Chapitre 3 : Clustering (classification automatique)

1. Introduction

- ***Définition* :** technique visant à grouper des objets dans des classes a priori inconnues de sorte à ce que les objets d'une même classe soit similaires et les classes ne soit pas similaires.
- ***Objectifs* :** (1) Abstraire les objets en classes (conceptualisation) ayant les mêmes caractéristiques. (2) utiliser les classes obtenues à des fins diverses : (i) marketing (ciblage de groupes clients, profiling), (ii) navigation Web (personnalisation du contenu par classe d'utilisateurs), (iii) Détection de communautés, etc.

2. Recherche de clusters

- On dispose d'un ensemble d'objets (individus, voitures, etc.) qu'on veut regrouper.
- Les objets ne sont pas classés initialement et on ne dispose pas de classes prédéfinies.
- Tous les objets sont décrits par un ensemble de caractéristiques (variables).
- Les variables sont de divers types mais ne sont pas toutes adéquates au clustering.
- Chaque objet peut être considéré comme un vecteur dans un espace multidimensionnel.
- Les objets similaires peuvent être proches géométriquement.

3. Algorithme K-means (K-moyennes)

Il existe beaucoup d'algorithmes de classification automatique, dont le plus ancien est celui des K-means (k-moyennes) avec beaucoup de variations.

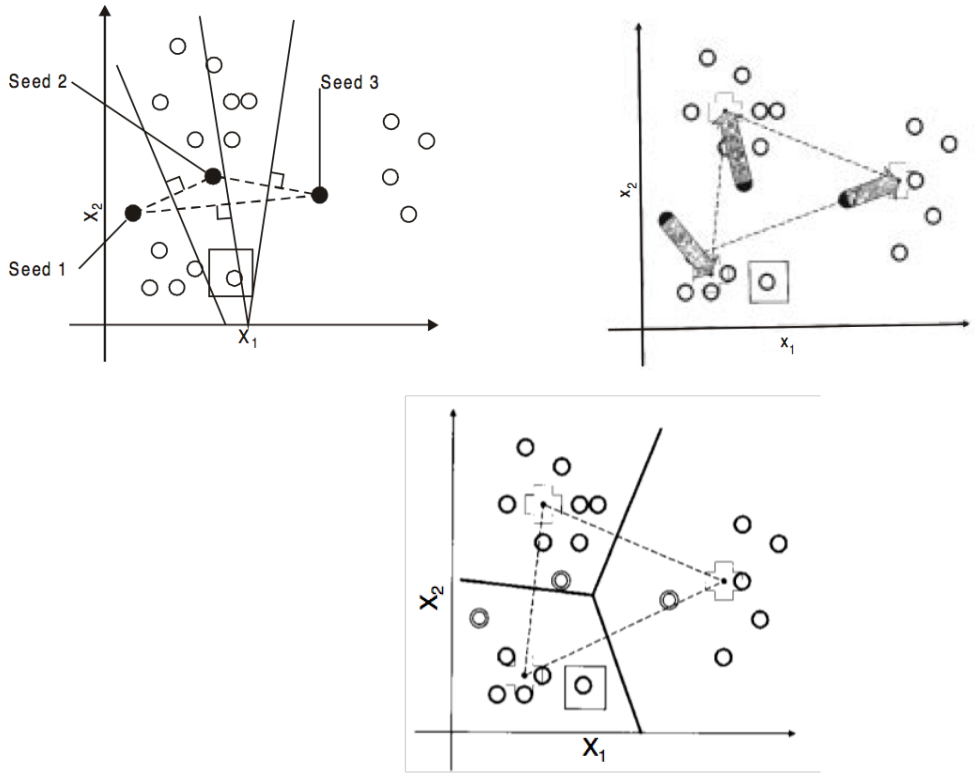
a. ***Principe de base :*** le principe de base de K-means est regrouper les objets en se basant à leur distance par rapport à un centre (centroid). Initialement, les centres sont choisis aléatoirement, mais ils seront mis à jour par re-calcul. Les cluster initiaux seront affinés, voire changés. La classification finale correspondra à une stabilisation des clusters : aucun objet ne change d'appartenance à son cluster.

b. ***Description de l'algorithme :*** Soit les N objets à classifier (pour les besoins d'analogie géométrique, on désigne les objets comme étant des points)

- Choisir K centres aléatoires (seeds).
- Calculer la distance entre chacun des points restants et chaque seed.
- Calculer le centre de chaque cluster. La valeur de chaque caractéristique de chaque centre est égale à la moyenne des valeurs de la caractéristique pour tous les points du cluster correspondant.
- Réaffecter les points selon leur distance par rapport au centre de chaque cluster : de nouveaux clusters peuvent être générés.
- Recalculer le centre C de chaque cluster
- S'arrêter lorsqu'aucun point ne change de cluster (stabilisation des clusters).

Illustration : On suppose que les objets sont décrits par deux caractéristiques X1 et X2. Les trois figures suivantes (de Nagabhushana) montrent les premières itérations de k-means. La figure à

gauche montre le choix des seeds initiaux (cercles noirs) et la formation des premiers clusters (trois. La figure à droite montre les centroids (symbole +). La troisième figure montre les nouveaux clusters. Le point entouré d'un carré est le point ayant changé de cluster après le calcul des centres.



4. ***Exemple illustratif*** (ref. Kardi) : soit les quatre objets décrits par deux caractéristiques avec les valeurs suivantes : $X_1(1, 1)$, $X_2(2, 1)$, $X_3(4, 3)$, $X_4(5, 4)$ (voir figure 1). On choisit les X_1 et X_2 comme seeds (étoiles) et on les désigne par C_1 et C_2. On calcule la distance entre chaque point et chaque centroid. On utilise la distance euclidienne (voir plus loin). On obtient la matrice suivante :

X_1	X_2	X_3	X_4	
0	1	3,61	5	C_1
1	0	2,83	4,24	C_2

On remarque que les points X_2, X_3 et X_4 sont proches à C_2 et X_1 à C_1. On obtient deux clusters composés de (X_1) et (X_2, X_3, X_4). Le calcul des nouveaux centres donne $C'_1(1,1)$ et $C'_2(11/3, 8/3)$. On recalcule les nouvelles distances. On obtient le tableau suivant.

X_1	X_2	X_3	X_4	
0	1	3,61	5	C'_1
3,14	2,36	0,47	1,89	C'_2

Selon cette matrice, on déplace X_2 au cluster à gauche et on obtient les clusters de la figure 3. On recalcule les centres (étoiles de la figure 3). On trouve les centres $C''_1(1.5, 1)$ et $C''_2(4.5, 3.5)$. On calcule les distances et on obtient les distances suivantes

X_1	X_2	X_3	X_4	
0,5	0,5	3,20	4,61	C''_1
4,30	3,54	0,71	0,71	C''_2

La matrice montre qu'aucun point ne change de cluster. On arrête les itérations.

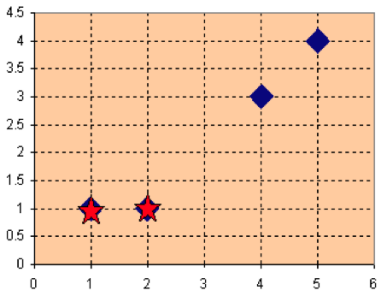


Figure 1

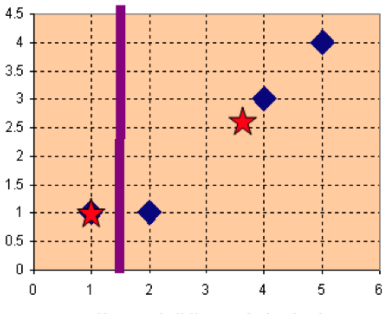


Figure 2

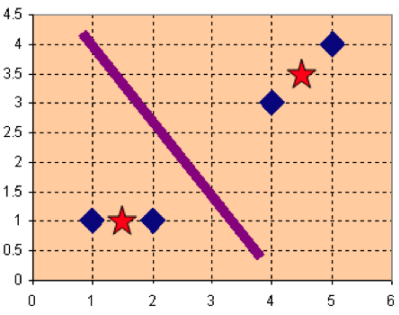


Figure 3

5. **Types de variables pour le clustering :** les types de données pour le clustering ne sont pas tous adéquats. On les classe par ordre croissant d'adéquation comme suit :

   - Variables symboliques (couleurs, noms, etc.)
   - Rangs (1, 2, 3, etc.)
   - Les intervalles (degrés de températures, etc.)
   - Les vraies mesures (âge, longueurs, volumes, etc.)

Les intervalles et les vraies mesures sont les plus adéquats au clustering. Les variables symboliques et les rangs doivent être transformées pour être utilisées. Cependant, les transformations ne donnent pas toujours des intervalles ou mesures ayant un sens et comparables.

6. **Mesures de similarité :** pour trouver les objets associés, on utilise différentes mesures de similarité.
   a. *Distance* : le calcul de distance dépend de la nature des données

- Intervalles et vraies mesures : tout d'abord, nous standardisons les données. On calcule l'écart absolu moyen par la formule $s_f = \frac{1}{n}(|x_{1f} - m_f| + |x_{2f} - m_f| + ... + |x_{nf} - m_f|)$ où m représente la moyenne des valeurs. Ensuite, on divise l'écart entre chaque mesure et sa

$$z_{if} = \frac{x_{if} - m_f}{s_f}$$

moyenne par l'écart-type moyen.

Une fois les mesures standardisées, nous calculons les distances. Les formules sont

- Formule de Minkowski $d(i,j)=\sqrt[q]{(|x_{i_1}-x_{j_1}|^q+|x_{i_2}-x_{j_2}|^q+...+|x_{i_p}-x_{j_p}|^q)}$
- q = 1 : distance de Manhattan :

$$d(i,j)=|x_{i_1}-x_{j_1}|+|x_{i_2}-x_{j_2}|+...+|x_{i_p}-x_{j_p}|$$

- q = 2 : distance euclidienne : $d(i,j)=\sqrt{(|x_{i_1}-x_{j_1}|^2+|x_{i_2}-x_{j_2}|^2+...+|x_{i_p}-x_{j_p}|^2)}$

- Variables binaires : on élabore une table de contingence (voir exemple en cours) pour calculer quatre valeurs, désignées par a, b, c et d. « a » désigne le nombre de cas où l'objet i possède un 1 et l'objet i possède un 1. Pour b, c et d, on cherche les respectivement le nombre des cas où on a (1, 0), (0, 1), (0, 0).
  - Une fois les valeurs de a, b, c et d trouvées, on aura deux mesures :
    - Coefficient d'appariement simple :

$$d(i,j)=\frac{b+c}{a+b+c+d}$$

    - Coefficient de Jaccard :

$$d(i,j)=\frac{b+c}{a+b+c}$$

Remarque : lors du codage des variables booléennes en binaire, l'assignation des valeurs 1 et 0 dépendra de la symétrie des objets.

- Si les objets sont symétriques (pas de valeur booléenne fréquente), on code au hasard les deux valeurs booléennes
- Si les objets sont asymétriques (ex : test de virus), on code par 1 la valeur la moins fréquente.

La distance sera calculée en utilisant les variables asymétriques.

b. *Angle entre les vecteurs :* ce calcul ne tient pas compte des grandeurs des mesures.
c. *Nombre de caractéristiques communes* : on calcul le nombre d'appariements

$$d(i,j) = \frac{p-m}{p}$$

où *m* est égal au nombre de caractéristiques communes et *p* est le nombre total de caractéristiques.

7. **Choix du nombre K** : il n'existe pas de meilleurs K ou de mauvais K pour le clustering, puisque ce nombre est fourni par l'utilisateur pour spécifier le nombre de clusters souhaités. Cependant, les clusters obtenus peuvent être évalués. Les bons clusters seront le résultat d'un bon choix de *k* et inversement.