# Proposition d'une solution au problème d'initialisation cas du K-means

Z.Guellil[1] et L.Zaoui[2]

[1,2]Université des sciences et de la technologie d'Oran MB, Université Mohamed Boudiaf USTO -BP 1505 El Mnaouer -ORAN - Algérie
[1] g.zouaoui@gmail.com, [2] Zaoui_Lynda@yahoo.fr

**Résumer.** Dans cet article nous présentons une simple technique d'initialisation du k-means dans le but de maximiser la séparabilité et la compacité des groupes et ceci en choisissant l'objet le plus mal classé comme étant le nouveau centre du groupe. L'utilisation de cette technique peut se faire soit par intégration directe dans la phase d'initialisation du classique k-means ou bien dans une approche incrémental comme dans le cas du global k-means. Nos expériences montrent que cette technique produit des groupes qui sont similaire a celle du global k-means en un temps très réduit.

**Mot-clés :** Analyse des données, clustering, optimisation, k-means, Global optimisation.

## 1 Introduction

Le partitionnement des données est une tache importante en analyse de données, elle divise un ensemble de données en plusieurs sous ensembles, ces sous ensembles appelés groupes ou clusters. Ces groupes sont caractérisés idéalement par une forte similarité à l'intérieur et une forte dissimilarité entre les membres de différents groupes [3].

L'usage de cette technique vise à identifier un résumé de la structure interne de ces données, sans aucune connaissance a priori sur les caractéristiques des données [1]. Cela touche plusieurs domaines dont la reconnaissance des formes, l'imagerie, la bioinformatique et l'indexation des bases d'images.

Dans ce cadre plusieurs méthodes ont été développées, la plus populaire est celle des k moyennes (K-means), elle doit sa popularité à sa simplicité et sa capacité de traiter de larges ensembles de données [4].

Cependant, la principale limite de cette méthode est la dépendance des résultats des valeurs de départ (centres initiaux). À chaque initialisation correspond une solution différente (optimum local) qui peut dans certain cas être très loin de la solution optimale (optimum global). Une solution naïve à ce problème consiste à lancer l'algorithme plusieurs fois avec différentes initialisations et retenir le meilleur regroupement trouvé. L'usage de cette solution reste limité du fait de son coût et que l'on peut trouver une meilleure partition en une seule exécution.

Dans cet article, nous présentons dans la section 2 les algorithmes k-means et global k-means. La section 3 décrit nos solutions au problème d'initialisation de ces algorithmes, enfin les résultats de notre expérimentation sont décrits dans la section 4.

## 2 k-means et le Global k-means

K-means définit par McQueen [2] est un des plus simples algorithmes de classification automatique des données. L'idée principale et de choisir aléatoirement un ensemble de centres fixé a priori et de chercher itérativement la partition optimale. Chaque individu est affecté au centre le plus proche, après l'affectation de toutes les données la moyenne de chaque groupe est calculé, elle constitue les nouveaux représentants des groupes, lorsqu'ont abouti à un état stationnaire (aucune donnée ne change de groupe) l'algorithme est arrêté.

**Algorithme** 1 : K-means
**Entrée**
Ensemble de N données, noté par x
Nombre de groupes souhaité, noté par k
**Sortie**
Une partition de K groupes $\{C_1, C_2, ...C_k\}$
**Début**
1) Initialisation aléatoire des centres $C_k$ ;
**Répéter**
2) Affectation : générer une nouvelle partition en assignant chaque objet au groupe dont le centre est le plus proche ;

$$x_i \in C_k \; si \; \forall j \left| x_i - \mu_k \right| = \min_j \left| x_i - \mu_j \right| \quad (1)$$

Avec $\mu_k$ le centre de la classe K ;
3) Représentation : Calculer les centres associe à la nouvelle partition ;

$$\mu_k = \frac{1}{N} \sum_{x_i \in C_k} x_i \quad (2)$$

**Jusqu'à** convergence de l'algorithme vers une partition stable ;
**Fin.**

Ce processus tente de maximiser la similarité intra-class représentée sous forme d'une fonction objective :

$$J = \sum_{i=1}^{K} \sum_{x_j \in C_i} d(x_j, C_i) \quad \textbf{(3)}$$

Dans le cas de la distance euclidienne cette fonction est appelée fonction d'erreur quadratique.

$$J = \sum_{i=1}^{K} \sum_{x_j \in C_i} \left\| x_j - C_i \right\|^2 \quad \textbf{(4)}$$

Global k-means [5] est une solution au problème d'initialisation du k-means, elle est fondé sur les données et vise à atteindre une solution globalement optimale. Elle consiste à effectuer un clustering incrémental et à ajouter dynamiquement un nouveau centre suivi par l'application du k-means jusqu'à la convergence.

Les centres sont choisis un par un de la façon suivante : le premier centre est le centre de gravité de l'ensemble des données (résultat de l'application du k-means avec k=1), les autres centres sont tirés de l'ensemble de données ou chaque donnée est une candidate pour devenir un centre, cette dernière sera testée avec le reste de l'ensemble, le meilleur candidat est celui qui minimise la fonction objectif (4), l'algorithme suivant permet d'illustrer le principe :

**Algorithme 2 :** Global k-means
**Entrée**
Ensemble de N données, notés par x ;
Nombre de groupes souhaiter, noté par k ;
**Sortie**
Une partition de K groupes $\{C_1, C_2, ...C_k\}$
**Début**
1) $C_1$ = Centre de gravité de l'ensemble des données ;
**Répéter**
2) Initialiser les centres i-1 par le résultat de l'étape précédente ;
3) Trouver l'$i^{éme}$ centre :
**Pour chaque** donnée x **faire**
3.1) Considère x comme étant le $i^{éme}$ centre ;
3.2) Affecter les données aux plus proche centre ;
3.3) Calculer l'erreur quadratique pour $C_i$=x ;
**Fin faire**
3.4) Garder le centre $C_i$ = x qui minimise l'erreur quadratique ;
4) Appliquer le k-means jusqu'à la convergence ;
**Jusqu'à** obtenir une partition en k groupes ;

**Fin.**

Les auteurs ont remarqué que cette solution été lourde à cause de la stratégie de choix du nouveau centre, ils ont proposé le Fast global k-means avec une nouvelle stratégie permettant d'accélérer le global k-means, cette stratégie garde la même philosophie que sa précédente (toutes les données peuvent être candidates pour devenir un centre), mais évite d'affecter les données aux centres le plus proche (centres déjà existant en plus du centre candidat) et de calculer l'erreur quadratique, sachant que l'erreur quadratique diminue en fonction du nombre de centre par un taux bn, le nouveau centre sera le candidat qui maximise ce taux.

$$\mathrm{b_n} = \sum_{i=1}^{N} \max\left(d_{k-1}^{i} - \left\|x_n - x_i\right\|^2, 0\right) \quad \textbf{(5)}$$

Avec $d_{k-1}^{i}$ la distance entre $x_i$ et son plus proche centre parmi les k-1 centres. Selon les expérimentations des auteurs, cette technique améliore le temps d'exécution et assure de bons résultats presque aussi bon que ceux fournis par la stratégie précédente.

## 3 Stratégie d'initialisation

Nous proposant une stratégie d'initialisation qui se base sur l'individu le plus mal classé, l'algorithme d'initialisation peut s'appliquer lors de l'initialisation d'une classification ou au cour de la classification (voir section 3.2).

### 3.1 Calcul des nouveaux centres

L'absence d'un signe indiquant si l'optimum global est atteint ou pas fait penser à la possibilité d'améliorer les résultats.

Observant l'équation (1), un objet est affecté à un groupe s'il lui est le plus proche, plus la distance diminue plus la probabilité d'appartenance à ce groupe augmente, dans le cas contraire, l'objet le plus loin de son groupe d'appartenance est considéré comme étant mal classé, il fera certainement un bon candidat afin de former le nouveau centre.

Le global k-means est amorcé par un seul groupe ayant pour représentant le centre de gravité de l'ensemble des données, dans certain cas, cette partie de l'espace est vide (figure 1) ce qui permet de dégrader la classification, nous proposons d'amorcer l'initialisation du k-means avec deux groupes, les centres de ces groupes doivent assurer la séparabilité des données au cours de classification, il est évidant de choisir les deux données les plus éloignées.

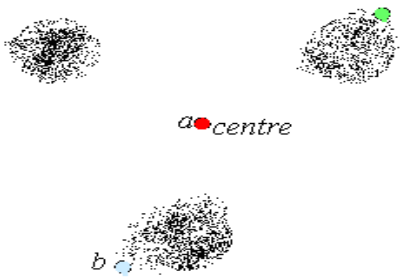


**Fig. 1.** (a) Le centre des données en rouge, (b) le bleu et le vert représentent les deux objets les plus éloigné.

Ce principe est illustré par l'algorithme suivant :

```
Algorithme 3 initialisation par le mal classé.
Début
    1) Création d'une matrice de distance
    2) Choisir les deux éléments les plus éloignés
       (ils représentent les deux premiers centres) ;
    TANT QUE le nombre de classes souhaité n'est pas
    atteint Faire
       3) Affecter les individus aux noyaux disponibles ;
       4) Sélectionner un élément mal classé (celui qui
          possède la plus grande distance de son centre
          le plus proche) ;
       5) Ajouter cet individu à l'ensemble des
          noyaux ;
       6) Augmenter le nombre des noyaux ;
    Fin TANT QUE
Fin.
```

### 3.2 L'approche incrémental

Notre approche incrémental de classification est similaire à celle du globale k-means, la différence entre elles réside dans les points suivant :

- Le nombre de points initiaux, dans notre cas deux au lieu de un seul dans le global k-means.
- La recherche du nouveau centre ce limite à la recherche de l'élément le mal classé au lieu de testé toutes les données.

```
Algorithme 4 : Approche incrémentale de classification
Entrée
     Ensemble de N données, notés par x ;
     Nombre de groupes souhaiter, noté par k ;
Sortie
```

Une partition de K groupes $\{C_1, C_2, ...C_k\}$

**Début**

1) $C_1 = x_1;$ $C_2 = x_2;$ Avec $d(x_1, x_2) = \max_{\substack{i,j \in [1..N] \\ i \neq j}} (d(x_i, x_j))$

**Répéter**

2) Initialiser les centres i-1 par le résultat de l'étape précédente ;

3) Trouver l'$i^{éme}$ centre $C_i$:

$$C_i = x : x = \max_{i \in [1,N]} (d_{k-1}^i) \quad (6)$$

4) Appliquer le k-means jusqu'à la convergence ;

**Jusqu'à** obtenir une partition en k groupes ;

**Fin.**

Grâce au faible cout de la stratégie de choix du nouveau centre, il est clair que l'approche proposée est plus rapide que le global k-means.

## 4 TESTE ET RESULTAT

L'algorithme proposé a été testé avec la version rapide du globale k-means et le k-means initialisé avec l'algorithme (3), Init k-means, sur des données artificielles et sur des images couleurs.

Dans les deux cas nous avons exécuté le faste global k-means et le modified fast global k-means pour un nombre de groupe k = 15, le "init k-means" est lancé pour k = 2, 3, …16, Pour chaque valeur de k nous nous calculons le temps d'exécution et des indices de la qualité, la qualité est évaluée en fonction des deux indices suivant :

Le premier indice est l'erreur quadratique moyenne, c'est l'indice le plus couramment utilisé pour mesurer la compacité des groupes, de faibles valeurs de cet indice indique une fort compacité des groupes.

$$E = \frac{1}{N} \sum_{i=1}^{K} \sum_{x_j \in C_i} d(x_j, C_i) \quad (7)$$

Le deuxième indice est celui de Davies-Bouldin (DB), il permet de mesurer la compacité et la séparabilité des groupes, de petites valeurs du DB sont indicatives de la présence de groupes compacts et bien séparés.

$$DB = \frac{1}{k} \sum_{i}^{k} \max_{j \neq i} \left( \frac{S(C_i) + S(C_j)}{d(C_i, C_j)} \right) \quad (8)$$

Où

$$S(C_i) = \frac{1}{card(C_i)} \sum_{x \in C_i} d(x_j, C_i) \qquad \textbf{(9)}$$

D'après notre expérimentation on, on constate que le Fast global k-means est plus coûteux en terme temps d'exécution, ce temps augmente en fonction du nombre de classes. Or dans l'algorithme init k-means, ce nombre n'influe pas sur le temps d'exécution, ceci étant dû au nombre d'itérations effectuées pour obtenir une stabilisation (notons que ce nombre dépend des points de départ et du type de données, il augmente lorsque les données sont denses).

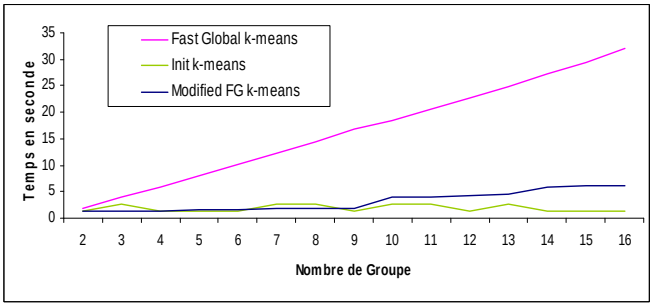


**Fig. 2.** Temps d'exécution en fonction du nombre de groupes.

En terme erreur quadratique, les solutions des trois approches possèdent la même qualité avec de légères différences, nous avons remarqué que lorsque l'erreur quadratique d'une des deux approches proposées est plus élevée que celle du global k-means (cluster moins compact), l'Indice DB diminue ce qui correspond à une forte séparation des clusters (voir nombre de groupe égal à 6 pour init k-means et 8 pour le modified fast global k-means).

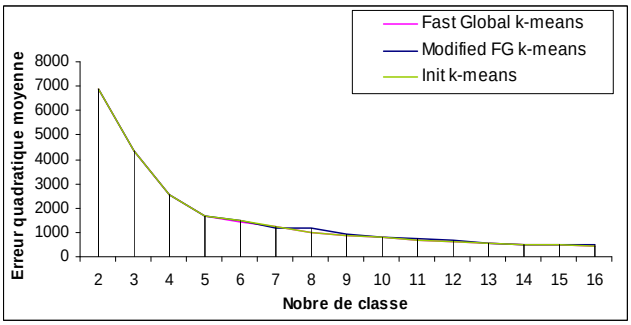


**Fig. 3.** Erreur quadratique moyenne en fonction du nombre de groupes.

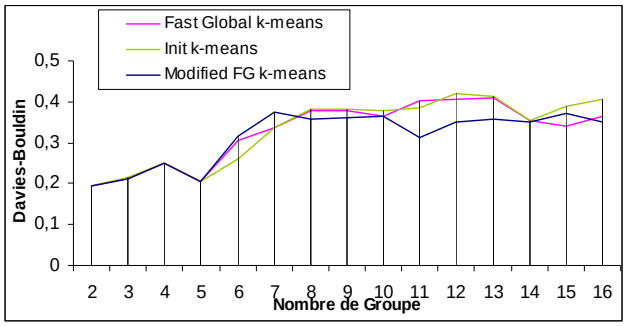


**Fig. 4.** Indice DB en fonction du nombre de groupes.

La stratégie de choix du nouveau centre de groupe du global k-means favorise l'élément qui peut attirer le maximum d'objet ce qui pénalise des petits groupes qui sont plus importants dans certain cas, alors que le choix de l'élément mal classé permet de les détecter, comme le montre la figure 5, a gauche l'image originale suivit du résultat de classification en six classes par l'algorithme Fast global k-means et le modified Fast global k-means.

**Fig. 5.** Exemple de classification en 6 classes.

## 5 Conclusion

Dans ce papier, nous avons proposé une solution au problème d'initialisation. Sachant que le principe du k-means est de minimiser la similarité intra classe (compacité des groupes), ce qui ne conduit pas forcément à une maximisation de similarité inter classe (séparabilité), l'approche proposée vise à maximiser la séparation des groupes ainsi que la compacité en appliquant le k-means. Notons, enfin, que cette méthode est applicable avec tout algorithme nécessitant des valeurs de départ (EM, PAM,…).

Malheureusement, cette technique est sensible au bruit, ceci arrive lorsque le nouveau centre choisi est un bruit (l'objet le plus loin du centre de la classe d'appartenance), le résultat du groupement tend à produire des groupes sans significations, ajoutant à cela le problème de choix du nombre de groupes qui conduit à de mauvaises solutions (figure 4), ces deux problèmes nous permettent d'envisagées des perspectives d'amélioration qui seront le sujet de nos futures études.

## Références


1. Boris Mirkin. Clustering for Data Mining: A Data Recovery Approach, Chapman & Hall/CRC, Taylor & Francis Group, Boca Raton, 2005.
2. Celeux G,.Diday E., Govaert G., Lechevallier Y., Ralam-Bondrainy H. Classification Automatique des Données. Bordas, Paris, 1989.
3. Daniel T. Larose. Discovering Knowledge in Data: An Introduction to Data Mining, John Wiley & Sons, Inc., Hoboken, New Jersey. 2005
4. Jacob Kogan, Introduction to Clustering Large and High-Dimensional Data, Cambridge University Press, Cambridge, 2007.
5. Likas A., Vlassis M. & Verbeek J., The global k-means clustering algorithm, Pattern Recognition, 36, pp. 451-461.,2003