TP Clustering

Intervenants : MJ. Huguet – M. Siala

Le but de ces TP est de développer une méthode de clustering et de la comparer à une méthode proposée dans scikit-learn.

Le travail est à réaliser en binôme.

Première partie

- Récupérez les jeux de données pour tester les performances d'algorithmes de clustering basés sur des méthodes par voisinage. Ces jeux de données ont été proposés par George Karypis, Eui-Hong (Sam) Han et Vipin Kumar (voir référence sur la méthode Chameleon).
- Appliquez la méthode DBSCAN de scikit-learn sur ces données.
- Analysez les résultats obtenus (paramétrage de la méthode, qualité du clustering, forme des clusters, détection d'anomalies, ….)
- Comparez les résultats obtenus aux résultats connus de la littérature (à mettre en relation avec les caractéristiques de la méthode).

Deuxième partie

- Implémentez l'algorithme SNN basé sur un graphe des k plus proches voisins proposé dans l'article : https://www-users.cs.umn.edu/~kumar/papers/siam_hd_snn_cluster.pdf
- Analyser les résultats obtenus (paramétrage de la méthode, qualité du clustering, forme des clusters, sensibilité aux anomalies, ….)
- Comparez les résultats obtenus aux résultats connus de la littérature : comparaison avec la méthode DBSCAN testée en première partie ou avec d'autres méthodes de clustering.

- Pour aller plus loin (facultatif) : testez votre approche sur d'autres jeux de données en plus grande dimension.

Rendu des TP

- Rapport (par exemple sous forme de notebook) avec explication de la démarche, évaluation et analyse des résultats
- Date de rendu : ** **
- Modalité d'envoi : ** **

Pour aller plus loin

1. Article sur la méthode Chameleon : http://glaros.dtc.umn.edu/gkhome/node/152
2. Variante parallèle de la méthode SNN : http://cds.iisc.ac.in/faculty/vss/courses/PPP2015/projects/Nikhilesh_Suguna_SNNparallel.pdf