# Clustering

CHEN Jing

Polytech’Tours

# Définition

- Objectif : il décrit des méthodes statistiques de classification de données. Il divise un ensembles de données en différents classes homogènes
- Clustering : est l'apprentissage non supervisé. C'est une technique d'analyse exploratoire des données servant à résumer les informations sur les données ou à déterminer des liens entre les points.

# Demande de Clustering

- Télescopable: certain algo. marche bien sur le petit témoin, mais il va avoir des erreurs sur le grand témoin.
- Type de données:  certain algo. est désgné vers certain type de données. Comme binary, nominal, ordinal ect.
- Forme de clustering: certains algo. mesure de la Manhattan de distance ou la Euclidean distance
- Paramètre : Pendant le traitement, certain algo. faut configurer des paramètre manuellement, comme le nombre de cluster

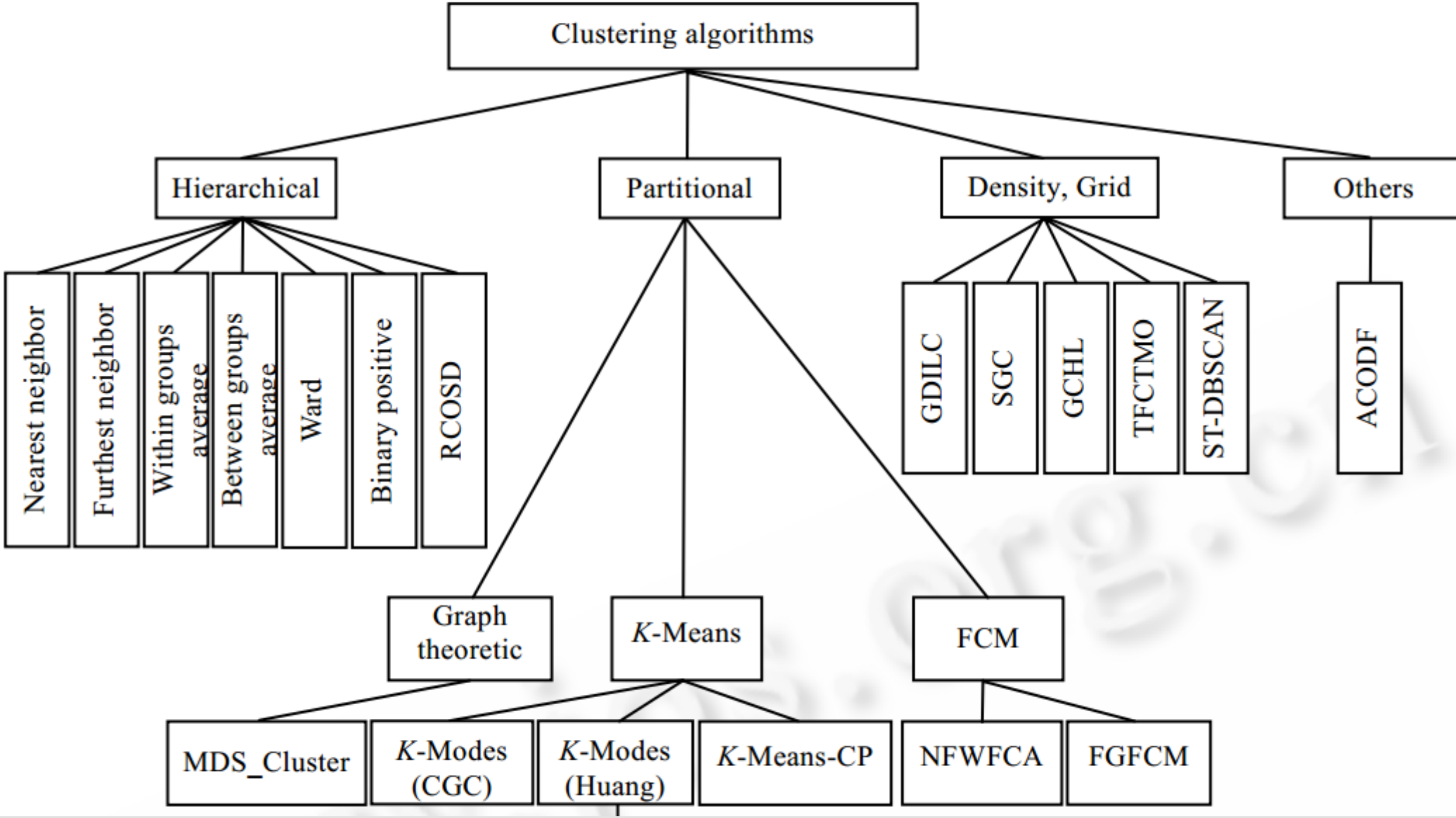
# Demande de Clustering

- Traitement de noise : la plupart des données bruites contient des points isolés ou des erreurs données
- Séquence d'enregistrement : certain algo est sensible à la séquence d'enregistrement.
- Haut dimensions : des données contiennent plusieurs propriétés/dimensions.
- Explication sémantique: client souhaite que le résultat de clustering est explicable.

# Usage et Catégorisation

- Usage : par exemple, il peut sert à analyser les différents groupes de client. En le même temps, il peut résumer la propension à consommer de chaque groupe.
- Étant un module d'analyse données, clustering peut être un outil indépendant qui sert à découvrir des informations de distribution de base de données et faire un bilan de caractéristique de chaque cluster. Il peut aussi analyser un cluster principalement.
- Catégorisation: maintenant il existe principalement des méthodes:

Méthode partitionnée, méthode hiérarchique, méthode en base de densité, méthode en base de gril, méthode en base de modèle


Clustering algorithms
Hierarchical
Partitional
Density, Grid
Others
Nearest neighbor
Furthest neighbor
Within groups average
Between groups average
Ward
Binary positive
RCOSD
GDILC
SGC
GCHL
TFCTMO
ST-DBSCAN
ACODF
Graph theoretic
K-Means
FCM
MDS_Cluster
K-Modes (CGC)
K-Modes (Huang)
K-Means-CP
NFWFCA
FGFCM

# Méthode de regroupement hiérarchique

- Ascendante hiérarchique: elle part d'une situation où tous les individus sont seuls dans une classe (CAH). Chaque cluster est progressivement absorbé par le cluster le plus proche jusqu'à la fusion des 2 derniers clusters.
- Descendante hiérarchique: Pour réaliser la subdivision, il faut souvent faire une classification hierarchique ascendante pour savoir quelle est la meilleure façon de séparer les points.

- Algo : Birch, Cure, Chameleon

# Méthode de partitionnement de données

Objective :

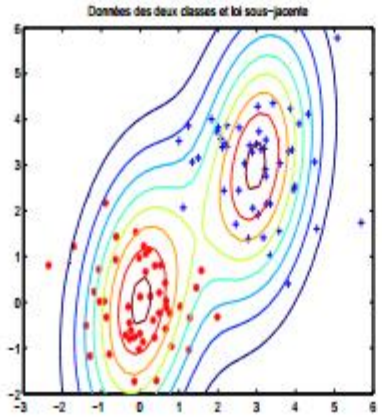
- Minimiser l'inertie intra-classe pour  obtenir des cluster les plus homogènes possibles
- Maximiser l'inertie inter-classe afin d'obtenir des sous-ensembles bien différenciés
- Recherche d'une partition en K ( K < N données) clusters . Construire toutes les partitions possibles et évaluer la qualité de chaque clustering et retenir la meilleure partition.
- Algo: k-means, k-medoids, clarans

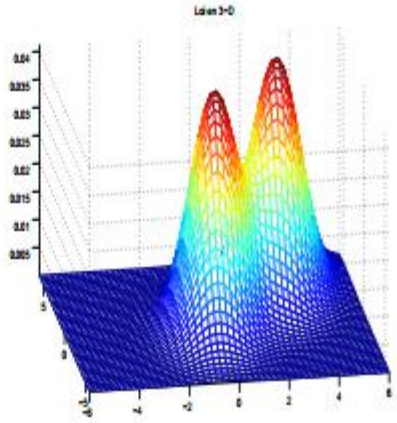
# Méthode de modélisation

- Données objectives suivent une modèle de distribution statique

Modèle de mélange gaussien

$$f(X) = \pi_1 \mathcal{N}(X; \mu_1, \Sigma_1) + \pi_2 \mathcal{N}(X; \mu_2, \Sigma_2) \quad \text{avec} \quad \pi_1 + \pi_2 = 1$$



- Algo: notion de modèles de mélange, algo. EM et variantes