TP Clustering sous Weka

Rappel Modalité : *Vous effectuerez un compte rendu écrit des TPs. Ce compte rendu pourra être fait en binôme. Les éléments obligatoires apparaîtront au fur et à mesure des TP, ces éléments ne sont pas limitatifs et nous vous encourageons vivement à aller plus loin (interprétation des résultats, …)*

Jeux de données *: Les jeux de données avec lesquels vous allez travailler sont ceux fournis par l'UCI http://www.ics.uci.edu/~mlearn/MLSummary.html et ceux disponibles sur http://cs.joensuu.fi/sipu/datasets/*

Objectif TP : Prise en main de Weka, découverte de kmeans, interprétation des résultats, visualisation graphique.

Mission : Vous devez réaliser le meilleur clustering possible pour différents jeux de donnée. Pour cela vous testerez différents algorithmes avec différents paramètres et comparerez les résultats obtenus.

Prise en main de Weka

Weka est disponible sous linux, vous pouvez le télécharger dans votre environnement préféré.
En salle TP, Weka se trouve dans le menu Applications -> Sciences.

1- Choisir Explorer pour commencer
2- Charger un jeu de données (bouton Open file, cf figure 1).
3- A partir de l'onglet Cluster (cliquez sur cluster dans la barre du haut) ; on peut observer différents algorithmes implémentés et changer d'algorithme en cliquant sur choose (cf fig2).
La boîte Cluster mode permet de choisir la méthode d'évaluation du modèle extrait :

- **Use training set** : effectue et teste le clustering sur le même jeu de données ;
- **Supplied test set** : teste le clustering sur un jeu de données à spécifier ;
- **Percentage split** : effectue le clustering sur le pourcentage indiqué du jeu de données et teste sur le pourcentage restant ;
- **Classes to clusters evaluation** : permet d'assigner une classe à un cluster pendant la phase de test. La classe assignée est la plus fréquente dans le cluster ; une erreur de classement (taux de mal classés) est calculée ainsi que la matrice de confusion. Dans ce cas, l'algorithme ne prend pas en compte la valeur de cet attribut dans le calcul de distance.

L'accès aux options se fait en cliquant sur le nom de l'algorithme (voir figure 2).
4- A partir de la boîte Result List (click droit, *Visualize cluster assignement*), vous pouvez visualiser la répartition des exemples dans chaque cluster.

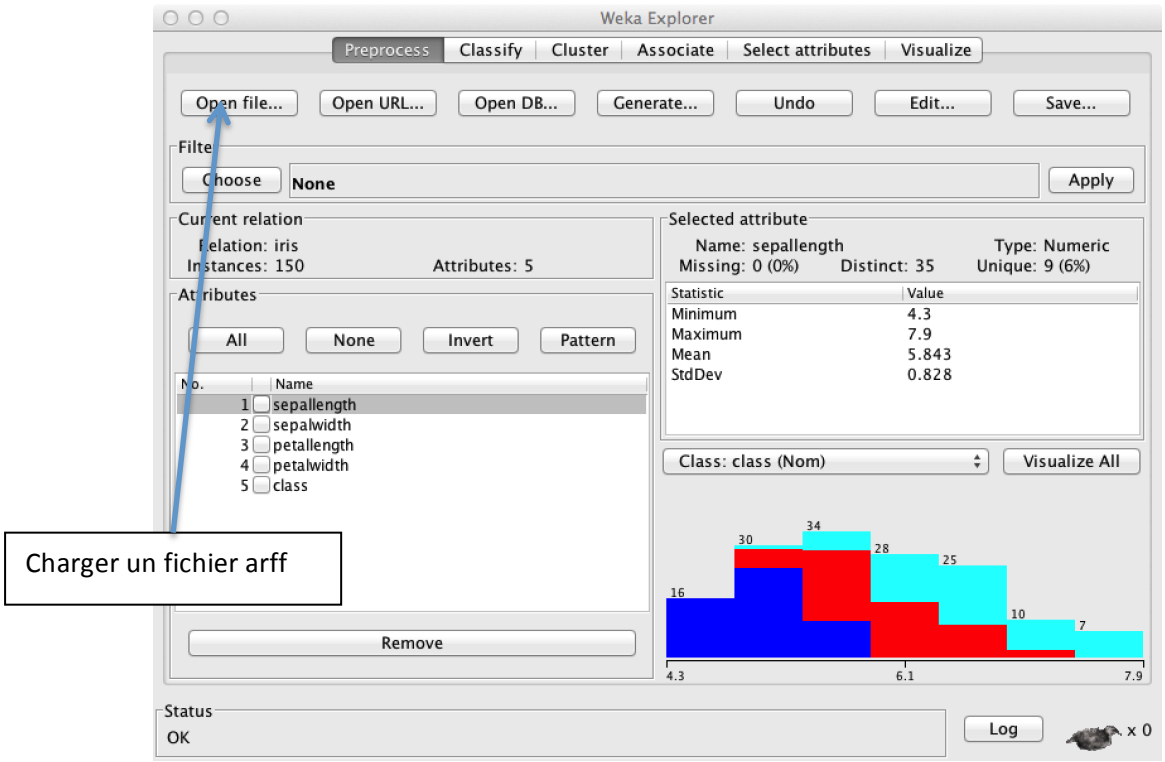


Figure 1 : ouverture d'un jeu de données

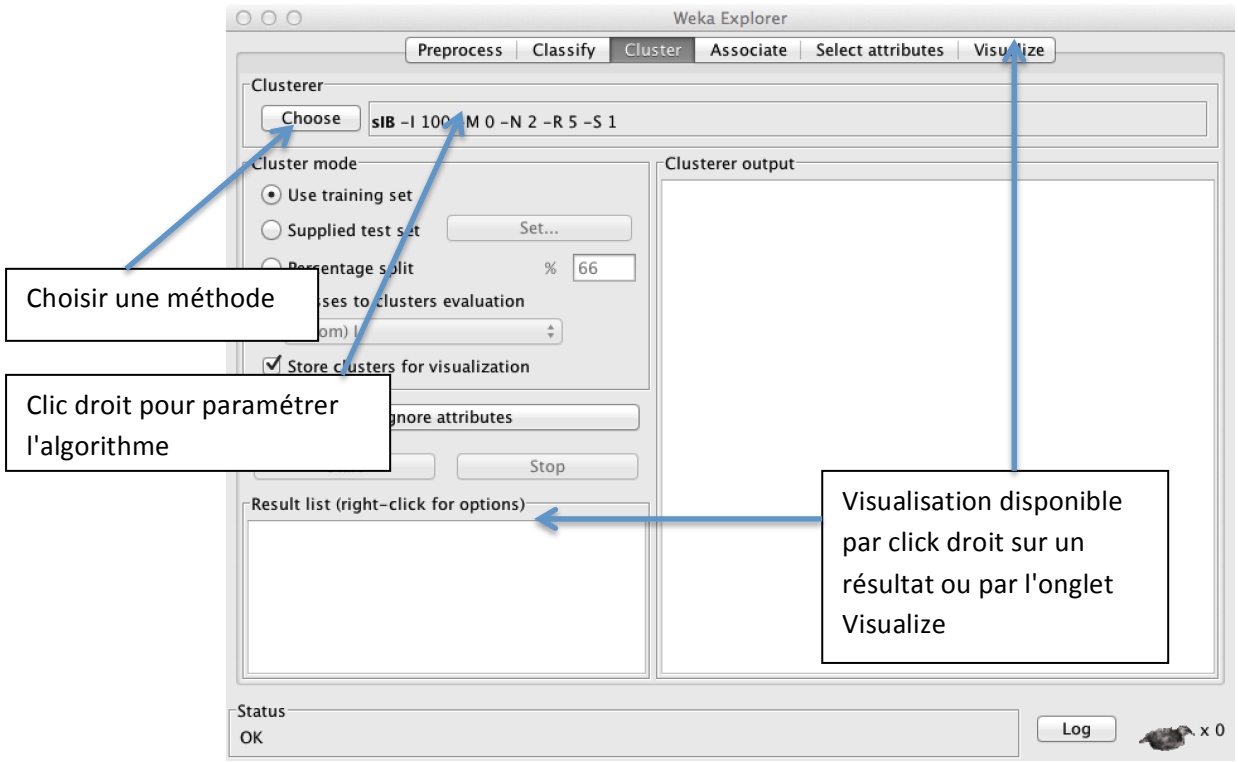


Figure 2 : paramétrage du clustering

Travail demandé

Jeux de données

Vous travaillerez avec les jeux de données suivants :

- **Iris** : décrit des mesures sur les sépales et pétales d'iris.
- **Spiral** : ensemble de points en coordonnées cartésiennes. Ce jeux de données est au format txt, il faut le convertir à la main en fichier arff, en ajoutant un entête (regardez l'entête de iris.arff pour comparer). Il faut éventuellement transformer l'attribut classe en nominal.
- **vote** : ce jeu de données décrit le résultat des votes de chaque représentant au Congrès des Etats-Unis sur les 9 questions clés identifiées par le Congressional Quarterly Almanac.

Question 1. Analyse exploratoire via l'explorer

Effectuez une première analyse des jeux de données. Combien d'instances ? Combien d'attributs ? Quels types ? Vous donnerez un tableau synthétisant la composition du jeu de données.

L'objectif du TP est de trouver le meilleur algorithme pour chaque jeu de données et de critiquer les algorithmes mis à disposition.

Découverte de Kmeans

Nous allons d'abord réaliser plusieurs kmeans sur différents ensembles de données. Toutes les données qui vous sont rendues disponibles, possèdent un attribut indiquant la classe. Cet attribut servira uniquement (dans le cas du clustering) pour réaliser des statistiques supplémentaires sur les résultats. Il faut donc ignorer cet attribut avant de lancer le kmeans (nommé *SimpleKmeans* dans Weka), sauf dans le mode *"Classes to clusters evaluation"*.

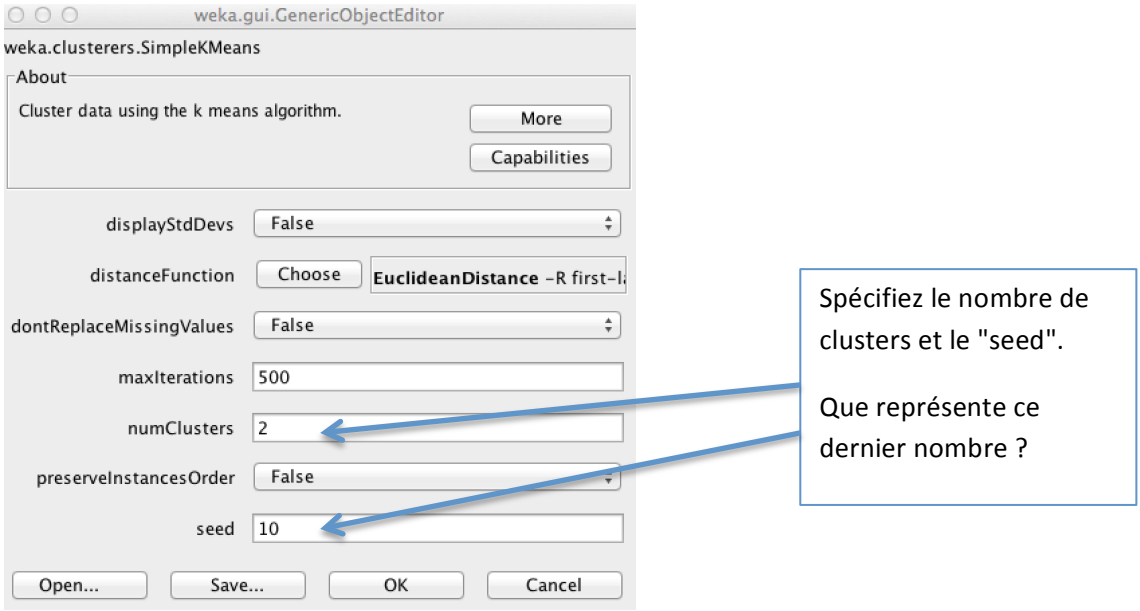


Figure 3 : paramètres de l'algorithme KMeans

Q0 : Lecture des résultats obtenus

Q0.1 Exécutez un Kmeans sur le jeu de données IRIS.

Quels paramètres devez vous fixer ? Comment les fixer ? (cf figure 3)

Q0.2 Voici un exemple de résultats obtenus sur le jeu de données IRIS. Que retrouve t-on comme information ?

```
=== Run information ===

Scheme:weka.clusterers.SimpleKMeans -N 3 -A "weka.core.EuclideanDistance -R first-last" -I 500
-S 10
Relation:     iris
Instances:    150
Attributes:   5
              sepallength
              sepalwidth
              petallength
              petalwidth
Ignored:
              class
Test mode:Classes to clusters evaluation on training data
=== Model and evaluation on training set ===


kMeans
======

Number of iterations: 6
Within cluster sum of squared errors: 6.998114004826762
Missing values globally replaced with mean/mode

Cluster centroids:
                         Cluster#
Attribute      Full Data          0          1          2
                   (150)       (61)       (50)       (39)
=========================================================
sepallength       5.8433     5.8885      5.006     6.8462
sepalwidth         3.054     2.7377      3.418     3.0821
petallength       3.7587     4.3967      1.464     5.7026
petalwidth        1.1987      1.418      0.244     2.0795
```

Métrique de qualité du clustering notée WC

Centres des clusters

```
Time taken to build model (full training data) : 0.01 seconds

=== Model and evaluation on training set ===

Clustered Instances

0       61 ( 41%)
1       50 ( 33%)
2       39 ( 26%)


Class attribute: class
Classes to Clusters:

  0  1  2  <-- assigned to cluster
  0 50  0 | Iris-setosa
 47  0  3 | Iris-versicolor
 14  0 36 | Iris-virginica

Cluster 0 <-- Iris-versicolor
Cluster 1 <-- Iris-setosa
Cluster 2 <-- Iris-virginica

Incorrectly clustered instances :    17.0     11.3333 %
```

Tableau de contingence → nécessite de connaître la vraie répartition

Weka assigne à chaque cluster trouvé la classe → comment procède weka ici ?

Taux d'erreurs par rapport à la classe de la vraie répartition

Q1 . Interprétation des résultats

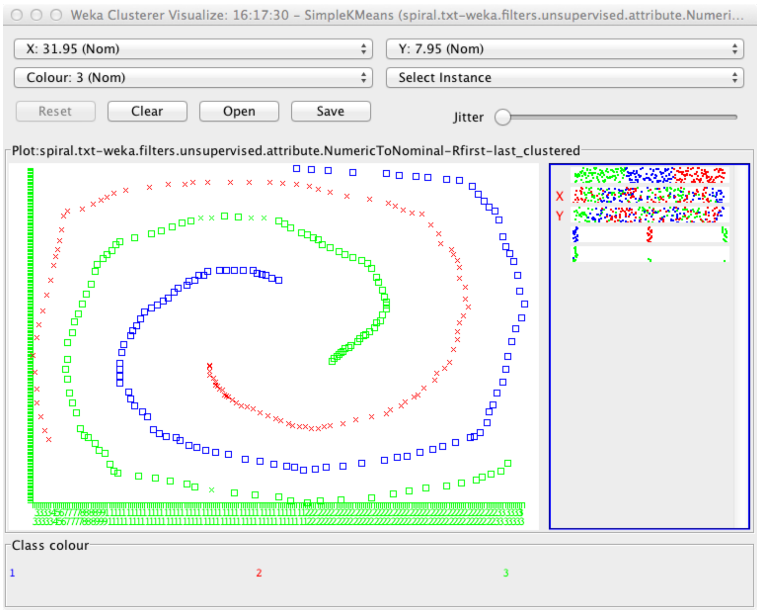
Q1.1 Comment interprétez vous le centre des clusters, quelles informations vous donne t-il ?
Q1.2 Comment est calculé le « Incorrectly clustered instances » ?
Q1.3 Comment feriez-vous pour retrouver l'appartenance des instances à un cluster ?

Q2 . Faites varier les paramètres de kmeans sur les trois jeux de données

Q2.1 Que remarque t-on lorsqu'on fait varier le seed ?
Q2.2 Que remarque t-on lorsqu'on fait varier le nombre de clusters ?
Q2.3 Quels désavantages du kmeans sont montrés dans ce simpleKMeans de Weka (justifier) ? Que proposez vous pour palier à ces désavantages ?

Q3. Visualisation 2D du clustering

A partir de la boîte Result List (bouton droit, "Visualize cluster assignement"), visualiser la répartition des exemples dans chaque cluster. Les croix représentent les instances classées dans le "bon" cluster et les carrés représentent les instances classées dans le "mauvais" cluster.



Q3.1 Recherchez des visualisations 2D intéressantes.
Q3.2 Comment représenteriez-vous les frontières des clusters?
Q3.3 Comment utiliser cette représentation 2D pour classer une nouvelle instance ?

Comparaison de résultats - Tableau à rendre comme compte-rendu de TP.

Remplissez le tableau ci-dessous avec différents algorithmes et différents paramètres de chaque algorithme, sur les 3 jeux de données.
Déduisez-en les meilleurs choix à faire pour chaque jeu de donnée.

	Spiral		Iris		Vote	
	WC	Taux d'erreur	WC	Taux d'erreur	WC	Taux d'erreur
Algorithme 1 et paramètres						
Algorithme 2 et paramètres						
Algorithme 3 et paramètres						
….						