Introduction

Afin d'orienter les étudiants arrivant en première année de licence vers un parcours qui leur semble le plus adapté, il serait intéressant de pouvoir leur fournir un aperçu des chemins suivis par leur prédécesseur et éventuellement le poste auquel ceux-ci travaillent désormais. C'est pourquoi ce document s'oriente vers la modélisation de carrière des étudiants et leur catégorisation à l'aide de méthodes de clustering. Le premier obstacle étant la catégorisation des métiers qui sont le plus souvent vagues et de certains titre tel que consultant pouvant couvrir une grande variété de professions. Par conséquent, la première partie de mon mémoire se concentrerait sur les différentes méthodes de clustering, ce qui permettrait de pouvoir affecter un tag aux différents métiers occupés par les anciens étudiants. Dans un second temps, les algorithmes utilisables seront étudiés suivis par les multiples types de cluster existant dans le soucis de choisir une structure adaptée à notre cas et ainsi obtenir une représentation adaptée des parcours d'étudiants.

Table des matières

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 Les types de clustering

1.1.1 Clustering hiérarchique

Parmis les différents types de clustering existant [3], le premier étudié sera le clustering hiérarchique. Très utilisé comme outil d'analyse de données, l'idée principale du clustering hiérarchique 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 méthode de clustering possède ses propres outils de visualisation qui sont le dendrogramme et la classification double. Le dendrogramme permet d'illustrer l'arrangement des clusters (figure 1.1) tandis que la classification double est une technique d'explorations de données non-supervisée permettant de segmenter simultanément les lignes et les colonnes d'une matrice. Les avantages du clustering hiérarchique sont sa facilité d'implémentation dans des algorithmes tel que K-Means en plus de fournir une représentation comme dit précédemment. Cependant sa complexité le rend inefficace sur de larges jeux de données [1]. De plus, la première injection de données ainsi que leur ordre 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 [2]. 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

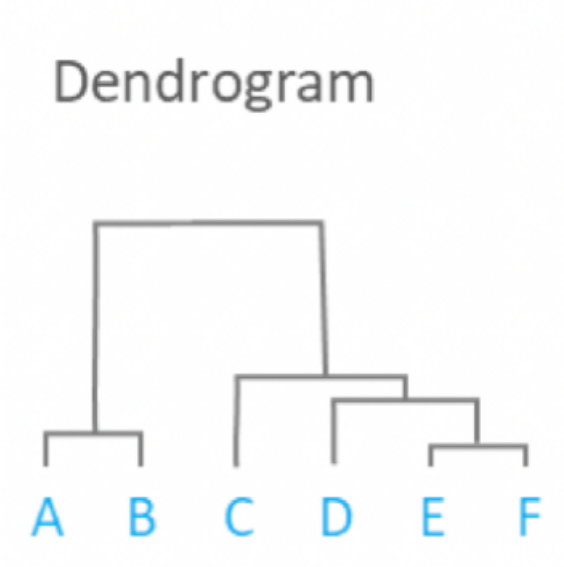

FIGURE 1.1 – Exemple de dendrogramme

est comment gérer efficacement les 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.

1.1.2 Clustering partionné

LOREM IPSUM

1.2 Les algorithmes

LOREM LE IPSUM

1.2.1 K-means

1.2.2 Agglomerative Hierarchical

1.2.3 DBSCAN

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

1.3.2 Prototype-Based

1.3.3 Graph-Based

1.3.4 Well-Separated

Chapitre 2

Implémentation

2.1 Problématique

Bibliographie

[1] M. Santini. Advantages and disadvantages of k-means and hierarchical clustering (unsupervised learning). *Machine Learning for Language Technology*, 2016.

[2] D. Sonagara1 and S. Badheka2. Comparison of basic clustering algorithms. *International Journal of Computer Science and Mobile Computing*, 2014.

[3] P.-N. Tan, M. Steinbach, A. Karpatne, and V. Kumar. *Cluster Analysis : Basic Concepts and Algorithms.* Pearson ; 2 edition (January 4, 2018), 2018.