Département IA & SD le 07/04/2024

M1 BioInfo. MODULE : FD1

TP N° 3

La fouille de données ou "data mining" est un processus qui permet de découvrir des connaissances inconnues au préalable, enfouies dans une masse volumineuse de données.

Dans le cadre du tp N°3 il est question de :

- lire un benchmark dans le but d'effectuer le clustering de ses données : il faut evidemment proceder à la phase de preprocessing des données d'abord

- tracer la courbe d'elbow dans le but de trouver un nombre de clusters enfouis dans la population

- appliquer l'algorithme AGNES puis DIANA

Pour determiner autant de clusters que trouvés par la courbe d'Elbow

-appliquer l'algorithme DBSCAN , faire varier Minpts et le rayon EPS jusqu'à trouver autant de clusters que ceux determinés par la courbe d'Elbow .

Puis calculer les mesures de performances des clusters obtenues par chacun des 3 algorithmes .

-etablir un histogramme des inerties des 5 methodes (K_Means, K_Medoids, Agnes, Diana et DBScan) et comparer ainsi les performances de chaque methode.

# Rapport FINAL à remettre le 29/04/2024

Ce rapport final DOIT contenir TOUS LES RESULTATS du TP depuis le début du semestre S2 M1BIOINFO , ie : le preprocessing, le clustering par les 5 algorithmes de clustering vus en cours

Ainsi que l'interface graphique IHM qui integre toutes ces méthodes : (la classification supervisée sera un bouton « vide » présent dans l'interface en prévision du tp de FD2).

-le bouton « benchmark » doit contenir une liste de datasets que l'utilisateur peut selectionner pour une tache de data mining

- Le bouton « preprocessing doit contenir toutes les méthodes de preprocessing des benchmarks.

-le bouton 'clustering' doit contenir tous les algorithmes vus en cours