Clustering Basé sur la Densité et par Décomposition en Grilles

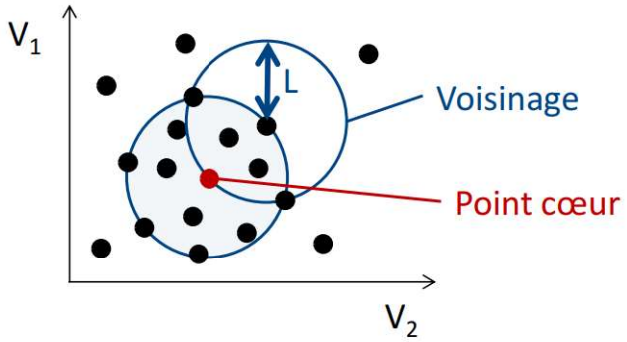
Nicolas PASQUIER
Laboratoire I3S (UMR-7271 UCA/CNRS)
Département Informatique
Université Côte d'Azur
http://www.i3s.unice.fr/~pasquier

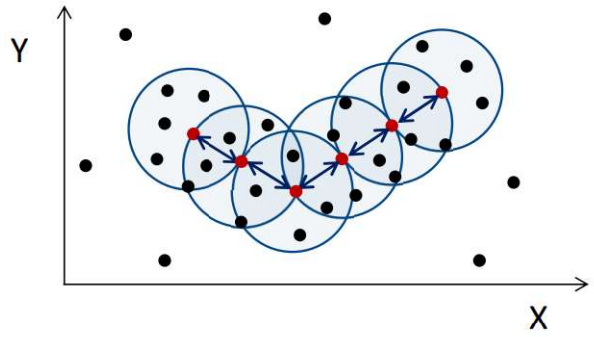
Clustering Basé sur la Densité

- Principe : identifier les régions denses de l'espace de données séparées par des régions faiblement peuplées (non denses)
- Chaque région dense de l'espace des données formera un cluster
- La densité des différentes régions de l'espace de données est évaluée en considérant le nombre de points qui sont « voisins » dans cette région
- Un point cœur dans l'espace de données est un point dont la taille du voisinage est au moins égale à un seuil
- Paramètres de l'approche définis a priori par l'utilisateur
  - Distance de voisinage : valeur seuil maximale pour la mesure de distance évaluant la similarité entre deux instances
  - Nombre de voisins : valeur seuil minimale pour le nombre d'instances dans le voisinage d'une instance pour qu'elle soit un point cœur

Clustering Basé sur la Densité

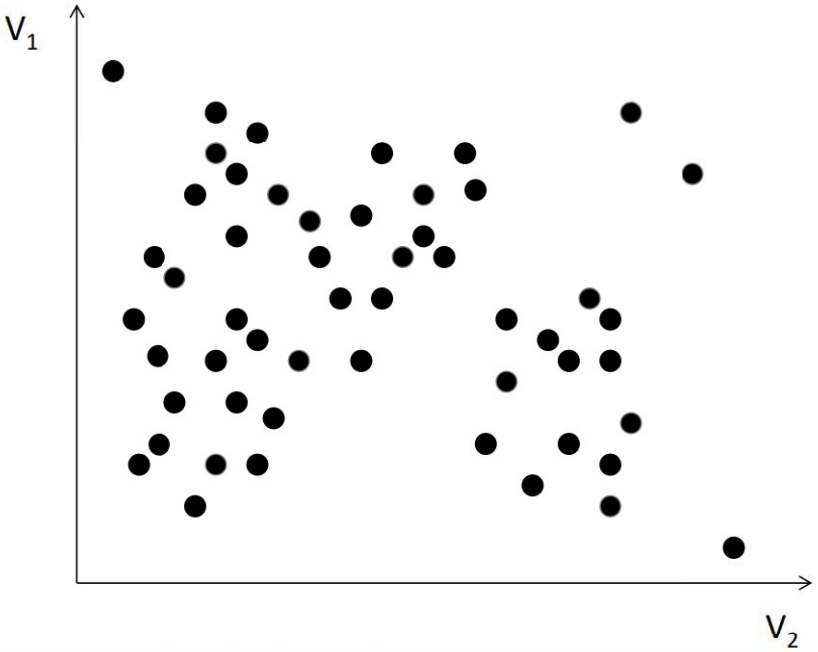
- Exemple : espace des données bidimensionnel
  - Distance de voisinage dans l'espace de données L = 0.5
  - Taille minimale du voisinage d'un point cœur N = 12
- Principe algorithmique
  1. Sélectionner aléatoirement des points dont le voisinage sera calculé
  2. Identifier les points cœurs parmi ceux-ci
  3. Fusionner les points cœurs mutuellement atteignables (voisinages communs)





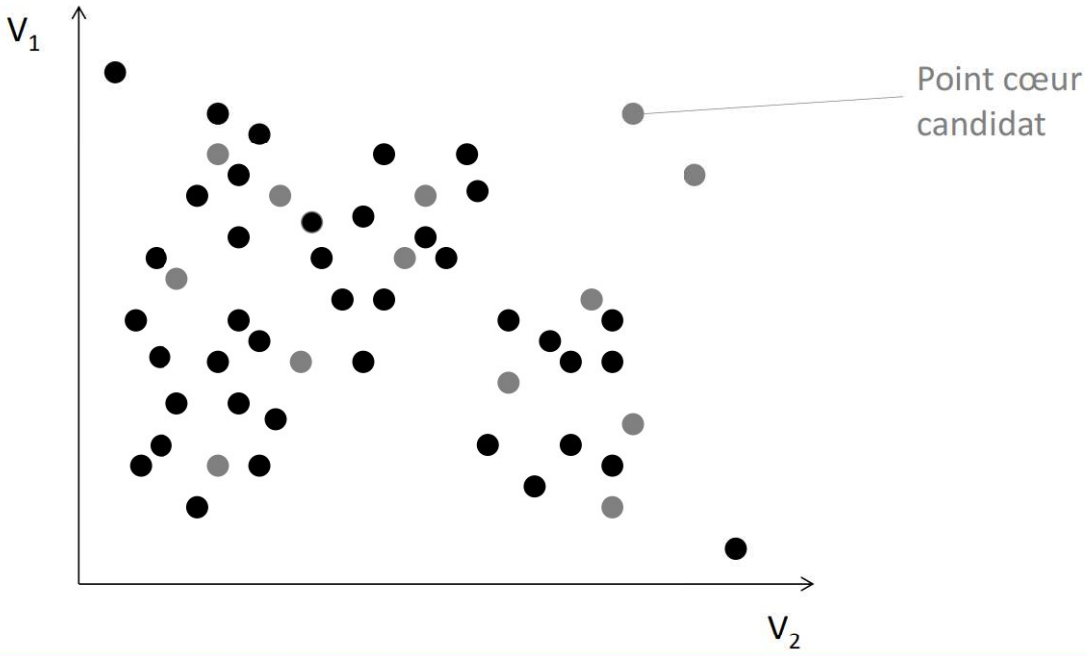
Clustering Basé sur la Densité

- Exemple d'espace des données bidimensionnel
- Paramètre N = 3



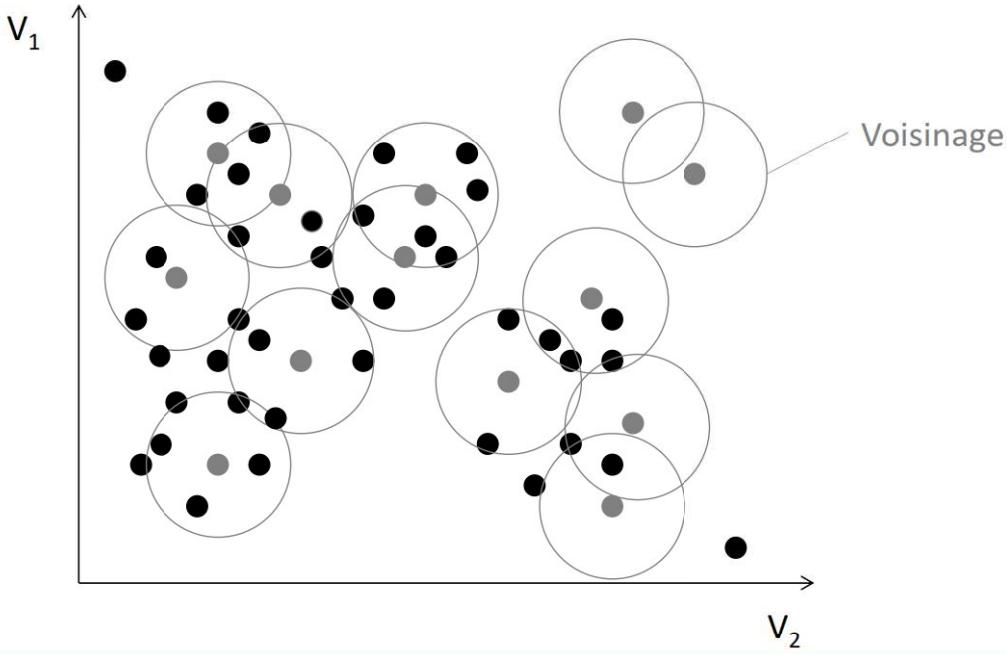
Clustering Basé sur la Densité

- Initialisation : sélection aléatoire des instances qui constitueront les points cœurs candidats



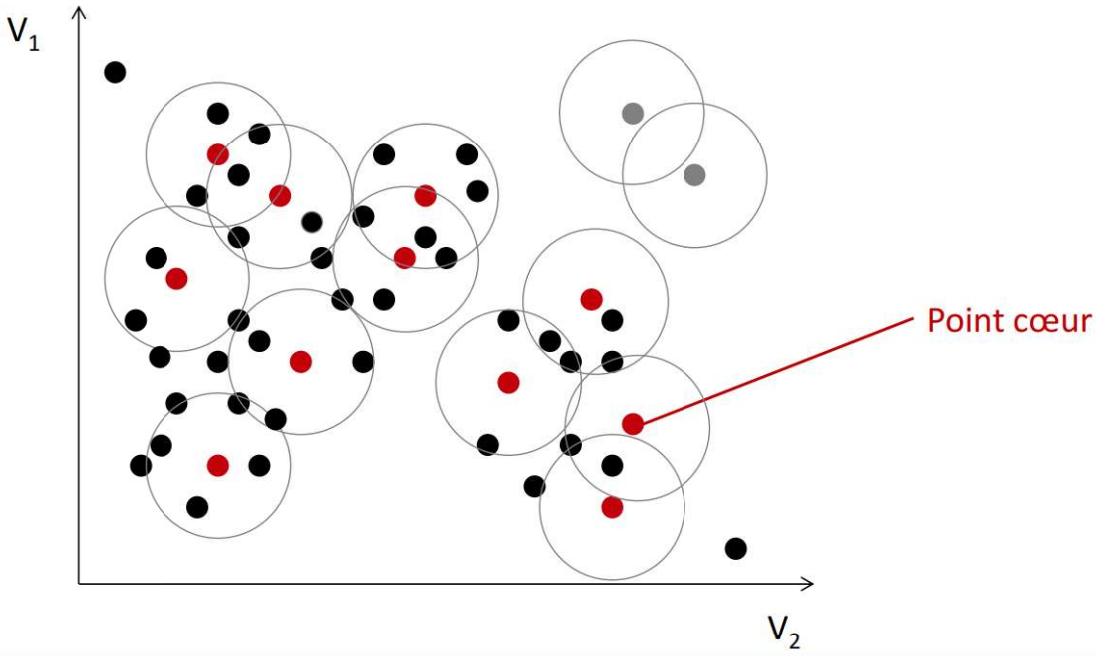
Clustering Basé sur la Densité

- Calcul du voisinage des instances points cœurs candidats



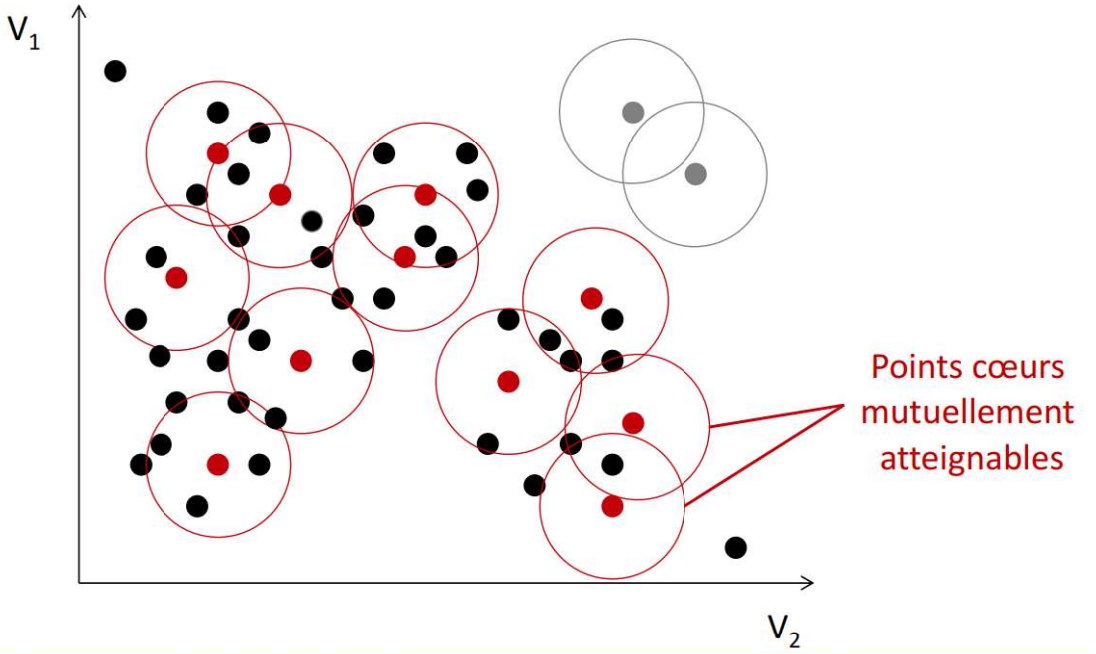
Clustering Basé sur la Densité

- Identification des instances points cœurs parmi les candidats



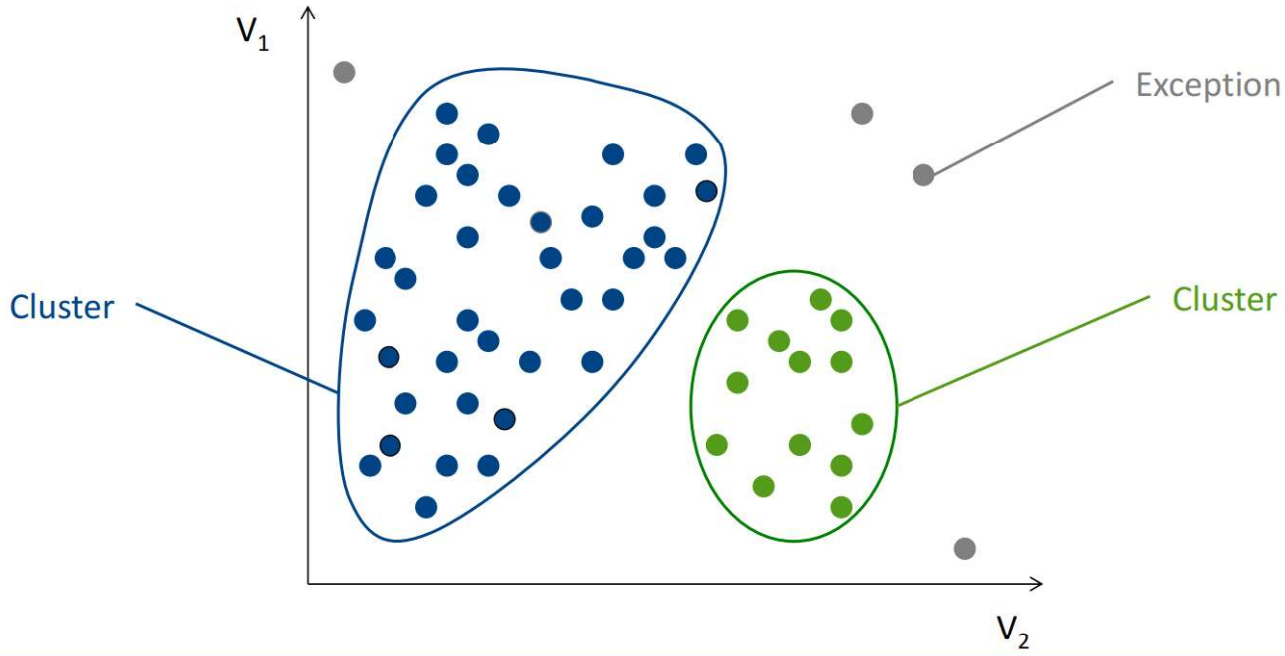
Clustering Basé sur la Densité

- Fusion des clusters des points cœurs qui sont mutuellement atteignables (clusters recouvrants)



Clustering Basé sur la Densité

- Les clusters obtenus peuvent avoir différentes tailles et formes, et les données bruitées sont implicitement identifiées (points isolés)



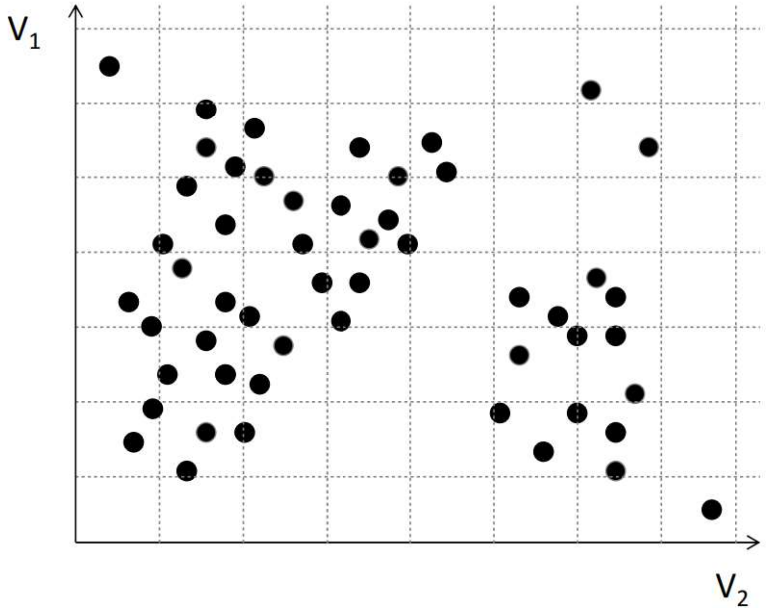
Clustering Basé sur la Densité

- Forces
  - Ne nécessite pas de définir a priori le nombre K de clusters
  - Les clusters peuvent avoir différentes tailles et formes
  - Robuste aux données bruitées et valeurs aberrantes
- Faiblesses
  - La complexité temporelle est $O(N^2)$ pour N instances, ou O (N.log(N)) si un index spatial est utilisée : le traitement de très grands ensembles de données peut être coûteux
  - Moins adéquate pour les variables discrètes que pour les variables continues
- Algorithmes les plus populaires : DBScan, DENCLUE, OPTICS

Clustering Basé sur les Grilles

- L'espace de données multidimensionnel est divisé en cellules
- Le calcul de la densité des régions (taille du voisinage) n'est plus centré sur les instances mais effectué pour chaque cellule

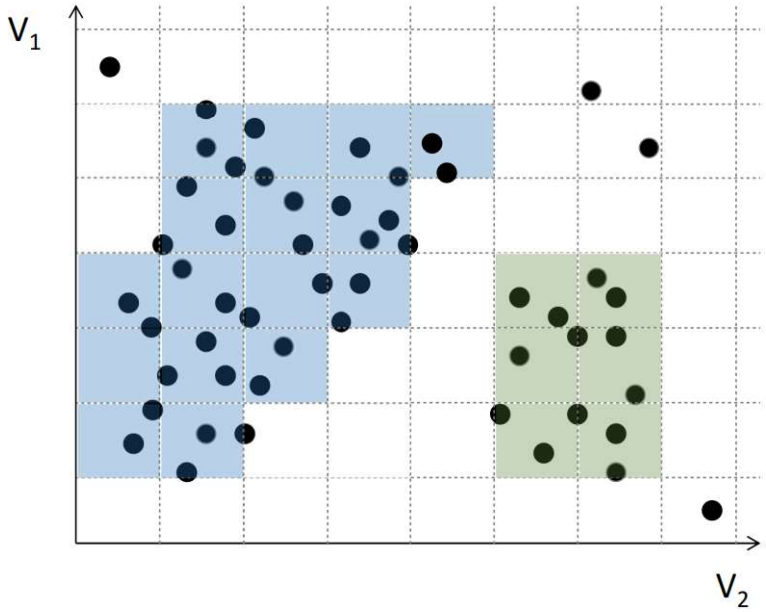
- Chaque variable est discrétisée en largeur (intervalles égaux)
- La largeur des intervalles de discrétisation détermine la taille et le nombre de cellules
- Son paramétrage permet de s'adapter à la distribution des valeurs de la variable



Clustering Basé sur les Grilles

- Les cellules denses sont les cellules qui contiennent un nombre d'instances au moins égal au seuil minimal défini par l'utilisateur (taille du voisinage)

- Les cellules denses adjacentes sont fusionnées pour former des clusters
- Les données bruitées et les valeurs aberrantes sont implicitement ignorées car elles sont contenues dans des cellules faiblement peuplées

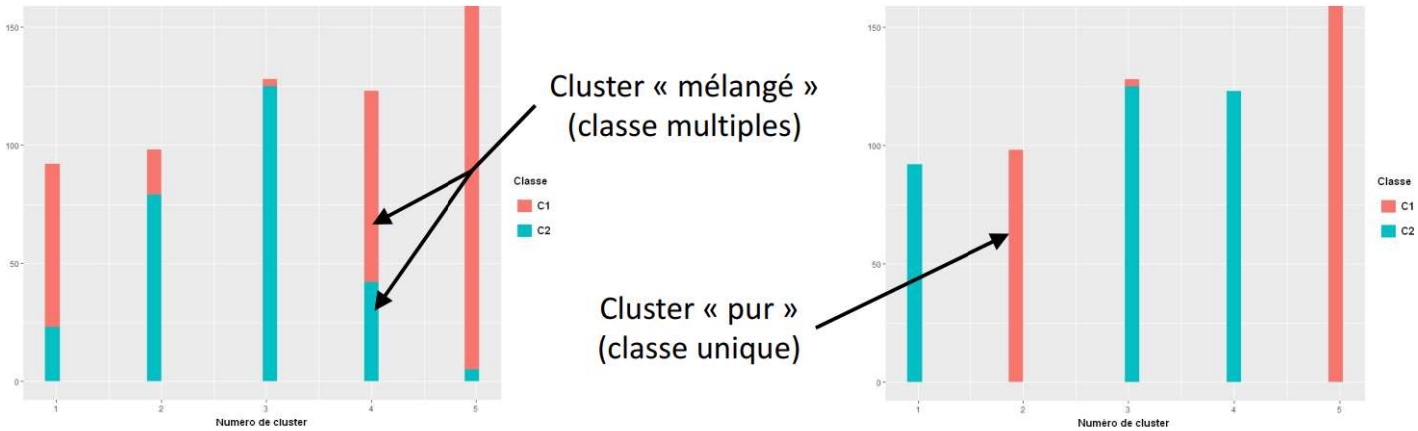


Clustering Basé sur les Grilles

- Forces
  - Approche basée sur la densité : pas de paramètre K, clusters de tailles et formes arbitraires, robustesse aux données bruitées et valeurs aberrantes
  - La paramétrisation de la résolution de la grille permet d'adapter le processus aux données : compromis entre efficacité et précision du résultat
  - Bonnes propriétés de passage à l'échelle par rapport au nombre d'instances et de dimensions
- Faiblesses
  - Les variables continues doivent être discrétisées, ce qui peut être difficile à optimiser automatiquement
- Algorithmes les plus populaires : STING, CLIQUE, WaveCluster

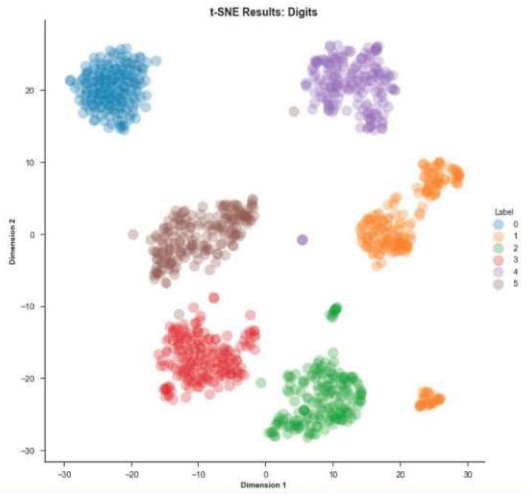
Évaluation des Clusters : Évaluation Externe

- **Évaluation externe** : dans le cas où nous disposons d'une variable de classe dans les données
- Évaluer la pertinence des clusters générés par dénombrement des instances de chaque classe dans chaque cluster
- Cas optimal : toutes les instances de chaque cluster sont de la même classe
- Exemple : histogrammes des effectifs des classes par cluster



Évaluation des Clusters : Évaluation Interne

- **Évaluation interne** : si nous ne disposons pas d'une variable de classe dans les données
- Comparaison des distribution des valeurs des variables entre clusters
  - Variables numériques : quartiles et moyenne
  - Variables discrètes : distribution des valeurs et mode (valeur la plus fréquente)
- Représentation bi/tri-dimensionnelle des données avec cluster en couleur
  - Calcul de deux ou trois composantes principales à partir des données ou de la matrice de distance (e.g. méthode t-SNE)
  - Affichage du nuage de points obtenu avec coloration des points par cluster



Références et Bibliographie

- Principales Librairies R
  - dbscan : implémentation optimisée de l'algorithme DBSCAN de clustering par densité utilisant une structure de données de type K-Dimensional Tree
  - pdfCluster : clustering par densité via l'estimation de la densité du noyau (clusters de composants maximalement connectées et de densité supérieure à un seuil, représentés dans une structure arborescente en sortie)
  - ADPclust : clustering de données de grande dimension sur la base d'un diagramme de décision bi-dimensionnel
  - fpc : fonctions de clustering par densité, ainsi que d'évaluation et de validation des clusters
  - Clustering : comparaison de différentes fonctions et approches de clustering (partitionnement, hiérarchique, etc.) afin de déterminer leur adéquation aux données fournies
  - tsne : visualisations bi-/tri-dimensionnelle des données par calcul de composantes pour affichage du nuage de points avec les clusters en couleur