MA431 : Mathématiques appliquées à la sécurité

Clustering : kmeans et DBSCAN

D. Barcelo

Grenoble INP ESISAR

2022/2023

1. Méthodes des k-means
2. Clustering par densité : DBSCAN
3. Comparaison des méthodes sur un exemple

Méthode de partitionnement

la méthode des k-means ou k-moyennes

Méthodes des k-means

Méthode des k-means

- Objectif : obtenir une partition.
- On fixe par avance le nombre de clusters attendus : k.
- Principe de l'algorithme : minimiser l'inertie intraclasse.
- Résolution exacte impossible : algorithme de Lloyd.

On considère un ensemble E de n individus caractérisés par p variables.
On suppose l'espace $\mathbb{R}^p$ muni d'une distance d appropriée.
On désire constituer k classes.

Algorithme

Etape 0 :
On détermine k centres aléatoires de classes : $\left\{g_1^0, \ldots, g_k^0\right\}$.
Les k centres déterminent une partition $C^0 = \left\{C_1^0, \ldots, C_k^0\right\}$ de E.
Un individu i appartient à C_j^0 s'il est plus proche de g_j^0 (au sens de la distance d) que de tous les autres centres.

Algorithme

Etape 1 :
On détermine k nouveaux centres de classes : $\left\{g_1^1, \ldots, g_k^1\right\}$.
g_j^1 est le centre de gravité de C_j^0 obtenu avec la partition C^0.
Les k nouveaux centres déterminent une partition $C^1 = \left\{C_1^1, \ldots, C_k^1\right\}$ de E. Un individu i appartient à C_j^1 s'il est plus proche de g_j^1 (au sens de la distance d) que de tous les autres centres.

Algorithme

Etape q :
On détermine k nouveaux centres de classes : $\left\{g_1^q, \ldots, g_k^q\right\}$.
g_j^q est le centre de gravité de C_j^{q-1} obtenu avec la partition C^{q-1}.
Les k nouveaux centres déterminent une partition $C^q = \left\{C_1^q, \ldots, C_k^q\right\}$ de E. Un individu i appartient à C_j^q s'il est plus proche de g_j^q (au sens de la distance d) que de tous les autres centres.

Algorithme

L'algorithme s'arrête :

- soit lorsque deux itérations successives conduisent à la même partition,
- soit lorsque l'inertie intraclasse est suffisamment faible,
- soit lorsqu'on a atteint un nombre d'itérations préalablement fixé.

Illustration

Lien vers une animation très bien faite pour comprendre la méthode des k-means

Ou

La même sur Chamilo

Exemple sous R

Exemple sous R

Avantages de la méthode

- Complexité algorithmique en $\mathcal{O}(np)$ donc linéaire.
- Utilisable pour un grand jeu de données.
- L'algorithme converge quelque soit le choix des centres initiaux.
- Sensibles aux données aberrantes qui forment des clusters isolés (*détection d'intrusion*).

Inconvénients

La solution obtenue est-elle optimale ?

Inconvénients

La solution obtenue est-elle optimale ?
Surement pas ! Mais pourquoi ?

Inconvénients

La solution obtenue est-elle optimale ?
Surement pas ! Mais pourquoi ?

- la détermination des premiers centres de gravité n'est pas optimale,
- forte sensibilité aux conditions initiales,
- pour trouver un k optimal, il faut en tester plusieurs et considérer l'apport d'informations supplémentaires en augmentant k (comment augmente la proportion d'inertie expliquée par la partition).

Exemple

Retour sur l'exemple avec R

Variantes

Il existe plusieurs variantes des k-means :

- k-means++ : pour éviter les plus mauvaises solutions, on utilise un algorithme stochastique qui disperse au maximum les k premiers centroïdes (en les éloignant).
- k-means|| : on ajoute plusieurs centres à la fois (un nombre aléatoire), puis au bout de k itérations on clusterise l'ensemble des centres en k clusters (avec un k-means).
- k-means avec noyau : afin de former des clusters non convexes, on utilise l'astuce du noyau.

Méthode mixte

Si le nombre d'individus est trop important on combine des méthodes hiérarchiques et non hiérarchiques.

1. On forme k classes par les k means.
2. On construit un arbre à l'aide de la partition précédente.
3. On coupe à un nombre de classes approprié.
4. On effectue une nouvelle partition par les k means.

Clustering par densité

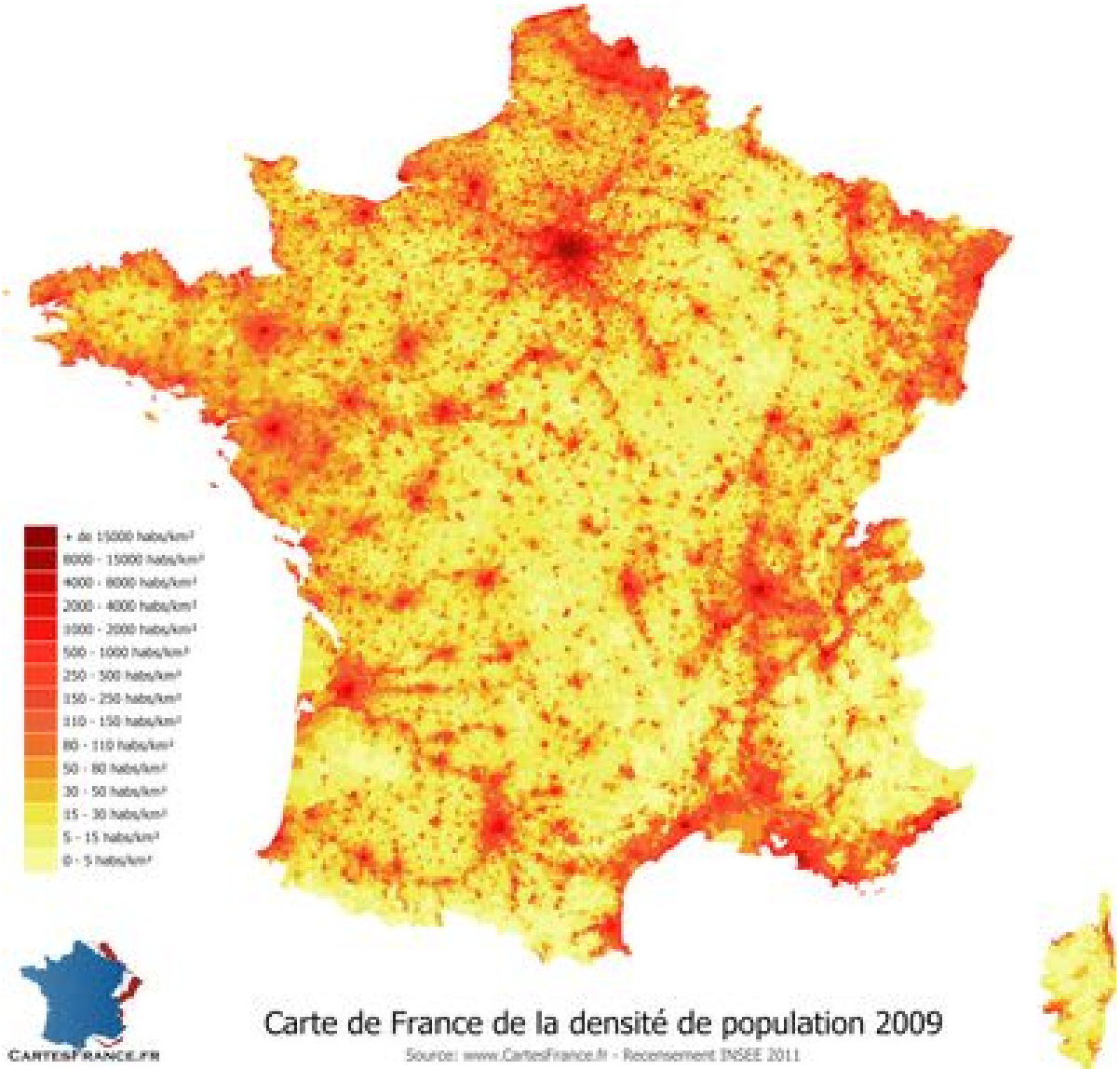
Density-Based Spatial Clustering of Applications with Noise : DBSCAN

Clustering par densité

Clustering par densité

- Objectif : détecter des clusters de formes irrégulières, non nécessairement convexes, et de tailles et variances inégales.
- Nombre de clusters attendues non fixé.
- Le cluster croît dans la direction où la densité est suffisante.
- On fixe deux paramètres de lissage : le nombre de voisins et le rayon de la sphère de voisinage.

Exemple


+ de 15000 habs/km²
8000 - 15000 habs/km²
4000 - 8000 habs/km²
2000 - 4000 habs/km²
1000 - 2000 habs/km²
500 - 1000 habs/km²
250 - 500 habs/km²
150 - 250 habs/km²
110 - 150 habs/km²
80 - 110 habs/km²
50 - 80 habs/km²
30 - 50 habs/km²
15 - 30 habs/km²
5 - 15 habs/km²
0 - 5 habs/km²
CARTESFRANCE.FR
Carte de France de la densité de population 2009
Source: www.CartesFrance.fr - Recensement INSEE 2011

DBSCAN

Paramètres de DBSCAN :

- On note MinPts le nombre minimal de points que doit contenir un cluster.
- On note ε le rayon d'un voisinage autour d'un point.
  Soit x un point alors $V_\varepsilon(x) = \{y / d(x, y) < \varepsilon\}$

DBSCAN

Définition :

- x est un point intérieur (core point) si son voisinage contient au moins MinPts éléments.
- x est un point frontière s'il n'est pas un point intérieur mais qu'il est dans le voisinage d'un point intérieur.
- x est un outlier (du bruit) sinon.

DBSCAN

Définition :

Un point y peut être :

- directement densité accessible s'il est dans le voisinage d'un point intérieur x.
- densité accessible à un point intérieur x s'ils sont reliés par une chaine de points directement densité accessibles.
- densité connecté à un point z si y et z sont densité accessibles à partir du même point intérieur x.

Algorithme

Etape O :
On fixe les paramètres MinPts et ε.

Algorithme

Etape 1 :

- On choisit aléatoirement un point x.
- On détermine le voisinage du point.
- S'il ne contient pas MinPts alors il est considéré comme un point frontière ou comme du bruit et on cherche un autre point.
- S'il contient MinPts points alors c'est le début d'un cluster C.

Algorithme

Etape 2 :

- On associe à C tous les points densités accessibles à x.
- On continue avec tous les points du cluster ainsi formé.
- à la fin de l'étape, tous les points du cluster sont densité connectés.

Algorithme

Etape 3 :

- On cherche un point extérieur à la classe C.
- On recommence l'étape 1.

Algorithme

L'algorithme s'arrête lorsque tous les points ont été affectés à un cluster ou bien identifiés comme un point frontière ou du bruit.

Illustration

Lien vers une animation très bien faite pour comprendre la méthode DBSCAN

Choix des paramètres

Choix de MinPts :

- MinPts=1 : tous les points sont des points noyaux.
- MinPts=2 : équivaut à CAH de saut minimum.
- MinPts$\geqslant$ 3 : préférable.
- Souvent MinPts $\geqslant p+1$ ou MinPts $\geqslant 2p$.
- En pratique MinPts prend souvent les valeurs 5, 10 et 15.

Choix des paramètres

Choix de ε :

- plus compliqué.
- En pratique : calcul des distances de chaque point à son MinPts plus proche voisin.
- Représentation graphique des distances par ordre croissant.
- Recherche d'un coude dans la représentation graphique.
- Au dessus de l'optimum : beaucoup de points sont noyaux donc clusters trop importants et moins nombreux.
- En dessous de l'optimum : beaucoup de points sont des point frontières ou du bruit.

Exemple

Exemple sous R

Avantages de la méthode

- Pas de choix à priori du nombre de clusters.
- Détecte des formes spéciales de clusters.
- Complexité algorithmique en $\mathcal{O}(n\ln(n))$.
- Utilisable pour un grand jeu de données.
- bruit non gênant, détecte les points hors-norme.

Inconvénients

- Paramètres pas si simples à fixer.
- L'affectation des points frontières à une classe est liée à l'ordre dans lequel les points sont étudiés.
- Suppose que les densités des clusters sont égales. Problématique en grande dimension (fléau de la dimension et points très éloignés)

Clustering par densité

Comparaison des méthodes sur un exemple

Exemple

Exemple sous R

Comparaison

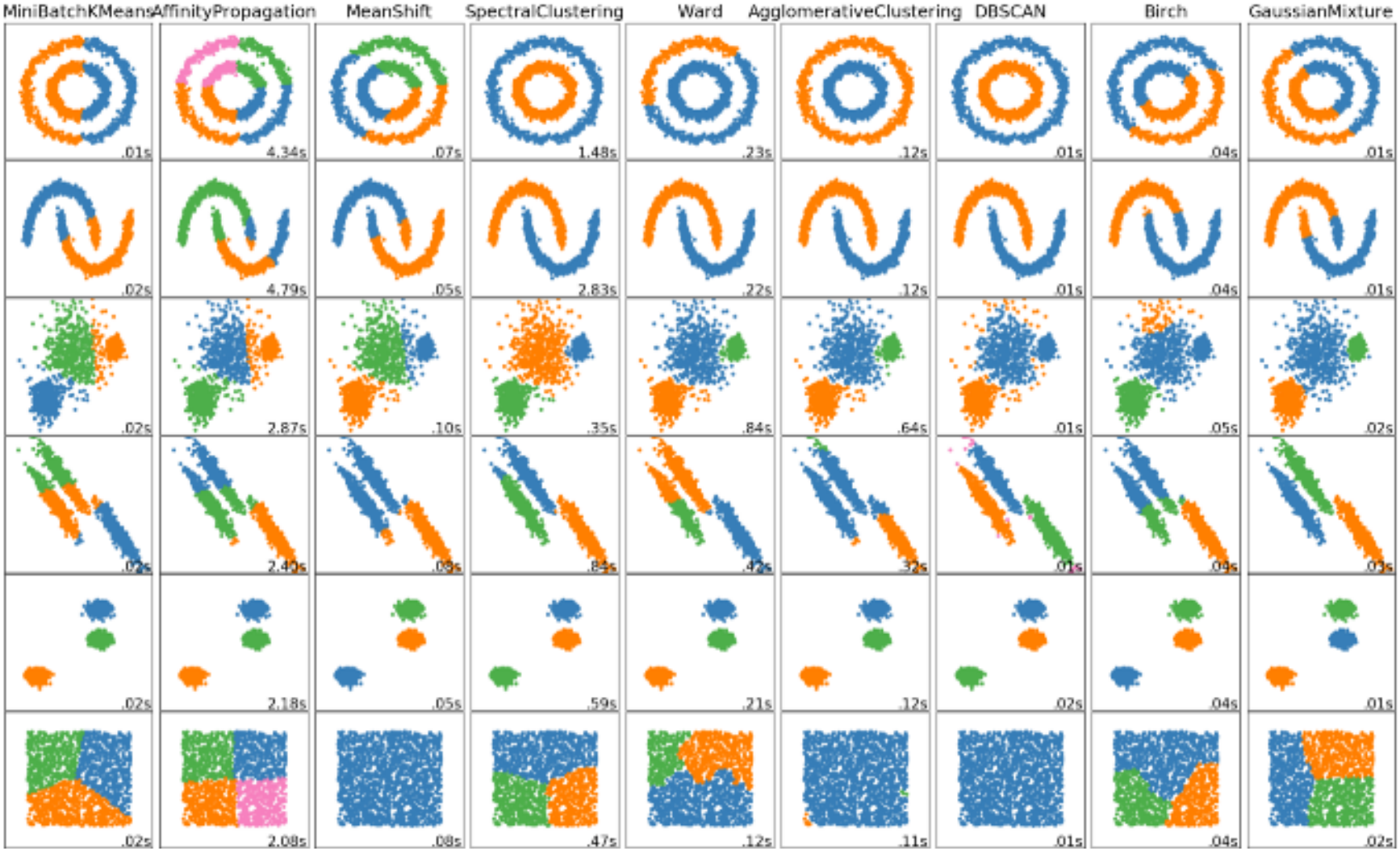


Image : https ://towardsdatascience.com/clustering-based-unsupervised-learning-8d705298ae51