TP4 : clustering sous Weka.

Ilhem Badreddine.
Armand Bour.
Groupe 1, Master Informatique.

1- Jeux de données :

Q1. Analyse exploratoire via l'explorer :

Jeux de données	Nb instances	Nb attributs	Types
Iris	150	5	Numeric, sauf class Nominal
Spiral	311	3	Numeric, sauf class Nominal
Vote	435	17	Nominal

2- Découvert de Kmean

Q0. Lecture des résultats obtenus.

Q0.1. Les paramètres à fixer :

- Nombre de clusters = 3 car il y a trois classes d'iris d'après le fichier iris.affr.
- Ignorer l'attribut class qui regroupe déjà les iris par espèce (et donc fausse les clusters)
- On fixe le mode à Classes to clusters evaluation, pour tester le clustering relativement à la classe class, à fin d'avoir les correspondance entre cluster et classe, ainsi que le taux d'erreur.

Q.0.2. Les informations trouvées :

- le nombre d'instances
- le nombre d'attributs
- les attributs ignorés
- le mode utilisé
- le nombre d'itérations
- les centres des clusters
- le Within cluster : wc, compasité.
- la répartition des instances dans les clusters
- la répartition des classes dans les clusters
- l'assignation des classes aux clusters
- le taux d'erreurs par rapport à la classe de la vraie répartition

Q.1. Interprétation des résultats :

Q.1.1. Les centres des clusters sont les barycentre (moyenne) des attributs de l'ensemble des instances du cluster.

Q1.2. Incorrectly clustered instances =

$$\frac{NbInstancesMalPlacéesParRapportALaVraieRepartition}{NbTotaleDInstances}$$

Q.1.3. On calcule les distances entre l'instance et le centre de chaque cluster, l'instance appartient au cluster dont la distance est la plus courte.

Q2.

Q.2.1. Variation de seed : seed est le nombre de fois ou weka fait un aléatoire avant qu'il fixe ses centres initiaux, donc il influe sur les centres initiaux des clusters par conséquence sur les cluster

finaux et le taux d'erreur.

Q2.2. Plus y a de cluster plus l'évaluation est faussé car le nombre de cluster est différent du nombre de classe.
Pour Iris : l'évaluation des classe est moins efficace et le taux d'erreur est plus haut.

Q2.3. Désavantages et solution possible:
- les résultats varient selon les valeurs des centres initiaux choisit aléatoirement. Pour y remédier : choisir un seed pas trop grand, pas plus que 10, selon notre expérience.
- Il faut donner le nombre de clusters à l'algorithme, ce qui altère éventuellement l'evaluation.Pour y remédier :si possible, une analyser des fichiers des données peut indiquer un nombre idéale de cluster comme dans iris avec classe.

Q.3. Visualisation 2D :

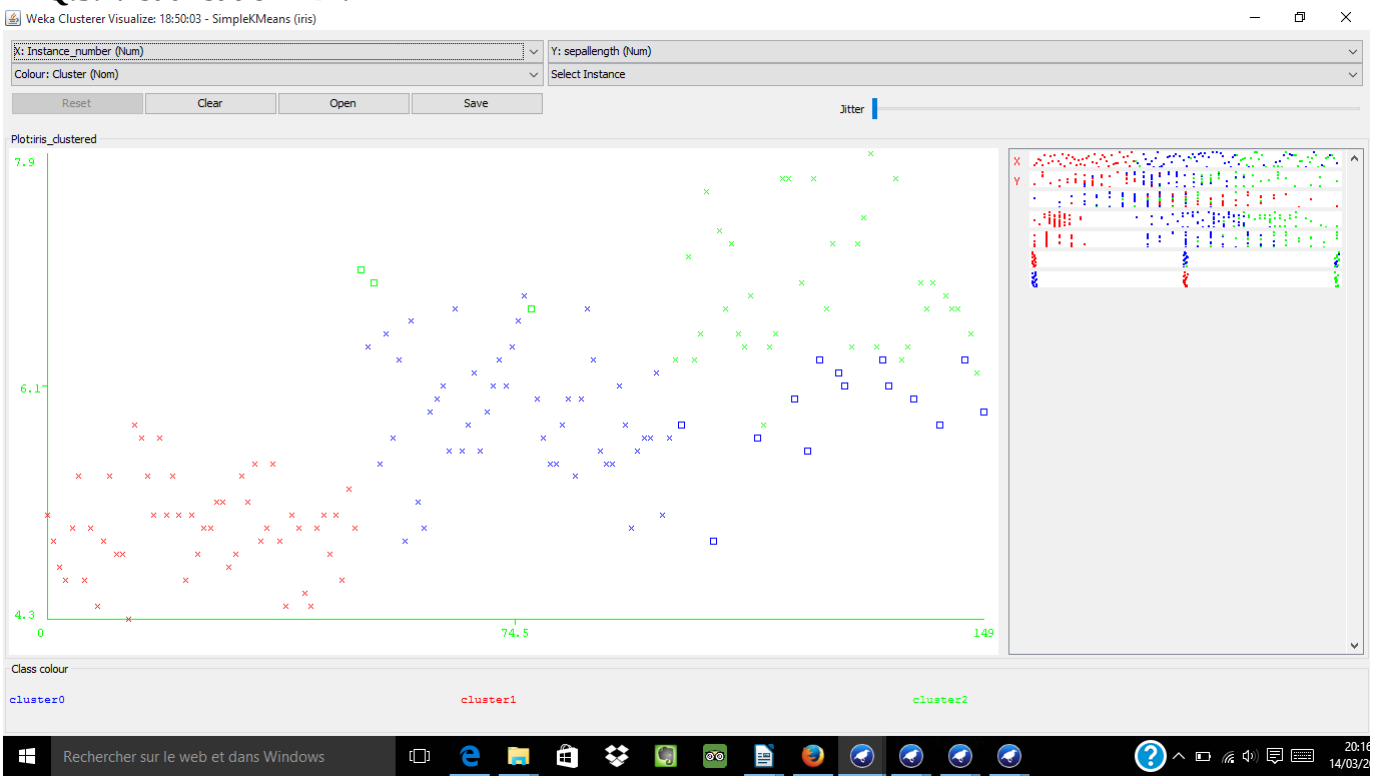

Figure 1 : SimpleKmean pour Iris.

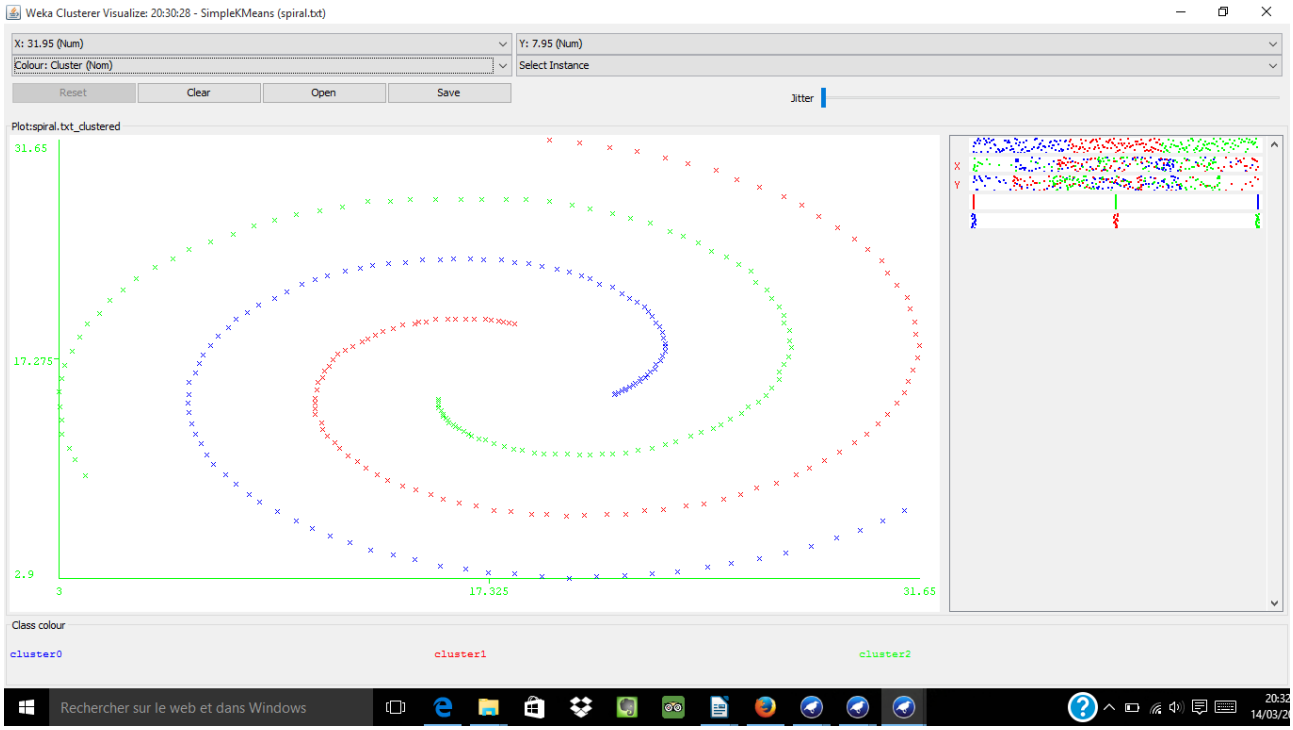

Figure 2 : SimpleKmean pour Spiral.

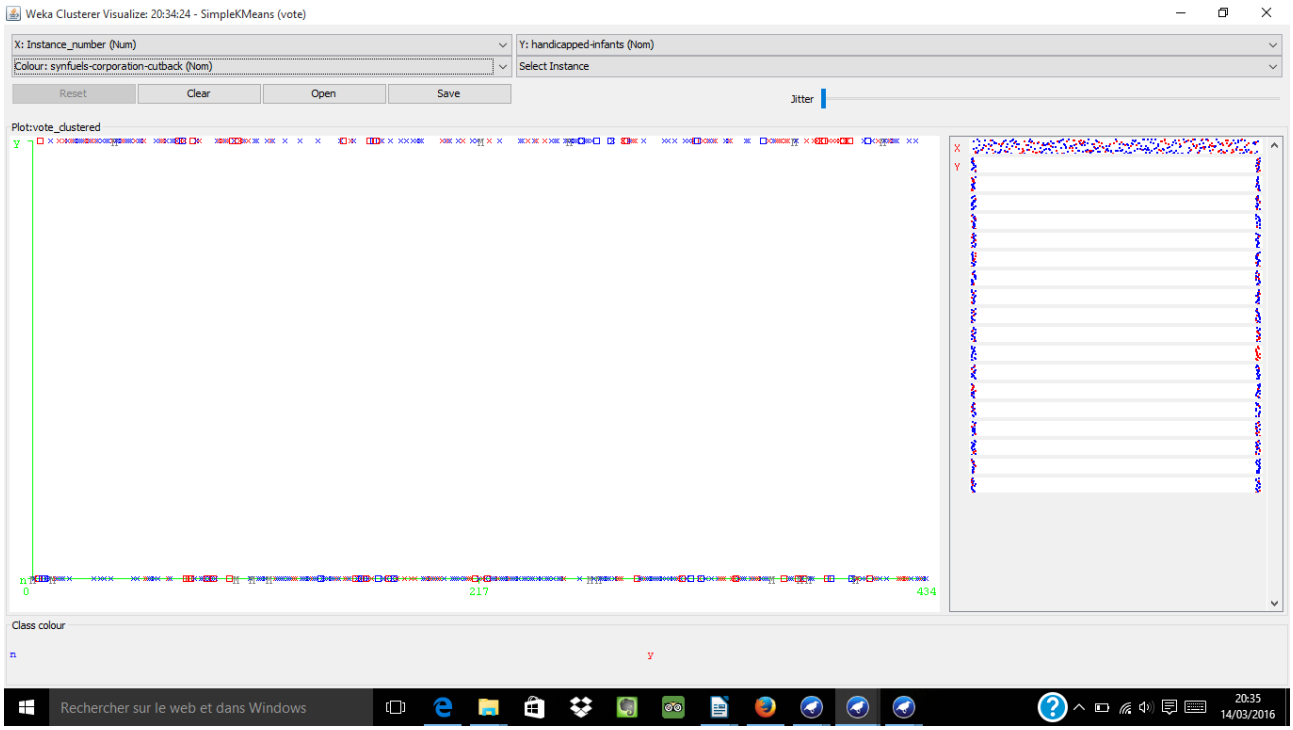

Figure 3 : SimpleKmean pour Vote.

Q3.3. En placant un point avec ses coordonnées, on peut observer directement le cluster auquel il appartient.

3. Comparaison des résultats :
Nous avons pris les algorithme présents pour les trois fichiers.

	Spiral (3 clusters)		Iris (3 clusters)		Vote (2 clusters)	
	WC	Taux d'erreur	WC	Taux d'erreur	WC	Taux d'erreur
CLOPE	-	-	-	-	-	57.0115 %
Cobweb	-	-	-	33.3333 %	-	92.6437 %
DBScan	-	-	-	66.6667 %	-	21.8391 %
EM	-	-	-	40 %	-	40.4598 %
Farthest First	-	-	-	33.3333 %	-	15.6322 %
Filtered Clusters	-	-	12.143688281579722	33.3333 %	-	14.023 %
Hierarchical Clusters	-	-	-	33.3333 %	-	38.6207 %
MakeDensityBasedClusterer	-	-	-	33.3333 %	-	12.6437 %
OPTICS	-	-	-	(0 cluster)	-	(0 cluster)
Sib	-	-	-	33.3333 %	-	-
Simple Kmeans	-	-	6.998114004826762	33.3333 %	1449.0	14.023 %
Xmean	-	-	-	33.3333 %	-	-