Unsupervised Learning

Johnny Nguyen[1]

***Abstract*— This electronic document permits me to synthesis the fifth course of data mining.**

I. INTRODUCTION

Lorsque nous parlons d'apprentissage non supervisé, c'est équivalent à traiter du clustering. C'est une collection d'objet. Nous allons définir ce qu'est un clustering puis avoir un aperçu des méthodes de clustering populaires.

II. CLUSTERING

C'est une collection de données. De manière générale, nous cherchons à calculer les similarités et les distances entre chaque groupe. Nous ne connaissons pas les classes à l'avance. La plupart du temps, c'est utilisé pour obtenir une distribution ou l'étape précédent l'implémentation d'un modèle. Les domaines d'applications sont la reconnaissance de pattern, d'analyse de données spatiales, l'analyse d'image, la science de l'économie et le web. Par exemple, profiler un client, établir les coûts, les catastrophes ou l'observation de la terre. Nous pouvons avoir des représentations différentes en fonction des contraintes que nous définissons. Qu'est-ce qu'un bon clustering ? Des groupes de tailles équivalentes, la mesure choisie et son abilité à trouver des patterns. La similarité est définie par la distance entre deux clusters. L'utilisation de la déviation moyenne absolue est plus robuste que la déviation standard. Le z-score est la mesure standard. Il y a une multitude de type de variables. (OBNR)

III. APERÇU DES CÉLÈBRES MÉTHODES DE CLUSTERING

Dans cette partie, nous prendrons connaissance des principaux algorithmes. L'idée principale est de partitionner le dataset en plusieurs clusters. En théorie, pour trouver les meilleurs paramètres, nous allons devoir tester tous les cas en spécifiant un nombre de cluster différent à chaque itération. A chaque itération, nous minimisons la distance entre les points au carré.

A. K-means

L'idée principale est de récupérer le centre de chaque cluster, représenté par la moyenne des coordonnées de ce cluster. Relativement efficace, de l'ordre de **O(tkn)** avec t correspondant au nombre d'itérations, k le nombre de cluster et n le nombre de points. Il peut-être couplé à un recuit simulé ou un algorithme évolutionnaire. Par contre, il n'est pas applicable sur le type de données catégorie. Il faut spécifier le nombre de cluster. Le bruit et les valeurs absolues ne sont pas gérées.

*This work was not supported by any organization

B. K-medoids

La différence avec le K-means, c'est qu'il est moins sensible aux données abérantes et se focalise sur les vrais points d'un cluster. Il est efficace sur une **petite quantité de données**.

C. PAM (Partionning Around Medoids

Plus robuste que K-means sur de **petites quantités de données**.

D. C-means

Cet algorithme nous donne les mêmes clusters et centroîds que celui de K-means. La **notion de dégré** est ajoutés pour définir l'appartenance d'un élément à un cluster.

E. Hiérarchical

Cette méthode permet de connaître pour chaque étape, les clusters disponibles. L'avantage est que nou savons toutes les combinaisons de clusters possibles. Cette fois, nous devons définir la **terminaison qui s'effectue sur le nombre d'étape**.

F. Density-Based Clustering

Cette méthode connecte les points les plus proches entre eux pour former des clusters. C'est très utile pour trouver un cluster possédant une forme complexe. Density-reachable correspond à une chaine. Density-connect correspond à des points.

G. Expectation maximization

Cet algorithme contrairement aux K-means cherche à maximiser son cluster le plus grand. En effet, les groupes ne sont pas équitables et nous obtenons un groupe dominant sur les autres.

IV. CONCLUSION

Bien que l'algorithme des K-means semblent le plus générique, le C-means et le hiérarchicals semblent intéressant par la précision de leurs données. Le Density-Based clustering et l'expectation maximization sont à utiliser si les clusters ont des exceptions. Le K-medoids est le meilleure algorithme pour de petites quantité de données.

ACKNOWLEDGMENT

Thanks to Andrea Tettamanzi for his work.

REFERENCES

[1] http ://www.i3s.unice.fr/ tettaman/