Université de Lorraine – Master 2 informatique

TP Clustering (Classification non supervisée)

Exercice 1

Nous allons utiliser les données **Iris** pour tester les algorithmes de clustering K-means (*simple K-means*) et HAC (*Hierarchical Clusterer*) implémentés dans Weka.

Explorer pour comprendre à quoi correspondent les différents paramètres de ces deux programmes et la façon dont les résultats sont restitués. Résultats textuels mais aussi la visualisation des affectations des instances du jeu de données aux clusters (*visualize cluster assignments*) et la visualisation du dendrogramme dans le cas de HAC.

1- Réaliser le clustering du jeu de données **Iris** avec chacun des programmes et les valeurs par défaut des paramètres. Il est inutile de normaliser les différents attributs avant le calcul de distance car c'est déjà dans le programme *simple K-means*

2- Pour aller plus loin, réaliser le clustering du même jeu de données en utilisant l'attribut **class** pour **évaluer** les clusters obtenus. Quels jeux de paramètres proposez-vous pour chaque programme pour que le clustering des données corresponde au mieux à leur classe ?

Rappelons que le clustering ne requiert pas d'attribut classe, cet attribut est simplement utilisé ici comme une vérité terrain et facilite l'évaluation et la comparaison des résultats du clustering.

Qu'est-ce que donne l'algorithme DBSCAN sur ce jeu de données ? Il faudra installer ce programme en utilisant le package manager de Weka.

Exercice 2

On va maintenant utiliser le jeu de données **Glass Identification**[1] qui comporte des données sur la composition en minéraux de plus de 200 échantillons de verre et correspondant à différentes types de verre (fenêtres, voitures, lampes, etc.).

1- Réaliser le meilleur clustering de ces données et évaluer la qualité de ce clustering par rapport à la valeur de SSE mais aussi par apport à l'attribut **Type.**

2- Proposer une caractérisation de chaque type de verre.

3- Est-ce que vous arrivez à peu près à la même conclusion en construisant un arbre de décision ? Pour cela, écrire les règles contenues dans l'arbre de décision (pour chaque type).

[1] https://archive.ics.uci.edu/ml/datasets/Glass+Identification