supervisée permettant de segmenter simultanément les lignes et les colonnes d'une matrice. L'autre avantage du clustering hiérarchique est sa facilité d'implémentation dans des algorithmes tel que K-Means en plus de fournir une représentation comme dit précédemment. Afin d'établir un arbre hiérarchique, le clustering hiérarchique à recours à deux méthodes qui sont la méthode agglomérative et la méthode divisive. Un regroupement agglomératif traite chaque objet comme un seul élément qui à chaque étape de l'algorithme est fusionné avec un second objet présentant le plus de similarités en un nouveau cluster de plus grande taille. Ce processus est répété jusqu'à que ce que tous les points soient membre d'un seul et même cluster. A l'inverse d'un regroupement agglomératif qui utilise une approche "bottom-up", les algorithmes divisifs utilisent une approche "top-down". Ces algorithmes débutent ainsi leur traitement à partir de la racine de l'arbre ou tous les objets sont regroupés en un seul cluster. A chaque ittération les cluster les plus hétérogènes sont divisés en deux jusqu'à ce que l'ensemble des objets fassent parti de leur propre cluster.Cependant sa complexité le rend inefficace sur de larges jeux de données [?]. De plus, la première injection de données ainsi que l'ordre de celles-ci à un fort impact sur le résultat final. En outre, il n'est pas possible de défaire ou modifier les étapes précédentes du traitement, c'est à dire qu'une fois une instance assignées à un cluster, il n'est plus possible de la déplacée pour effectuer d'éventuelles modifications ou corrections [?]. Dans notre cas la base de CV utilisée n'étant pas de taille importante le clustering hiérarchique reste une méthode applicable. Cependant la problématique à résoudre est la gestion des filières intégrant plusieurs domaines tel que la filière MIASHS de Nanterre qui possède une dimension mathématiques et une informatique. Les données étant représentées sous forme d'arbre cela entrainerait une répétition au niveau des résultats.

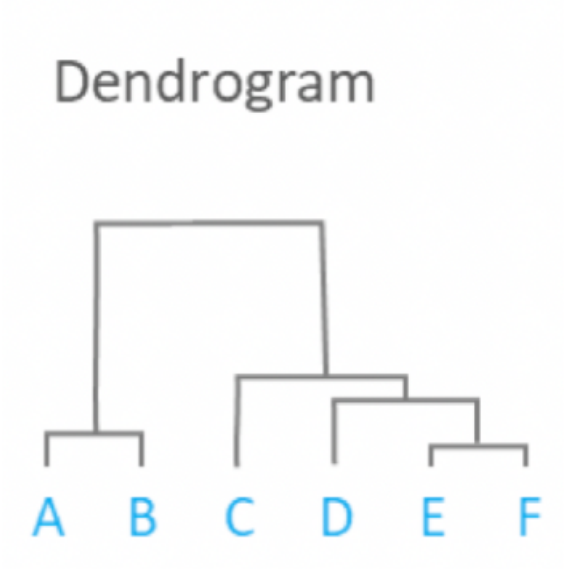

FIGURE 1.2 – Exemple de dendrogramme

1.2.2 Le clustering par partionnement

Le clustering par partionnement contrairement au clustering hiérarchique qui utilise un arbre afin de représenter les différents groupe de données va classifier les différents objets en groupe en fonction de leur similarités. Cependant ce mode de fonctionnement pause un problème concernant le choix de la "bonne représentation" en fonction d'un critère choisi, le but devient alors de recherche une représentation optimale de son critère à travers plusieurs itérations.[?] L'algorithme le plus utilisé pour ce type de méthode est K-means qui sera présenté dans la suite de ce document.

1.2.3 Autre méthodes de clustering

1.3 Les types de clusters

Le but du clustering étant de trouver des groupes d'objets présentant des similarités définies en fonction de l'objectif recherché. Il existe toutefois une multitude de types de cluster qui seront étudiés au sein de cette section chacun avec ses avantages et inconvénients en fonction de notre cas avant de statuer sur le type qui sera utilisé pour le reste de ce document.

1.3.1 Well-Separated

Un cluster "well-separated" est un regroupement de points de telle façon à ce que tous les points faisant parti d'un même cluster présentent de fortes similarités entre eux comparés aux points d'un cluster extérieur.

1.3.2 Prototype-Based

Un prototype-based cluster est un cluster dont les points qui le constitue sont plus proches ou similaire du prototype définissant le cluster traité que de tout autre prototype définissant d'autres clusters.

1.3.3 Graph-Based

Le graph-based cluster est utilisé dans les cas ou les données peuvent être représenté sous forme de graphe dont les nœuds sont des objets et les liens représentent les connexions entre ceux-ci. Dans cette situation un cluster peut être défini comme un composant connecté, c'est-à-dire un groupe d'objets liés les uns aux autres au sein du même groupe.

1.3.4 Density-Based

A travers l'utilisation d'un density-based cluster, le but est de détecter les zones ou les points formant des clusters sont concentrés et celles ou les points sont séparés par des zones vides ou par des zones contenant très peu de points. Les points ne faisant par partie d'un agrégat sont ici classés comme du bruit.

1.4 Les algorithmes

Dans cette section seront décrit les principaux algorithmes utilisés lorsque des techniques de clustering sont employées. Les avantages et inconvénients de chacun seront présentés en fonction du cas présenté dans l'introduction.

1.4.1 K-means

L'algorithme K-means est l'algorithme le plus populaire, celui-ci recherche la meilleure division possible au sein d'un jeux de données [?] en plus de sa facilité d'implémentation. Cependant celui-ci impose de savoir le nombre de clusters souhaités et par conséquent une bonne connaissance des données utilisées.

1.4.2 Agglomerative Hierarchical Clustering

Les techniques de clustering agglomératives partent d'un ensemble de points formant un cluster, par la suite, les deux clusters les plus proches sont fusionnés successivement jusqu'à ce qu'il n'y est plus qu'un seul cluster restant. [?]

1.4.3 DBSCAN

DBSCAN est un algorithme basé sur le partitionnement de données, celui-ci utilise deux principaux paramètres qui sont la distance et le nombre de points minimum devant se trouver dans un rayon donné afin qu'ils soient considérés comme un cluster.