LE DATA MINING

La fouille de données

Dr. Nedioui Med Abdelhamid
Université Echahid Hama Lakhder Eloued - Algérie
nedioui3904@gmail.com

Chapitre 9
Segmentation

1. Introduction

Le clustering (En l'appel aussi le partitionnement, categorisation, segmentation, regroupement…) est une classification non supervisée, visant a organiser un ensemble d'objets en groupes ou clusters, de façon a avoir des objets similaires en groupes et les objets différents organisés dans des groupes différents.

L'objectif de la segmentation est le suivant : on dispose de données non classée (étiquetées). On souhaite les regrouper par données ressemblantes.

2. Définition

Soit un ensemble X de N données décrites chacune par leurs P attributs.

La segmentation consiste a créer une partition ou une décomposition de cet ensemble en groupes telle que :

1. les données appartenant au même groupe se ressemblent ;

2. les données appartenant a deux groupes différents soient peu ressemblantes.

3. Les méthodes de clustring

La non connaissant préalable des critères de partitionnement a priori, nous à conduit a utiliser des algorithmes d'apprentissage non supervises : ils organisent les données sans qu'ils ne disposent d'information sur ce qu'ils devraient faire.

Il existe deux grandes classes de méthodes :

- Hiérarchique : on décompose l'ensemble d'individus en une arborescence de groups.
- Non hiérarchique : on décompose l'ensemble d'individus en **k** groupes ;

3. Les méthodes de clustring

3.1 Les méthodes hiérarchiques

Dans un clustering hiérarchique, un cluster peut être divise en sous clusters, l'ensemble des clusters étant généralement représente par un arbre. Un objet appartient a une et une seule feuille dans la hiérarchie, mais également a son nœud père, et ainsi de suite jusqu'a la racine.

Il existe deux types d'approches hiérarchique :

- Les approches par agglomération (ascendantes).
- Les approches par division (ou descendantes).

3. Les méthodes de clustring

3.1.1 Les approaches par agglomeration

Cette approche commence par des clusters formes d'un seul objet, puis les fusionne successivement jusqu'a ce que le critère d'arret soit atteint (Construction de K clusters)

Algorithme :

1. Initialement, mettre chaque objet dans son propre cluster ;
2. Parmi tous les clusters courants, sélectionner les deux clusters ayant la plus petite distance ;
3. Remplacer ces deux clusters par un nouveau cluster, formé par la fusion des deux clusters originaux ;
4. Répéter les étapes 2 et 3 jusqu'à atteindre la condition d'arrêt.

3. Les méthodes de clustring

3.1.1 Les approaches par agglomeration

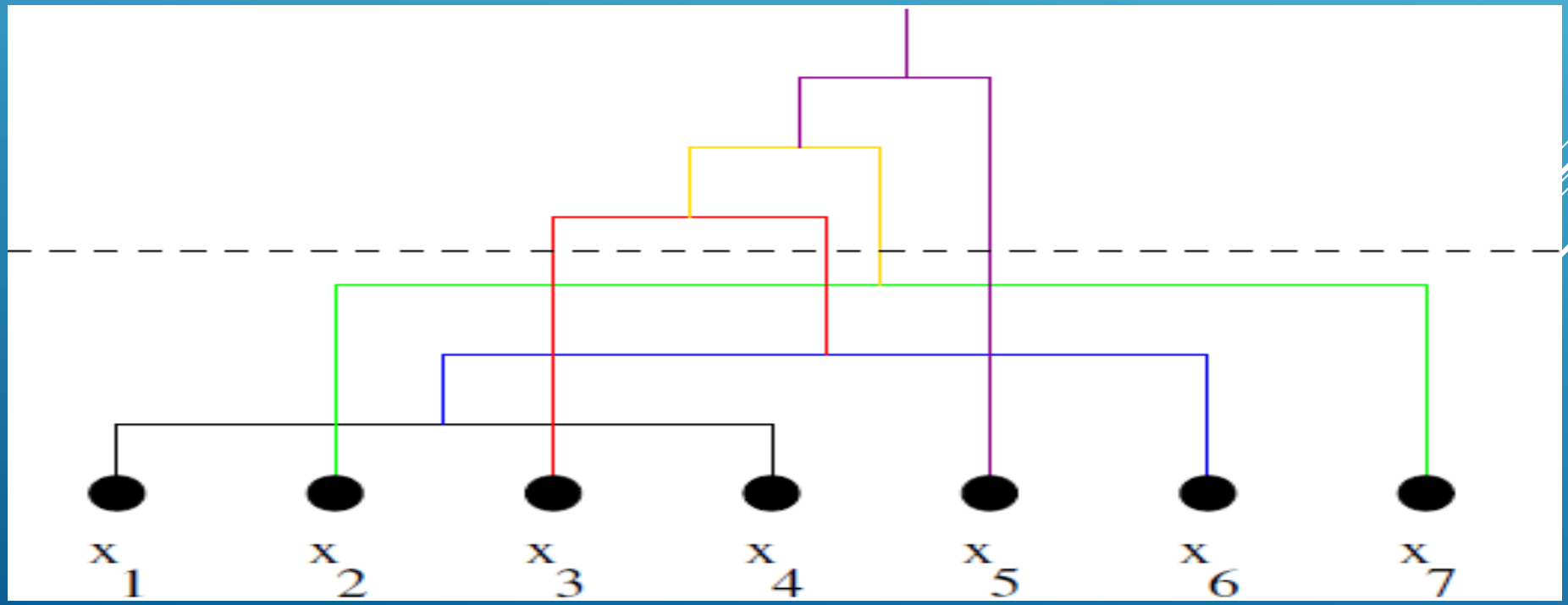
Le principe de la méthode de segmentation hiérarchique ascendante est le suivant :

On part de N groupes, que l'on réduit a N -1, puis a N -2,... On passe de k + 1 a k groupes en regroupant deux groupes. On obtient ainsi une hiérarchie (ou arbre de segmentation ou « **dendrogramme** ») dans lequel on peut facilement retrouver k groupes en le coupant a un certain niveau.

3. Les méthodes de clustring

3.1.1 Les approaches par agglomeration

Une fois obtenue cette hiérarchie, on peut obtenir une segmentation en 4 groupes en « découpant » l'arbre la ou il y a quatre branches verticales. Les 4 morceaux qui tombent constituent les 4 groupes.



3. Les méthodes de clustring

3.1.2 Les approaches par division

Cette approche commence par un cluster formé de tous les objets, qui sera ensuite divise en petits clusters jusqu'a atteindre une condition d'arret donnée par l'utilisateur.

Algorithme :

1. Mettre tous les objets dans un seul cluster ;
2. Répéter jusqu'à atteindre la condition d'arrêt :
   (*a*) Choisir un cluster à diviser ;
   (*b*) Remplacer le cluster choisi par le sous cluster obtenu.

3. Les méthodes de clustring

3.1.3 Les algorithmes

Il existe de nombreux algorithmes de type hiérarchique qui sont proposes dans la littérature, les plus connus sont : BIRCH (Balanced Iterative Redu-cing and Clustering using Hierarchies) pour le clustering de type agglomeratif et DIANA (Divisive Analysis) pour le clustering de type divisif. Il existe très peu d'algorithmes divisifs, notamment a cause de la difficulté a définir un critère de séparation d'un cluster. En effet pour un cluster de taille n, il y a $(2^{n-1}-1)$ possibilites pour diviser ce cluster en deux sous clusters. Dans le cas agglomeratif, chaque fusion de 2 clustre parmis n , offre selement n(n-1)/n possibilité

3. Les méthodes de clustring

3.1.4 Les avantages méthodes hiérarchiques

Avantages :

- Facilite de manipuler toutes formes de similitude ou de distance.
- Applicabilité a tout type d'attribut.

Inconvenients :

- Imprécision sur les critères d'arret.
- La plupart des algorithmes hiérarchique ne revisitent pas les clusters une fois construits en vue de l'amélioration des résultats.

3. Les méthodes de clustring

3.2 Les méthodes non hiérarchiques

Les méthodes de **hiérarchiques** ont généralement comme résultat un ensemble de K clusters, chaque objet appartenant a un seul cluster. Chaque cluster peut être représenté par un centroïde, qui correspond a la moyenne ou le centre de gravité de l'ensemble des objets contenus dans le cluster. La forme précise de cette description dépendra du type des objets qui sont groupes.

3. Les méthodes de clustring

3.2 Les méthodes non hiérarchiques

3.2.1 Algorithme du K-Means (Macqueen,1967)

L'algorithme du k-means est le plus populaire des algorithmes de clustering, il est utilise dans des applications aussi bien scientifiques que techniques.

Dans cette méthode, un cluster est représenté par son centroïde qui est une moyenne des points situes a l'intérieur du cluster, cette approche ne fonctionne convenablement qu'avec les attributs numérique et le résultat final peut être négativement affecté par la présence de bruits.

3. Les méthodes de clustring

3.2 Les méthodes non hiérarchiques

3.2.1 Algorithmedu K-Means (Macqueen,1967) :

La somme des écarts entre un point et son centroïde , exprimée avec une mesure appropriée, est utilisée comme fonction objectif. Chaque point est assigne au cluster dont le centroïde est le plus proche.

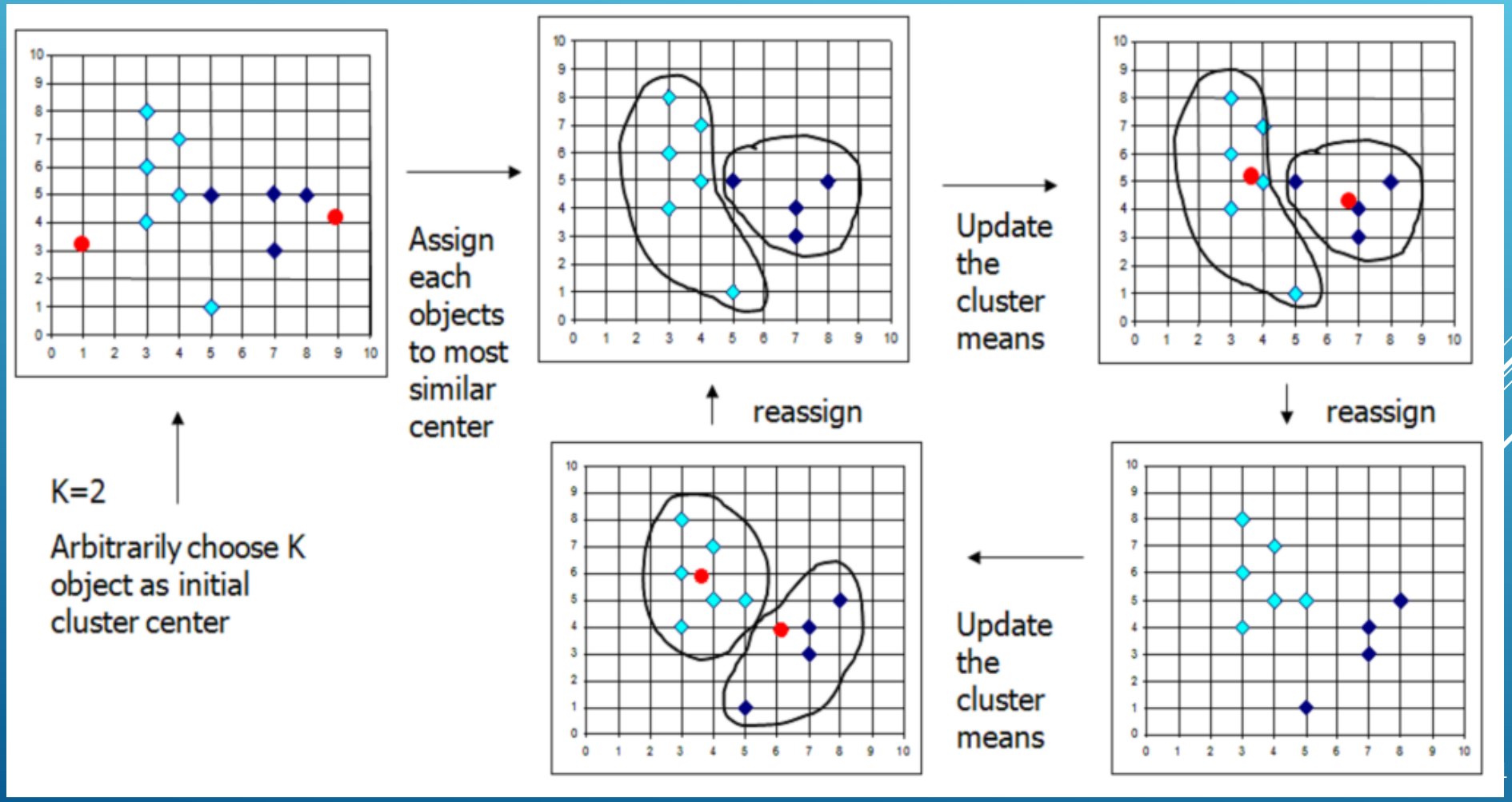
Algorithme : K-means

1. Sélectionner K points comme centroïdes initiaux ;
2. Former K clusters en assignant chaque point au centroïde le plus proche ;
3. Recalculer le centroïde de chaque cluster nouvellement formé ;
4. Répéter les étapes 2 et 3 jusqu'à ce qu'aucun centroïde ne change.

3. Les méthodes de clustring

3.2 Les méthodes non hiérarchiques

3.2.2 Exemple



3. Les méthodes de clustring

Remarques :

La segmentation obtenue dépend des centres initiaux

Lors de l'initialisation de l'algorithme, on prend K points dans l'espace de données au hasard. La segmentation calculée par les centres mobiles dépend de cette initialisation.

Pour contrer ce problème, on exécute plusieurs fois l'algorithme en prenant a chaque fois des centres initialises différemment. On compare les segmentations obtenues a chaque itération et on retient celle dont l'inertie intr-aclasse est la plus faible.

3. Les méthodes de clustring

Remarques :

Le nombre de groupes

Le nombre de groupes K choisi peut être mauvais. On peut tester plusieurs valeurs de K en exécutant plusieurs fois l'algorithme avec des K croissants.

3. Les méthodes de clustring

3.2 Les méthodes non hiérarchiques

3.2.2 Les algorithmes des k- médoïdes :

Les algorithmes des k-médoides sont différents de ceux des k-moyennes (ou k-means) par l'utilisation de médoides plutôt que des centroides pour représenter les clusters. Le principe de fonctionnement de ces méthodes ressemble a celui des k-moyennes sauf que, contrairement aux algorithmes des k-means ou le cluster est repesent par une valeur moyenne, un cluster dans l'algorithme des k-médoides est représente par un de ses objets prédominants appel le médoide.

3. Les méthodes de clustring

3.2 Les méthodes non hiérarchiques

3.2.2 Les algorithmes des k- médoïdes :

L'algorithme représentatif de cette famille est l'algorithme PAM (Partioning Around Medoids) qui a et introduit initialement par Kauf-man et Roussew . L'avantage de cet algorithme par rapport a celui des k-means est qu'il est moins sensible aux points bruits (points isoles). Il peut aussi être applique a n'importe quel type de variables (numérique ou catégorielle).

3. Les méthodes de clustring

3.2 Les méthodes non hiérarchiques

3.3 Avantages et inconvénients

Les algorithmes bases sur le principe de partitionnement des données présentent l'avantage d'être facile a implémenter en plus de leur rapidité et leur faible exigence en taille mémoire. Ainsi, ils sont applicables sur des bases de données volumineuses en choisissant une fonction de distance.

3. Les méthodes de clustring

3.2 Les méthodes non hiérarchiques

3.3 Avantages et inconvénients

Ces algorithmes présentent des inconvénients, dont nous citons quelques uns :

- Les conditions d'arrêt de l'exécution de l'algorithme doivent être fixées au préalable par l'utilisateur. En e et, ces algorithmes peuvent s'exécuter jusqu'a atteindre par exemple un nombre d'itérations prédéfini ou atteindre une stabilité des données (cas de l'algorithme K-means et l'algorithme K-medoids).

3. Les méthodes de clustring

3.2 Les méthodes non hiérarchiques

3.3 Les Avantages et inconvénients

- Ces algorithmes sont influences par les paramètres choisis aléatoirement dans la partition initiale, tel que le nombre de clusters k qui est un paramètre indispensable.
- Les clusters résultants d'une approche floue contiennent plus d'informations que dans le cas du clustering dur. Cependant, la visualisation et l'interprétations des clusters qui sont beaucoup plus difficiles.