Université Paris 8

Master 2 Big Data et fouille de données

Date : 15/10/2016

Rapport
Projet K-means

Réalisé par

Redha Nabil BELKHOUS

Numéro d'étudiant : 16705491

Enseignants

Nourredine ALIANE & Athmane MENADE

Table des matières

1 Introduction **3**

2 Algorithme K-means **3**
2.1 Description de l'algorithme 3
2.2 Implémentation de K-means 4
2.3 Résultats obtenus 4
2.3.1 Exemple du cours 4
2.3.2 K-means avec les données du fichier texte data.txt 5

3 Remarques **8**

1 Introduction

Dans le cadre du cours de **Modèles formels pour le Big Data**, l'étude d'algorithmes de Datamining est une étape importante pour comprendre l'aspect théorique et pratique des différentes techniques existantes.

Dans cette optique, il nous a était demandé d'implémenter l'algorithme K-means avec les indices **CH** et **DB**.

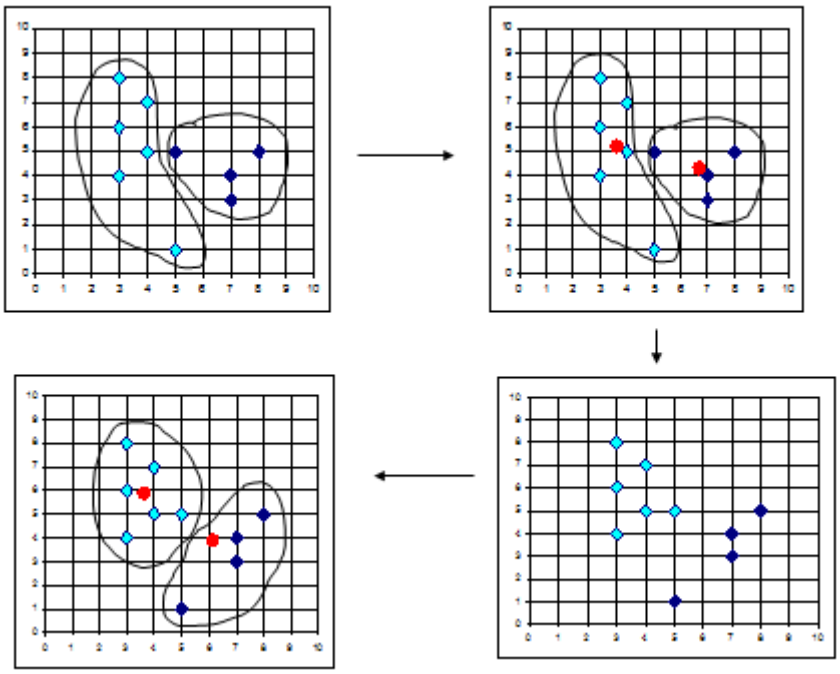
2 Algorithme K-means

2.1 Description de l'algorithme

C'est un algorithme qui appartient à la famille de l'apprentissage non supervisé où les clusters ne sont pas prédéfinis préalablement.

L'algorithme k-means est en 4 étapes :

— Choisir k objets formant ainsi k clusters.
— (Ré)affecter chaque objet O au cluster Ci de centre Mi tel que dist(O,Mi) est minimale.
— Recalculer Mi de chaque cluster (le centroide).
— Aller à l'étape 2 si il n'y a pas convergence.



2.2 Implémentation de K-means

Le choix technique de l'implémentation qui a été choisi est le langage **Java** car c'est une simple manipulation algorithmique et non pas un programme qui traitera de grands volumes de données.

Le programme est enrichi avec un rendu visuel à la fin du déroulement de l'algorithme K-means qui montre les centrioids et les clusters finaux.

Explication de la méthode d'implémentation utilisée :

- Une classe principale **KMeans** qui est le coeur de l'algorithme et qui reprend les étapes de ce dernier.
- La classe **Data** qui représente la structure de nos données (des points dans notre cas).
- La classe **Centroid** qui fait référence au centre de gravité qu'on peut trouver lors du déroulement de l'algorithme k-means.
- pour ce qui est de l'algorithme K-means, il suit scrupuleusement les étapes présenté précédemment dans la description avec en plus le calcul d'indice de qualité **CH** qui requière le calcul préalable de **la mesure de cohérence W** ainsi que **la mesure de séparation B**
- Il faut que l'utilisateur détermine le nombre de clusters K.
- pour plus de rapidité et de lisibilité la console affiche les clusters et les centroids finaux.
- Un affichage des clusters et des centroids dans un plant à deux dimensions s fait grace à la bibliothèque open source XYChart de java

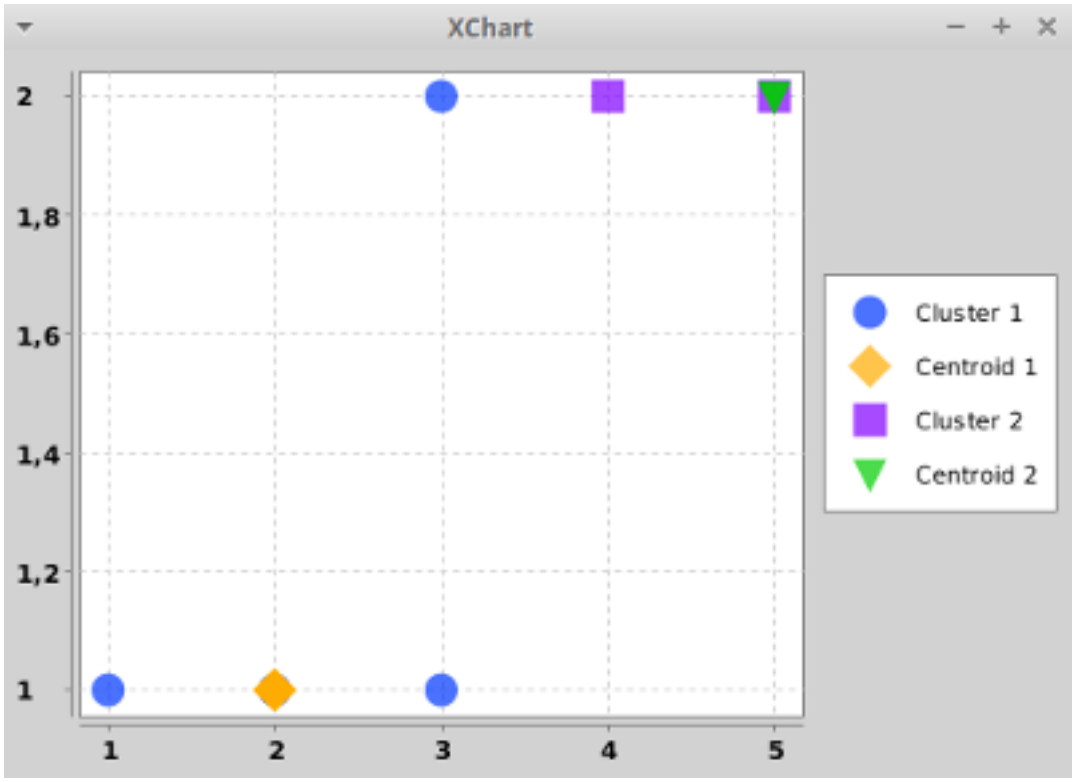
2.3 Résultats obtenus

2.3.1 Exemple du cours

Afin de s'assurer du bon fonctionnement de l'implémentation de l'agorithme K-means, on va prendre l'exemple traité en cours :

- Soient les points : **A(1.0,1.0)** , **B(2.0,1.0)**, **C(3.0,1.0)** ,**D(3.0,2.0)** , **E(4.0,2.0)**, **F(5.0,2.0)**.
- En prenant le nombre de clusters K=2, et en fixant les centroides à $\mathbf{C}_1(4.0, 2.0) et \mathbf{C}_2(5.0, 2.0)$.

```
Donner le K compris entre 1 et 8 pour le nombre de clusters
2
Centroid 1 x: 4.0 y: 2.0
Centroid 1 x: 5.0 y: 2.0
Cluster 0 includes:
      (1.0 , 1.0)
      (2.0 , 1.0)
      (3.0 , 1.0)
      (3.0 , 2.0)

Cluster 1 includes:
      (4.0 , 2.0)
      (5.0 , 2.0)

W = 4.414213562373095
B = 8.595241580617241
CH = 1.5577391459051042
DB = 0.48
**************************************************************

Centroids finalized at:
      (2.0) , (1.0)
      (5.0) , (2.0)
```



On remarque que les résultats obtenus sont bien les mêmes obtenus en cours avec l'exemple de présentation de l'algorithme.

2.3.2 K-means avec les données du fichier texte data.txt

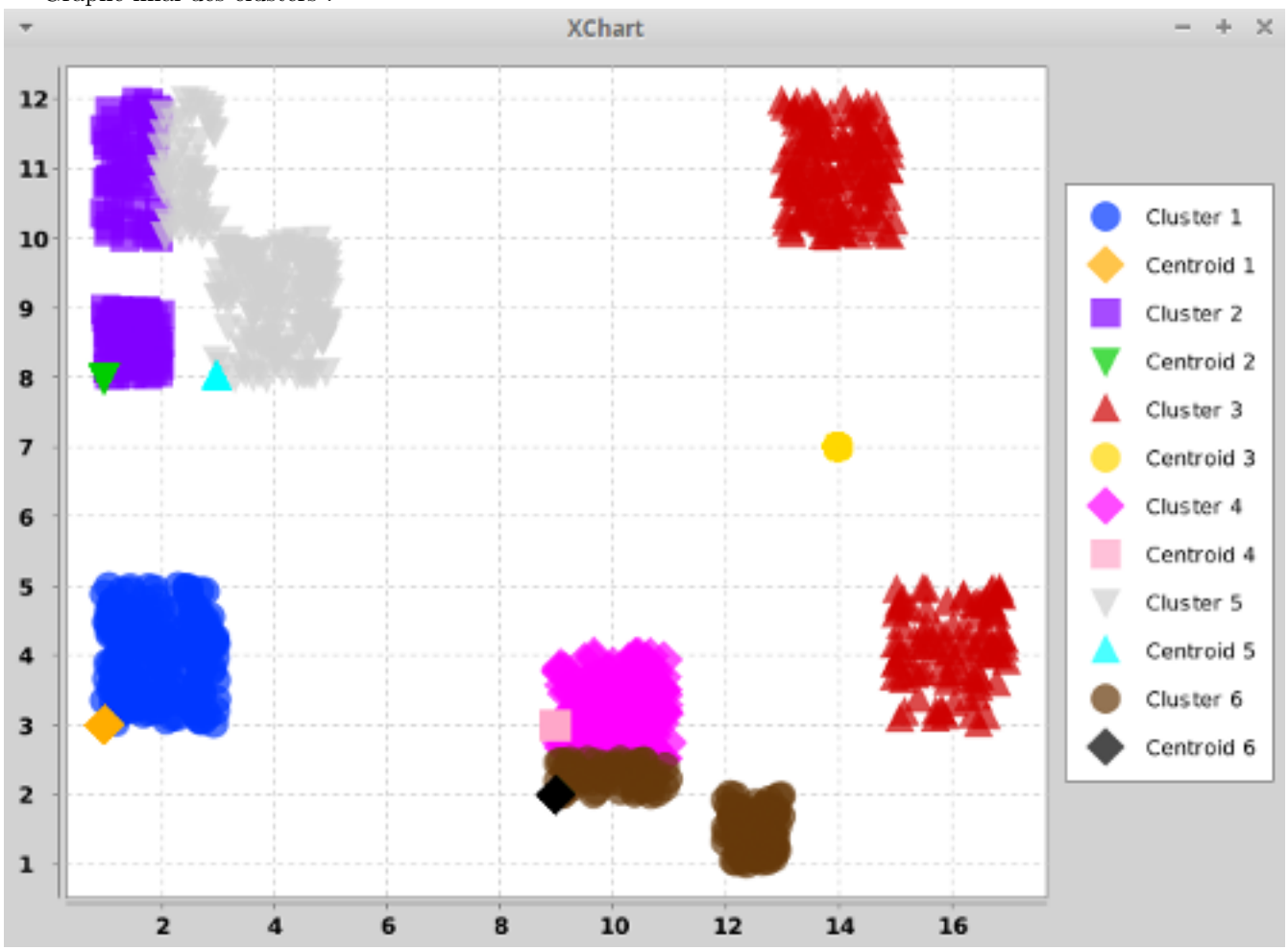
Avec 6 clusters Le choix aléatoire des centroids initiaux donné :

```
Donner le K compris entre 1 et 8 pour le nombre de clusters
6
Centroid 1 x: 0.68 , y: 1.9

Centroid 2 x: 1.22 , y: 7.46

Centroid 3 x: 9.74 , y: 9.23

Centroid 4 x: 6.67 , y: 6.15

Centroid 5 x: 5.51 , y: 7.68

Centroid 6 x: 5.84 , y: 3.27
```

Le résultat des clalculs d'u=indice DB et CH :

```
W = 2460.6227529900666
B = 6247.488365486002
CH = 2.527539478663387
DB = 0.2
*****************************************

Centroids finalized at:
     (1.0) , (3.0)
     (1.0) , (8.0)
     (14.0) , (7.0)
     (9.0) , (3.0)
     (3.0) , (8.0)
     (9.0) , (2.0)
```

Graphe final des clusters :



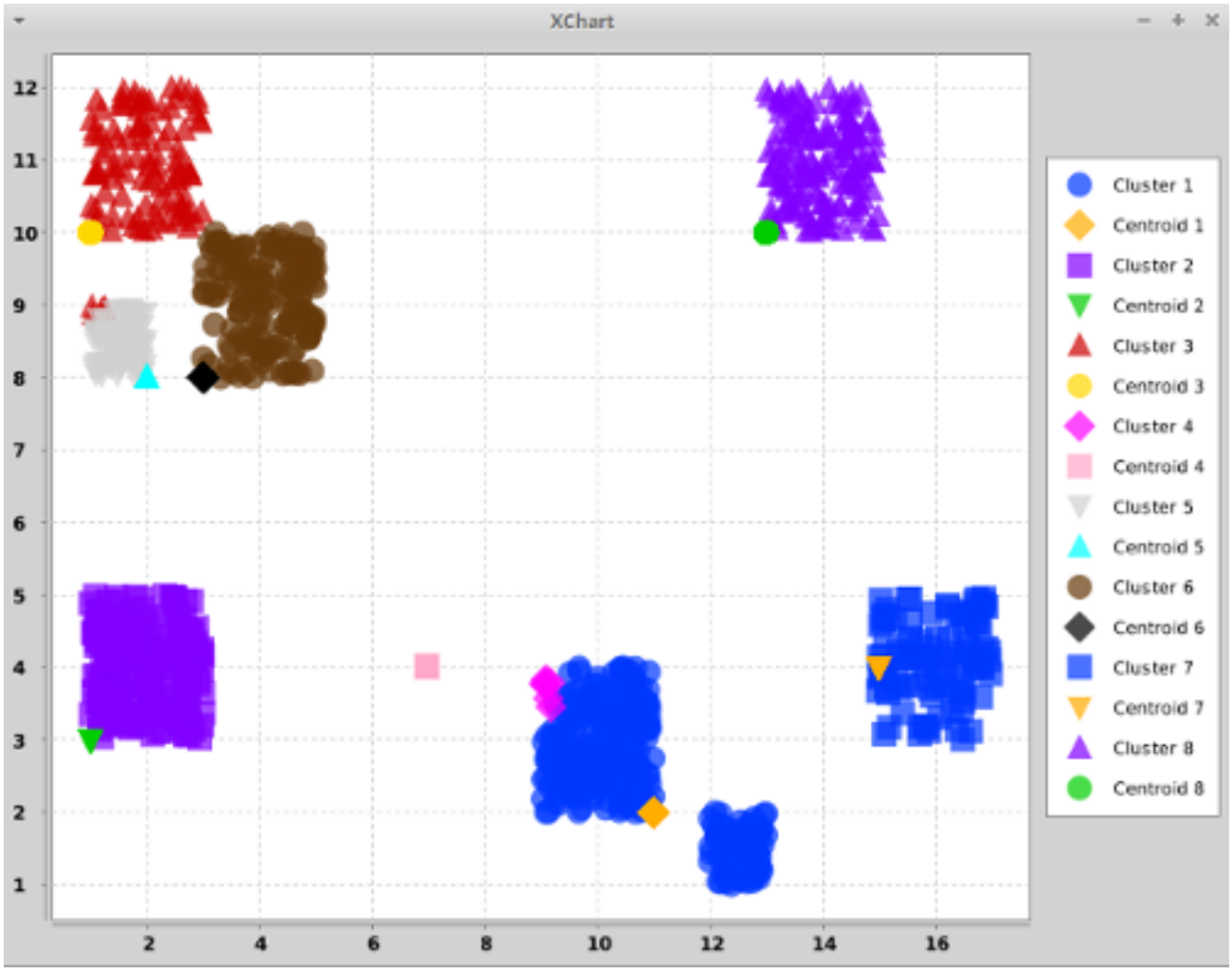
Avec 8 clusters Le choix aléatoire des centroids initiaux donné :

```
Donner le K compris entre 1 et 8 pour le nombre de clusters
8
Centroid 1 x: 9.86 , y: 2.13

Centroid 2 x: 4.05 , y: 5.69

Centroid 3 x: 1.72 , y: 9.76

Centroid 4 x: 6.98 , y: 4.01

Centroid 5 x: 2.44 , y: 8.5

Centroid 6 x: 2.76 , y: 7.64

Centroid 7 x: 9.81 , y: 6.57

Centroid 8 x: 9.82 , y: 9.17
```

Le résultat des clalculs d'u=indice DB et CH :

```
W = 1608.5042542151527
B = 6981.583029656157
CH = 4.313022679148038
DB = 0.53
**********************************

Centroids finalized at:
      (11.0) , (2.0)
      (1.0) , (3.0)
      (1.0) , (10.0)
      (6.98) , (4.01)
      (2.0) , (8.0)
      (3.0) , (8.0)
      (15.0) , (4.0)
      (13.0) , (10.0)
```

Graphe final des clusters :



Commentaires sur les résultats :

- J'ai constaté que les centres de gravité ont changé tout au long du déroulement de l'algorithme jusqu'à arrivé au résultat présenté ci-dessus.
- J'ai pu remarquer que l'indice CH était plus petit d'itération en itération.
- J'ai remarqué que quand l'indice DB était petit la qualité de la segmentation était mailleur.
- J'ai aussi remarqué que l'indice DB ne dépondait pas du nombre de clusters.
- les indices DB et CH change d'une distribution à une autre, c.a.d que même pour un nombre de clusters, le résultat dépond des centroids initiaux qui vont être choisis aléatoirement au début de l'algorithme.
- au niveau des graphes de résultats, pour un même nombre de clusters, la distribution de chacun n'est pas la même lors de chaque exécution.
- J'ai eu quelques exemples d'exécution ou j'avais un centroid qui ne contenait aucun point.
- La qualité du clustering dans l'algorithme K-Means dépond beaucoup du choix des centroids initiaux.
- J'ai remarqué que la plus l'indice DB est petit meilleur est la qualité de clustering.

3 Remarques

Vous trouverez les détails d'implémentations dans le code sources qui est soigneusement documenté (Classes, variables, méthodes, ...)